Figure 2. Some of the same speaker's high vowels from a read list of Swedish words. Reference values from Eklund & Traunmüller (1997).

References

Wim A. van Dommelen
Department of Language and Communication Studies, NTNU
wim.van.dommelen@hf.ntnu.no

Abstract
This paper looks into the question of how to quantifiy rhythm in Norwegian spoken as a second language by speakers from different language backgrounds. The speech material for this study was taken from existing recordings from the Language Encounters project and consisted of sentences read by natives and speakers from six different L1s. Measurements of syllable durations and speech rate were made. Seven different metrics were calculated and used in a discriminant analysis. For the five utterances investigated, statistical classification was to a large degree in congruence with LI group membership. The results therefore suggest that L2 productions differed rhythmically from Norwegian spoken as LI.

1 Introduction
During the last few years a number of attempts have been made to classify languages according to rhythmic categories using various metrics. To investigate rhythm characteristics of eight languages, Ramus, Nespov & Mehler (1999) calculated the average proportion of vocalic intervals and standard deviation of vocalic and consonantal intervals over sentences. Though their metrics appeared to reflect aspects of rhythm structure, also considerable overlap was found. Grabe's Pairwise Variability Index (PVI; see section 2.2) is a measure of differences in vowel duration between successive syllables and has been used by, e.g., Grabe & Low (2002), Ramus (2002) and Stockmal, Markus & Bond (2005). In order to achieve more reliable results Barry, Andreeva, Russo, Dimitrova & Kostadinova (2003) proposed to extend existing PVI measures by taking consonant and vowel intervals together. The present paper takes an exploratory look into the question of how to quantify speech rhythm in Norwegian spoken by second language users. Seven metrics will be used, five of which being based on syllable durations. Two metrics are related to speech rate, and the last one is Grabe's normalized Pairwise Variability Index with syllable duration as a measure.

2 Method
2.1 Speech material
The speech material used for this study was chosen from existing recordings made for the Language Encounters project. These recordings were made in the department's soundinsulated studio and stored with a sampling frequency of 44.1 kHz. Five different sentences were selected consisting of 8, 10, 11, 11, and 15 syllables, respectively. There were six second language speaker groups with the following LI's (number of speakers in parentheses); Chinese (7), English (4), French (6), German (4), Persian (6) and Russian (4). Six native speakers of Norwegian served as a control group. The total number of sentences investigated was thus 37 x 5 = 185.
2.2 Segmentation and definition of metrics
The 185 utterances were segmented into syllables and labeled using Praat (Boersma & Weenink, 2006). Syllabification of an acoustic signal is not a trivial task. It was guided primarily by the consideration to achieve consistent results across speakers and utterances. In words containing a sequence of a long vowel and a short consonant in a context like V:CV (e.g., fine [nic-e]) the boundary was placed before the consonant (achieving fi-ne), after a short vowel plus long consonant as in minne (memory) after the consonant (miin-e). Only when the intervocalic consonant was a voiceless plosive, the boundary was always placed after the consonant (e.g. in mat-et [fed]).

To compare temporal structure of the L2 with the L1 utterances, seven different types of metrics were defined. In all cases calculations were related to each of the seven groups of speakers as a whole. The first metric was syllable duration averaged over all syllables of each utterance, yielding one mean syllable duration for each sentence and each speaker group. Second, the standard deviation for the syllable durations pooled over the speakers of each group was calculated for each of the single utterances' syllables. The mean standard deviation was then taken as the second metric, thus expressing mean variation of syllable durations across each utterance.

For the definition of the third and fourth metric let us look at Figure 1. In this figure, closed symbols depict mean syllable durations in the sentence To barn matet de tamme dyrene (Two children fed the tame animals) produced by six native speakers. The syllables are ranked according to their increasing durations. Similarly, the open symbols give the durations for the same syllables produced by the group of seven Chinese speakers. Note that the order of the syllables is the same as for the Norwegian natives. Indicated are regression lines fitted to the two groups of data points. The correlation coefficient for the relation between syllable duration and the rank number of the syllables as defined by the Norwegian reference is the third metric in this study (for the Chinese speaker group presented in the figure r= 0.541). Further, the slope of the regression line was taken as the fourth metric (here: 18.7). The vertical bars in Figure 1 indicate ± 1 standard deviation. The mean of the ten standard deviation values represents the second metric defined above (for Norwegian 27 ms; for Chinese 63 ms).

![Figure 1. Mean duration of syllables in a Norwegian utterance ranked according to increasing duration for six native speakers (closed symbols with regression line). Open symbols indicate mean durations for a group of seven Chinese subjects with syllable rank as for the L1 speakers. Vertical bars indicate ± 1 standard deviation.](image)

Table 1. Mean syllable durations and standard deviations in ms for six groups of L2 speakers and a Norwegian control group. Means are across five utterances and all speakers in the respective speaker groups (see 2.2).

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
<th>French</th>
<th>German</th>
<th>Persian</th>
<th>Russian</th>
<th>Norwegian</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>285</td>
<td>227</td>
<td>238</td>
<td>200</td>
<td>255</td>
<td>224</td>
<td>178</td>
</tr>
<tr>
<td>sd</td>
<td>115</td>
<td>107</td>
<td>98</td>
<td>91</td>
<td>102</td>
<td>111</td>
<td>84</td>
</tr>
<tr>
<td>n</td>
<td>387</td>
<td>220</td>
<td>330</td>
<td>220</td>
<td>329</td>
<td>220</td>
<td>330</td>
</tr>
</tbody>
</table>

3.2 Discriminant analysis
In order to investigate whether rhythmic differences between utterances from the different speaker groups can be captured by the seven metrics defined above, a discriminant analysis was performed. It appears from the results that in the majority of cases the L2-produced utterances were correctly classified (Table 2). The overall correct classification rate amounts to 94.3%. Only one utterance produced by the Chinese speaker group was classified as Persian and one utterance from the French group was confused with the category Russian.
functions reached statistical significance, cumulatively explaining 96.4% of the variance. For
discriminant analysis using seven metrics (see section 2.2).
Predicted LI group membership (percent correct) of five utterances according to a
36

The present results suggest that the utterances spoken by the second language users differed in
rhythmical structure from those produced by the native speakers. It was shown that it is
possible to quantify rhythm using direct and indirect measures. Though the statistical analysis
yielded promising results, it should be kept in mind that the number of utterances investigated
was relatively small. Therefore, more research will be needed to confirm the preliminary
results and to refine the present approach.

Acknowledgements
This research is supported by the Research Council of Norway (NFR) through grant
158458/530 to the project Språkmøter. I would like to thank Rein Ove Sikveland for the 
segmentation of the speech material.

References
measures tell us anything about language type? Proceedings 15th ICPiS, Barcelona, 2693-
2696.
Grabe, E. & E.L. Low, 2002. Durational variability in speech and the rhythm class hypothesis.
Prosody 2002, Aix-en-Provence, 115-120.
Cognition 73, 263-292.
quantitative language. Language and Speech 48, 55-63.

Abstract
This paper presents /nailon/ – a software package for online real-time prosodic analysis
that captures a number of prosodic features relevant for interaction control in spoken
dialogue systems. The current implementation captures silence durations; voicing, intensity,
and pitch; pseudo-syllable durations; and intonation patterns. The paper provides detailed
information on how this is achieved.

1 Introduction
All spoken dialogue systems, no matter what flavour they come in, need some kind of
interaction control capabilities in order to identify places where it is legitimate to begin to talk
to a human interlocutor, as well as to avoid interrupting the user. Most current systems rely
exclusively on silence duration thresholds for making such interaction control decisions, with
thresholds typically ranging from 500 to 2000 ms (e.g. Ferter, Shriberg & Stolcke, 2002).
Such an approach has obvious drawbacks. Users generally have to wait longer for responses
than in human-human interactions, but at the same time they run the risk of being interrupted
by the system. This is where /nailon/ – our software for online analysis of prosody and the
main focus of this paper – enters the picture.

2 Design criteria for practical applications
In order to use prosody in practical applications, the information needs to be available to the
system, which places special requirements on the analyses. First of all, in order to be useful in
live situations, all processing must be performed automatically, in real-time and deliver its
results with minimal latency (cf Shriberg & Stolcke, 2004). Furthermore, the analyses must
be online in the sense of relying on past and present information only, and cannot depend on
any right context or look-ahead. There are other technical requirements: the analyses should
be predictable and constant in terms of memory use, processor use, and latency. Finally,
although not a strict theoretical nor a technical requirement, it is highly desirable to use
concepts that are relevant to humans. In the case of prosody, measurements should be made
on psychoacoustic or perceptually relevant scales.

3 /nailon/
The prosodic analysis software /nailon/ was built to meet the requirements and to capture
silence durations; voicing, intensity, and pitch; pseudo-syllable durations; and intonation patterns. It implements high-level methods accessible through in Tcl/Tk and the low-level audio processing is handled by the Snack sound toolkit, with pitch-tracking based on the
ESPS tool get_f0. /nailon/ differs from Snack in that its analyses are incremental with
relatively small footprints and can be used for online analyses. The implementation is real-