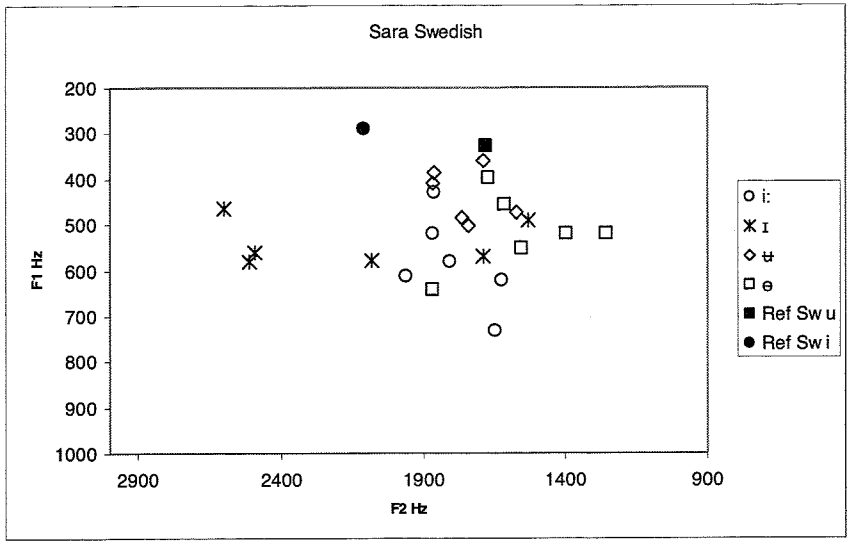

Figure 2. Some of the same speaker's high vowels from a read list of Swedish words. Reference values from Eklund & Traunmüller (1997).

References

Cunningham, U., 2003. Temporal indicators of language dominance in bilingual children. *Proceedings from Fonetik 2003, Phonum 9*, Umeå University, 77-80.

Cunningham, U., 2004. Language Dominance in Early and Late Bilinguals. *ASLA*, Södertörn.

Eklund, I. & H. Traunmüller, 1997. Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica 54*, 1-21.

Flege, J.E., C. Schirru & I.R.A. MacKay, 2003. Interaction between the native and second language phonetic subsystems. *Speech Communication 40*, 467-491.

Jenkins, J., 2004. Research in teaching pronunciation and intonation. *Annual Review of Applied Linguistics 24*, 109-125.

Kachru, B. (ed.), 1992. *The Other Tongue* (2nd edition). Urbana and Chicago: University of Illinois Press.

Mauranen, A., 2003. Academic English as lingua franca—a corpus approach. *TESOL Q. 37*, 513-27.

McAllister, R., J.E. Flege & T. Piske, 2002. The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics 30*(2), 229-258.

McArthur, T., 2003. World English, Euro-English, Nordic English? *English Today 73*(19), 54-58.

Phillipson, R., 1992. *Linguistic Imperialism*. Oxford: Oxford Univ. Press.

Regeringen, 2005. *Bästa språket – en samlad svensk språkpolitik*. Prop. 2005/06:2.

Seidlhofer, B., 2005. English as a lingua franca. *ELT Journal 59*(4), 339-341.

Skolverket, 2006. *Förslag till kursplan*.

Tarone, E., 1983. On the variability of interlanguage systems. *Applied Linguistics 4*, 142-163.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 33–36

Quantification of Speech Rhythm in Norwegian as a Second Language

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 quantify 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 L1 group membership. The results therefore suggest that L2 productions differed rhythmically from Norwegian spoken as L1.

1 Introduction

During the last few years a number of attempts have been made to classify languages according to rhythmical categories using various metrics. To investigate rhythm characteristics of eight languages, Ramus, Nespor & 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 rhythmic 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 sound-insulated 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 L1s (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* [nice]) the boundary was placed before the consonant (achieving *fi-ne*), after a short vowel plus long consonant as in *minne* (memory) after the consonant (*minn-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).

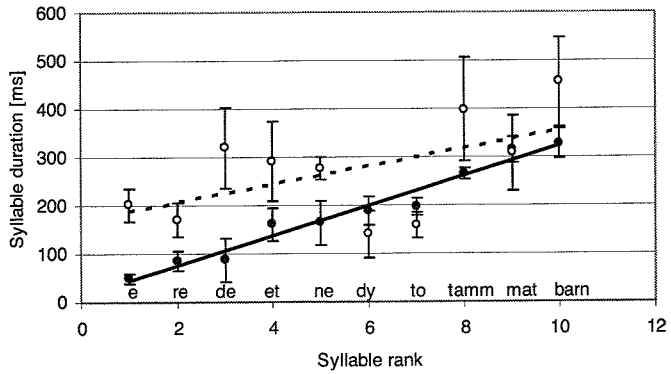

Figure 1. Mean duration of syllables in a Norwegian utterance ranked according to increasing duration for six native speakers (closed symbols with regression line). Open symbols indicate mean durations for a group of seven Chinese subjects with syllable rank as for the L1 speakers. Vertical bars indicate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

As metric number five speech rate was chosen, calculated as the number of (actually produced) phonemes per second. The standard deviation belonging to the mean number of phonemes served as the sixth metric. In both cases, there was one single value per utterance and speaker group. Finally, the seventh metric was the normalized Pairwise Variability Index (nPVI) as used by Grabe & Low (2002):

$$(1)\quad \text{nPVI} = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \right]$$

In this calculation the difference of the durations (d) of two successive syllables is divided by the mean duration of the two syllables. This is done for all (m-1) successive syllable pairs in an utterance (m= the number of syllables). Finally, by dividing the sum of the (m-1) amounts by (m-1) a mean normalized difference is calculated and expressed as percent.

3 Results

3.1 Mean syllable duration

Since the main temporal unit under scrutiny is the syllable, let us first see whether and to what extent the various speaker groups produced different syllable durations. As can be seen from Table 1, mean syllable durations vary substantially. Shortest durations were found for the natives (178 ms), while the subjects with a Chinese L1 produced the longest syllables (285 ms). The other groups have values that are more native-like, in particular the German speakers with a mean of 200 ms. For all speaker groups the standard deviations are quite large, which is due to both inter-speaker variation and the inclusion of all the different types of syllables. (Note that the standard deviation described here is different from the second metric; see 2.2.) According to a one-way analysis of variance, the overall effect of speaker group on syllable duration is statistically significant ($F(6, 2029) = 40.322$; $p < .0001$). Calculation of a Games-Howell post-hoc analysis resulted in the following homogeneous subsets (level of significance p= 0.05): (Chinese); (English, French, German, Russian); (French, English, Persian, Russian); (German, Norwegian, English, Russian); (Persian, French); (Russian, English, French, German); (Norwegian, German). It is thus obvious that syllable durations overlap considerably and do not really distinguish the speaker groups.

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).

	Chinese	English	French	German	Persian	Russian	Norwegian
mean	285	227	238	200	255	224	178
sd	115	107	98	91	102	111	84
n	387	220	330	220	329	220	330

3.2 Discriminant analysis

In order to investigate whether rhythmical 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.

Table 2. Predicted L1 group membership (percent correct) of five utterances according to a discriminant analysis using seven metrics (see section 2.2).

	Predicted L1 group membership						
L1 group	Chinese	English	French	German	Persian	Russian	Norwegian
Chinese	**80**	0	0	0	20	0	0
English	0	**100**	0	0	0	0	0
French	0	0	**80**	0	0	20	0
German	0	0	0	**100**	0	0	0
Persian	0	0	0	0	**100**	0	0
Russian	0	0	0	0	0	**100**	0
Norwegian	0	0	0	0	0	0	**100**

The results of the discriminant analysis further showed that three of the six discriminant functions reached statistical significance, cumulatively explaining 96.4% of the variance. For the first function, the metrics with most discriminatory power were slope (metric 4), speech rate (metric 5) and mean syllable duration (metric 1). The second discriminant function had also slope and speech rate, but additionally standard deviations for speech rate (metric 6) and for syllable duration (metric 2), and nPVI (metric 7) as important variables. Finally, of highest importance for the third function were metrics 5, 3 (correlation coefficient), 4, and 7, in that order.

4 Conclusion

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

Barry, W.J., B. Andreeva, M. Russo, S. Dimitrova & T. Kostadinova, 2003. Do rhythm measures tell us anything about language type? *Proceedings 15th ICPhS*, Barcelona, 2693-2696.

Boersma, P. & D. Weenink, 2006. Praat: doing phonetics by computer (Version 4.4.11) [Computer program]. Retrieved February 23, 2006, from http://www.praat.org/.

Grabe, E. & E.L. Low, 2002. Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7*. Berlin: Mouton, 515-546.

Ramus, F., 2002. Acoustic correlates of linguistic rhythm: Perspectives. *Proceedings Speech Prosody 2002*, Aix-en-Provence, 115-120.

Ramus, F., M. Nespor & J. Mehler, 1999. Correlates of linguistic rhythm in the speech signal. *Cognition 73,* 265-292.

Stockmal, V., D. Markus & D. Bond, 2005. Measures of native and non-native rhythm in a quantity language. *Language and Speech 48,* 55-63.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 37–40

`/nailon/` – Online Analysis of Prosody

Jens Edlund and Mattias Heldner
Department of Speech, Music and Hearing, KTH, Stockholm
{edlund|mattias}@speech.kth.se

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. Ferrer, 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 sufficiently general to work for many speakers and many domains, and 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-