

# Non-native durational patterns decrease speech intelligibility

Hugo Quené\*, L. E. van Delft

*Utrecht institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands*

---

## Abstract

In native speech, durational patterns convey linguistically relevant phenomena such as phrase structure, lexical stress, rhythm, and word boundaries. The lower intelligibility of non-native speech may be partly due to its deviant durational patterns. The present study aims to quantify the relative contributions of non-native durational patterns and of non-native speech sounds to intelligibility. In a Speech Reception Threshold study, duration patterns were transplanted between native and non-native versions of Dutch sentences. Results for non-transplanted, original versions show that intelligibility thresholds (critical speech-to-noise ratios) differed by about 4 dB between native and non-native versions. Results for manipulated versions with transplanted durations indicate that about 0.4 to 1.1 dB of this difference was due to the durational patterns, and that this contribution was larger if the native and non-native patterns were more deviant. The remainder of the difference must have been due to non-native speech sounds in these materials. This finding supports recommendations to attend to durational patterns as well as native-like speech sounds, when learning to speak a foreign language.

*Key words:* non-native speech, duration patterns, segmental durations,

---

\*Corresponding author

*Email address:* [h.quene@uu.nl](mailto:h.quene@uu.nl) (Hugo Quené)

## 1. Introduction

Many studies have attempted to relate the generally lower intelligibility of non-native (L2) speech to its phonetic characteristics. Among other factors, intelligibility has been found to depend on segmental errors (e.g. Anderson-Hsieh, Johnson, and Koehler, 1992; Derwing and Munro, 1997), intonation (e.g. Anderson-Hsieh et al., 1992; Chun, 2002; Holm, 2008), and rhythmic and durational patterns (Tajima, Port, and Dalby, 1997; Mareüil and Vieru-Dimulescu, 2006; Holm, 2008). Methodological difficulties have prevented researchers from evaluating the relative contributions of these characteristics (for discussions, see e.g. Rajadurai, 2007; Holm, 2008). The aim of the present study is to overcome these difficulties by employing the Speech Reception Threshold method (SRT, Plomp and Mimpen, 1979), and to quantify the relative contributions of segmental errors and of durational patterns to L2 intelligibility.

Durational patterns are of particular importance for L2 intelligibility, because they convey many linguistic distinctions in spoken languages such as English and Dutch (for overviews, see Klatt, 1976; Van Santen, 1994; Nootboom, 1997). Segmental durations may cue the distinction between voiced and voiceless consonants (see Slis and Cohen, 1969), as well as the location of word boundaries in the continuous stream of speech sounds (Quené, 1992; Shatzman and McQueen, 2006). Durations also indicate phrase boundaries (Klatt, 1976; Van Santen, 1994) and lexical and phrasal stress (Eefting and Nootboom, 1993; Sluijter and Van Heuven, 1995). The rhythm conveyed by stressed and unstressed syllables plays an important role in word recognition and word segmentation (for reviews, see Cutler, Dahan, and van Donselaar,

1997; Goetry and Kolinsky, 2000; Quené and Port, 2005; Patel, 2008), at least in English and Dutch.

Languages differ in their native durational patterns. Some languages have a contrast between phonologically long and short vowels (e.g. Finnish, Dutch), whereas others do not (e.g. Spanish). In many languages, vowels are longer before voiced than before voiceless consonants, but the magnitude of this difference varies from large (e.g. English) to medium (e.g. French) to very small (e.g. Arabic; Flege and Port, 1981; Laeuffer, 1992). These native durational patterns are assumed to carry over to L2 realizations, yielding timing patterns in L2 that deviate from native patterns. Indeed, many cross-linguistic studies confirm that durational patterns in L2 differ from native norms (e.g. Adams, 1979; Flege, 1993; White and Mattys, 2007a; Bent, Bradlow, and Smith, 2008).

Deviant durational patterns are indeed detrimental to speech intelligibility, as shown by a few prosody-transplantation studies involving deaf speech (Maassen and Povel, 1984) and non-native speech (Tajima et al., 1997; Mareüil and Vieru-Dimulescu, 2006; Holm, 2008). These studies show a positive effect of native-like durational patterns on the accuracy of listeners' responses, but the *relative* perceptual contributions by the native-like speech segments and by native-like segmental duration patterns cannot be assessed from the resulting accuracy percentages. To illustrate, let us consider a fictitious experiment, yielding average accuracy percentages of 98% (condition I: native segments, native durations), 88% (II: native segments, non-native durations), and 68% (III: non-native segments, non-native durations). Comparing differences in these percentages (I-II: 10% difference; II-III: 20% difference, i.e., twice as large) is misleading, because an arcsine

transformation of the percentages is required (Studebaker, 1985)<sup>1</sup>. After arcsine transformation, comparing differences in arcsine units (I-II: 0.29 arcsine units; II-III: 0.33 arcsine units) more appropriately suggests that both effects are approximately equally important for intelligibility. Similar problems arise when comparing accuracy percentages across studies.

The aim of the present study is to assess the relative contribution of native vs non-native durational patterns to intelligibility, relative to the native vs non-native speech sounds. Hence the nativeness of the speaker (L1 Dutch vs L2 Dutch) and of the durational patterns (L1 Dutch vs L2 Dutch) will be varied independently, yielding the following  $2 \times 2$  conditions:

S1D1 Native L1 Dutch speaker (S1) with original, native L1 duration patterns (D1)

S1D2 Native L1 Dutch speaker (S1) with manipulated, non-native L2 duration patterns (D2)

S2D1 Non-native L2 Dutch speaker (S2) with manipulated L1 duration patterns (D1)

S2D2 Non-native L2 Dutch speaker (S2) with original, non-native L2 duration patterns (D2)

(Note that the term “native” follows the perspective of the Dutch listeners in the present study.)

The four conditions above correspond to three planned contrasts or predictions. Contrast A predicts that native speech from S1 yields higher intelligibility than non-native speech of S2, irrespective of the durational patterns;

---

<sup>1</sup>Note that one should transform the underlying by-listener or by-item percentages, and not by-condition average percentages as we do here.

this trivial contrast captures the main effect of speaker nativeness on intelligibility (cf. e.g. Derwing and Munro, 1997; Tajima et al., 1997; Munro and Derwing, 1999; Van Wijngaarden, 2001; Van Wijngaarden, Steeneken, and Houtgast, 2002; Munro, 2008). Contrast B predicts that the intelligibility of the L1 speech by S1 decreases if the durational patterns in the L1 speech are similar to those in non-native L2 speech. Contrast C predicts that the intelligibility of the L2 speech by S2 will *increase* if the durational patterns in the L2 speech are similar to those in native L1 speech. This last contrast concerns the potential gain in intelligibility by producing correct durational patterns in one’s L2 speech.

Speech intelligibility will be quantified in this study by means of the Speech Reception Threshold method (SRT, Plomp and Mimpen, 1979). Briefly, this method returns the speech-to-noise ratio (SNR) at which 50% of a list of 13 sentences is understood correctly. Lower values (i.e., lower SNR) indicate better intelligibility. Because each threshold is determined over a list of multiple sentences, results may be generalized to other similar sentences, and they may also be compared between conditions and between experiments (Plomp, 1986). Previous studies using this SRT method (Van Wijngaarden, 2001; Van Wijngaarden et al., 2002) have found that the SRT values of native and non-native speech differ by about 2 to 4 dB, depending on the speaker. In other words, the SNR for native speech is about 3 dB lower (better) than that for non-native speech, for native listeners. (For reference, an increase of +3 dB in speech intensity corresponds to a reduction of the speaker–listener or speaker–microphone distance by a factor of  $1/\sqrt{2} \approx 0.71$ , under idealized conditions). In these studies, however, the effects of speaker nativeness and durational pattern nativeness were not separated.

The present study, then, combines the two approaches discussed above.

It uses stimulus materials with speaker and duration patterns varied orthogonally (similar to studies by Maassen and Povel, 1984; Tajima et al., 1997; Mareüil and Vieru-Dimulescu, 2006; Holm, 2008). The SRT method (Van Wijngaarden, 2001; Van Wijngaarden et al., 2002) is used to assess and compare the relative contributions of these two factors.

Since the SRT method requires many sentence recordings (130, instead of about 12), only one speaker was included for L1 Dutch and one speaker for L2 Dutch in the present study, similar to the one by Tajima et al. 1997. Of course, using a single L2 speaker (and a single L1-L2 language contrast) limits the generalizability of this study. Ideally, one would include multiple speakers from multiple languages (e.g., White and Mattys, 2007b; Holm, 2008). Other L2 speakers may indeed show other deviations in their segmental errors and in their durational patterns, and the effects of these two types of deviations on intelligibility may also vary between speakers. Nevertheless, we regard our results as relevant for our goals, for two reasons. First, the currently selected L2 speaker has previously been judged as having a strong foreign accent, and relatively poor intelligibility (Van Wijngaarden et al., 2002). Her durational patterns (reported in section 2.1.2 below) deviate from L1 patterns in some, but not all respects. Thus the present L2 speaker probably displays both segmental and durational errors. Second, the lower number of speakers (and the SRT method) allows us to include more sentence recordings, thus increasing generalizability over sentences.

## 2. Methods

### 2.1. Materials

#### 2.1.1. Original recordings

The Dutch SRT method (Plomp and Mimpen, 1979) uses 10 lists of 13 sentences, carefully composed, containing 8 or 9 syllables, and describing everyday situations. Two example sentences are: *De bal vloog over de schutting* “The ball flew over the fence”; *Eindelijk kwam de trein op gang* “Finally the train started moving”. These native Dutch sentences had been read by a female native speaker (S1) of Standard Dutch. These spoken sentences are used as test materials in Dutch speech audiometry; these recordings were kindly provided by Joost Festen (Vrije Universiteit, Amsterdam).

The same sentences of Plomp and Mimpen (1979) had been spoken in Dutch by a female native speaker (S2) of Polish, and recorded and used by Van Wijngaarden et al. (2002); these materials were kindly provided by Sander van Wijngaarden (formerly at TNO, Soesterberg). The L2 speech sounded fluent, but it had non-native intonation and duration patterns that were indeed different from those in the L1 speech, as assessed by the authors and one other expert. The intelligibility of this L2 speech was reported to be sufficiently high, viz. near 100% correct at +15 dB SNR in a pre-test (Van Wijngaarden, 2001), so that valid SRTs could be obtained.

#### 2.1.2. Duration patterns

In order to further investigate whether and how duration patterns differed, we analyzed the durations of vowel segments, as vowels are more susceptible to durational effects than consonants are (Klatt, 1976; Nootboom, 1997). Vowel durations in the realizations by S1 and S2 were taken from the segmental intervals described below ( $N = 2 \times 590$  vowels). Durations of vowel

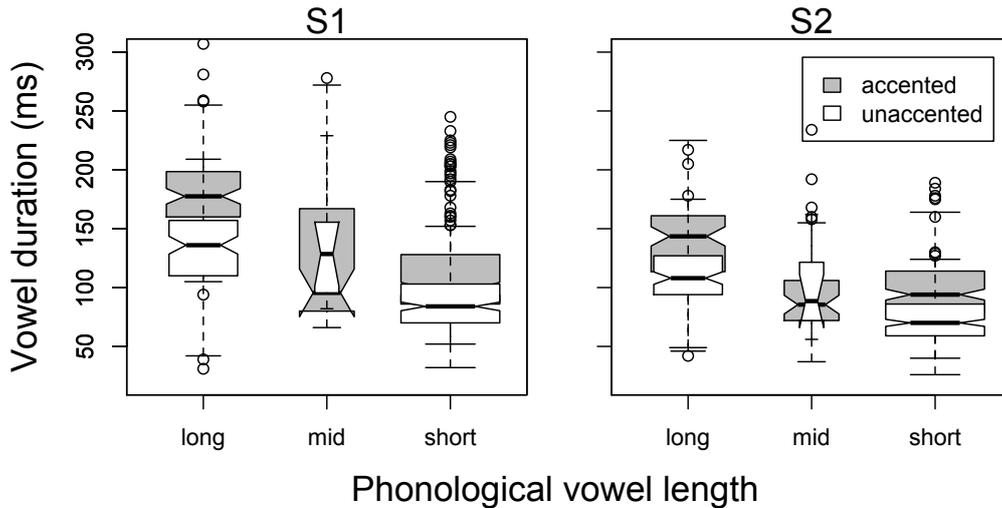
segments were compared by means of mixed effects modeling (Quené and Van den Bergh, 2004, 2008; Baayen, Davidson, and Bates, 2008). Speaker’s language (S1, S2), phonological vowel length (long, mid, short), sentence accent (unaccented, accented), and sentence-final position (non-final, final<sup>2</sup>) were included as 4 fixed effects; the 130 sentences were included as a random effect. The results showed a significant main effect of final position. On average, final vowels were 15 ms longer than non-final vowels ( $p < .001$ ); this main effect did not interact with speaker’s language. The main effects of phonological vowel length and accent, however, did interact with speaker’s language, as illustrated in Fig. 1.

The segmental durations summarized in Fig. 1 show that speaker S2 produced smaller differences due to phonological vowel length than S1 did. The phonologically long vowels of S2 are significantly shorter than those of S1 ( $p < .001$ ). S1’s long vowels have longer durations than her phonologically short vowels ( $p < .001$ ), whereas this difference is not significant for S2. Secondly, speaker S1 produced significantly longer durations for accented long vowels than for unaccented long vowels ( $p < .001$ ), whereas this difference was not significant for S2. As a result of these two differences, vowel durations of S1 have somewhat higher standard deviation (S1:  $\Delta V = 0.050$ ; S2:  $\Delta V = 0.037$ ) and higher coefficient of variation (S1:  $\text{Varco}V = 41.5$ ; S2:  $\text{Varco}V = 38.3$ ; e.g. White and Mattys, 2007a). These findings confirm that duration patterns are indeed different between the two speakers included in this study. The non-native speaker S2 produced phonologically long vowels that are too short, and even more so if these long vowels are accented. Hence,

---

<sup>2</sup>Syllables between and including the last accented syllable and the end of sentence were marked as final (cf. Cambier-Langeveld, 2000, Ch.2); most sentences had an accented single final syllable, and most of the unaccented final syllables contained schwa vowels.

Figure 1: Boxplots of vowel durations (in ms), for the native (left) and non-native speaker (right), broken down for phonologically long, mid and short vowels (horizontal axis) and accented (gray) and unaccented (white) vowels (with final and non-final vowels pooled). Box width corresponds with number of observations; box notches correspond with approximate 95% confidence intervals of the box' median.



both phonological vowel length and sentence accent are indicated less clearly in the L2 speech used in this study, as compared to the matching L1 speech.

### 2.1.3. Manipulations

The purpose of the phonetic manipulations was to exchange the patterns of segmental durations between the native and non-native realizations of each test sentence. In addition, differences in tempo (speaking rate) and in pitch contours were removed, because these might act as confounds in the intelligibility comparisons. All analyses and manipulations were performed by means of Praat software (Boersma and Weenink, 2008).

In step 1, the matching S1 and S2 versions of each sentence were segmented into the same number of intervals, by means of visual and auditory feedback, using criteria derived from Peterson and Lehiste (1960) and Klatt

(1976). Each speech sound and silent pause was marked as a separate interval. If adjacent sounds were strongly coarticulated and difficult to segment, then these were considered as a single interval in both of corresponding S1 and S2 realizations; approx. 10% of the resulting intervals contained multiple segments.

In step 2, the speaking rate was equalized among the two versions, by means of PSOLA manipulation as implemented in Praat (Moulines and Charpentier, 1990; Boersma and Weenink, 2008). This yielded an intermediate duration for both versions of a test sentence, while the durational pattern was preserved: the relative durations of the intervals were unaltered.

Step 3 was only required for the conditions with durational manipulation (S1D2 and S2D1), and not for the other conditions (S1D1 and S2D2). In this step, each interval in one version of a test sentence was lengthened or shortened to match the duration of the corresponding interval in the other version of that sentence. Manipulation was again done by means of PSOLA. In a few cases, intervals had to be lengthened by more than a factor of two, which yielded unnatural speech containing chirp or whistle sounds (Moulines and Charpentier, 1990). These few cases were corrected by hand, mostly by adjusting the target pitch values.

In step 4, the pitch contours of the S1 sentences were stylized (with occasional manual correction of pitch detection errors). This stylized pitch contour was also used to replace the pitch contours of the matching S2 sentences. Finally, all test sentences were set at a mean intensity level of 70 dB (SPL) and resampled to 44.1 kHz.

#### *2.1.4. Masking noise*

The masking noise used in the SRT method had a spectrum similar to the masked speech. The two female speakers in this study yielded slightly

different long-term average spectra; in order to avoid varying speech–noise interaction as a possible confound, separate noise maskers were used for each list by each speaker in each condition. To this end, the long-term average speech spectrum was determined for each combination of list, speaker and condition. The resulting spectrum was used to filter a white noise signal, yielding a masker noise signal matching the masked speech. The average spectra of the masking noise are indeed slightly different for the two speakers S1 and S2, as illustrated by the spectra for the two conditions with unmanipulated durations (S1D1 and S2D2) in Figure 2.

## 2.2. *Participants and procedure*

A total of 40 native Dutch listeners (3 male, 37 female; ages 18 to 28) listened to the test sentences. All participants had self-reported normal hearing and no speech or language disorder, and no passive or active knowledge of Polish (since few Polish speakers learn Dutch, it was assumed that listeners had not been previously exposed to Polish-accented L2 Dutch). Participants were recruited from a volunteer subjects pool at Utrecht University; they enrolled voluntarily and received €5 for their participation.

Participants were seated in a sound-attenuated booth wearing closed headphones. Each participant was instructed to repeat the Dutch sentences as perceived, with clear pronunciation. The experimenter decided immediately whether the spoken response was correct. (If the experimenter was in doubt, the participant was asked to repeat the last response. Minor deviations in unaccented words (e.g. *werd* “became” reported as *wordt* “becomes”) were accepted as correct, but all accented words had to be repeated correctly.

Listeners first responded to two practice lists with unmanipulated durations, with native speaker (S1D1) and non-native speaker (S2D2). The other 8 lists were used for data collection, with 2 lists per condition. Lists and

conditions were rotated over participants; the order of lists (and conditions) was randomized for each participant.

The speech-to-noise ratio (SNR) of the sentences was varied according to the up-and-down strategy of Plomp and Mimpen (1979), also employed by Van Wijngaarden (2001) and Van Wijngaarden et al. (2002). Adjustments were done automatically after feedback from the experimenter about whether the spoken response was correct or incorrect. (This feedback was not visible for the participant.)

For each list, the first sentence started at  $-15$  dB SNR, and it was presented repeatedly until the participant responded with the correct sentence. The SNR of this first sentence was increased by  $+4$  dB in the first and second presentation, and by  $+2$  dB on each subsequent presentation, to quickly approach the 50% threshold. Then each of the subsequent 12 sentences of the list was presented once. If responded correctly, the SNR of the next sentence in the list was adjusted by  $-2$  dB; if responded incorrectly, the SNR was adjusted by  $+2$  dB. After the last (13th) sentence, the SNR at which a subsequent sentence in the list would have been presented was also recorded. The 50% intelligibility threshold (SRT) for each presented list was obtained by averaging the last 10 SNRs of the list, i.e. the SNR used in sentences 4 to 14 (although no 14th sentence was presented)<sup>3</sup>.

### *2.3. Results*

The resulting Speech Reception Thresholds for each of the 4 test conditions are summarized in Figure 3. The 4 test conditions correspond to 3 orthogonal contrasts matching the predictions above.

---

<sup>3</sup>The sequence of 10 SNRs always contained multiple changes in the sign of the adjustment, so that the obtained SRT closely approximates the listener’s intelligibility threshold.

The first contrast (A) compares the native-speaker conditions (S1D1 and S1D2) against the non-native-speaker conditions (S2D1 and S2D2). As predicted, this contrast yields a large difference of about 3 dB in SRT, as shown in Figure 3. The second contrast (B) compares the two conditions with a native speaker, viz. with original (S1D1) and manipulated (S1D2) durational patterns. As predicted, the SRT increases (i.e., intelligibility decreases), by about 1 dB, if the native L1 speech is manipulated to have non-native durational patterns.

The third contrast (C) is most relevant for L2 acquisition. Does intelligibility of non-native speech improve if the non-native speech has native-like durational patterns? The third contrast compares the two conditions with the non-native speaker, viz. with original (S2D2) and manipulated (S2D1) durational patterns. The SRT decreases, by about  $-0.4$  dB, if the non-native L2 speech is manipulated to have native-like durational patterns.

These three effects, visible in Figure 3, were evaluated by means of mixed-effects modeling; this allows us to simultaneously include participants as well as sentence lists as random effects (for details and references, see e.g. Quené and Van den Bergh, 2004, 2008; Baayen et al., 2008), using the `lme4` package in R (Bates, 2005; R Development Core Team, 2008). The full model contains all three contrasts in the fixed part, here denoted as  $\gamma_A$ ,  $\gamma_B$  and  $\gamma_C$ , respectively. The random part contains crossed effects of participants  $u_{0(j0)}$  and of sentence lists  $v_{0(0k)}$ , as well as residual component  $e_{i(jk)}$ :

$$Y_{i(jk)} = \gamma_{0(00)} + \gamma_{A0(00)}L2 + \gamma_{B0(00)}S1manip + \gamma_{C0(00)}S2manip + (u_{0(j0)} + v_{0(0k)} + e_{i(jk)}) \quad (1)$$

Coefficients of this model were estimated using the full Maximum Likelihood criterion, to allow comparisons with other models described below (Pinheiro and Bates, 2000; Hox, 2002, Ch.3).

The estimated coefficients of this full model (1) are reported in Table 1. The fixed coefficients  $\gamma_A$  for contrast A and  $\gamma_B$  for contrast B are indeed highly significant, as shown by their large magnitudes and by the large decrease in deviance (in comparison to the intercept-only or “empty” model; see also Figure 3). The effect of contrast C ( $\gamma_C$ ), although relatively small, is also significant. To assess this critical contrast C further, the full model (1) containing this difference in thresholds was compared to a simpler model (2) without this critical effect (Hox 2002, Ch.3; Faraway 2006).

$$Y_{i(jk)} = \gamma_{0(00)} + \gamma_{A0(00)}L2 + \gamma_{B0(00)}S1manip \quad (2) \\ + (u_{0(j0)} + v_{0(0k)} + e_{i(jk)})$$

This so-called Likelihood Ratio Test yielded a significant result [ $\chi^2(1) = 4.08, p = .043$ ], as listed in Table 1.

However, this Likelihood Ratio Test of significance is regarded as anti-conservative (Pinheiro and Bates, 2000, §2.4.2), and the effect was therefore further investigated by a test procedure recommended by Pinheiro and Bates (2000, §2.4.1) and Faraway (2006, §8.2). For this test, responses were simulated with parametric bootstrapping under the simpler model (2). These simulated responses were then analysed by the full model (1) and by the simpler model (2), and outcomes were again compared with a Likelihood Ratio Test. These steps were repeated 1000 times. A proportion of .047 (s.e. 0.007) of the resulting likelihood ratios, from responses simulated from the simpler model, exceed the above-observed likelihood ratio of the full model. This proportion of .047 may be regarded as an unbiased bootstrap estimate of the  $p$  value of the critical contrast  $\gamma_C$  (Faraway, 2006, §8.2). In sum, the critical third contrast has a small but significant effect on intelligibility, so that speech reception thresholds of L2 speech are slightly lower (i.e., intelligibility is higher) if the duration patterns in the L2 speech are similar to

Table 1: Estimated parameters (with standard error of estimate in parentheses) of mixed-effects modeling with participants and sentence lists as two crossed random effects.

	empty model	Model (2)	Model (1)
<b>fixed</b>			
intercept, $\gamma_{0(00)}$	-1.030 (0.147)	-1.030 (0.166)	-1.030 (0.166)
Speaker, $\gamma_{A0(00)}$		3.155 (0.142) *	3.155 (0.141) *
S1manip, $\gamma_{B0(00)}$		1.065 (0.201) *	1.065 (0.200) *
S2manip, $\gamma_{C0(00)}$			-0.405 (0.200) *
<b>random</b>			
participants, $\sigma_{u_{0(j0)}}^2$	< 0.001	0.216	0.219
lists, $\sigma_{v_{0(0k)}}^2$	0.059	0.136	0.137
residual, $\sigma_{e_{i(jk)}}^2$	4.538	1.622	1.597
<b>evaluation</b>			
-2 log(lh)	1395	1103	1099
$\chi^2_{\text{deviance}}$		292.40 (df=2)	4.08 (df=1)
$p_{\text{deviance}}$		$p < .0001$	$p = .043$
* $p < .05$			

those in L1 speech.

In addition, one might predict that the *amount* of decrease of SRT due to manipulation should reflect the *amount* of difference in duration patterns between non-native and native realizations (S2D2 and S2D1). If an S2D2 sentence has relatively deviant duration patterns, then manipulating that sentence to S2D1 should yield a relatively large increase in intelligibility (decrease in SRT). This was further investigated by computing the *difference* in duration between two matching vowels in the S1 and S2 realizations, based on the raw vowel durations described in section 2.1.2 above. Since our experimental method yields a single SRT response for a list of 13 sentences, these duration differences were also aggregated per list. Per-list means of these differences would not be informative (as negative and positive differences would cancel each other out); hence we used the median absolute deviation (mad) of these differences. Figure 4 confirms that the amount of per-list change in intelligibility (decrease in SRT) between conditions S2D2 and S2D1 (i.e., contrast C above) is indeed correlated to the amount of per-list deviance in duration patterns ( $r = -.76, p = .028$ ). Lists that have relatively large median average deviations in vowel duration benefit from our manipulations (about  $-1$  dB shift in SRT), whereas lists with relatively small deviations do not.

### 3. Discussion and conclusion

The results show clear effects of both the speech segments (speaker nativeness) and the durational patterns. If both factors differ jointly (S1D1 vs S2D2), then the SRT shifts by about 4 dB in the present study. This difference is within the expected range (Van Wijngaarden et al., 2002, report an average shift of 3.6 dB for Polish speakers), and it is equivalent to a

reduction in speaker–listener or speaker–microphone distance by a factor of 0.63. Pitch contours and speaking rate were held constant among matching sentences across conditions, to exclude these as possible confounds. The shift can therefore only be due to segmental differences and/or different durational patterns.

The results for the conditions with transplanted durations (S1D2 and S2D1) show that durational patterns do indeed influence the resulting SRTs, as predicted. This reconfirms the perceptual importance of appropriate durational patterns (Maassen and Povel, 1984; Tajima et al., 1997; Mareüil and Vieru-Dimulescu, 2006; Holm, 2008; Bent et al., 2008). But quantifying the effect of durational patterns is somewhat complicated, because the effect of transplanting durations is larger for the native speech (1.1 dB) than for the non-native speech (0.4 dB). If the speech contains more non-native speech segments, then durations are apparently less relevant for intelligibility. Stated simply, this suggests that hearing a native-like speech sound in an inappropriate durational pattern may be relatively worse than hearing a non-native-like sound in an appropriate duration pattern. (However, L1 materials had originally been recorded for audiological measurements, with very clear pronunciation and slightly formal accent, which may have made durational manipulations in this material relatively more conspicuous than in the L2 materials.)

Moreover, the variation between sentence lists (each list contains 13 sentences) indicates that the effect of duration transplantations is larger as the L2 durations are more deviant from the L1 durations; this finding further confirms the perceptual contribution of duration patterns to intelligibility. For a few lists, the gain in intelligibility is about 1 dB SRT, or about one quarter of the difference in intelligibility between L1 and L2 speech. The

major part of the difference in intelligibility must be ascribed then to segmental errors, either phonemic or sub-phonemic, which are indeed known to impede speech communication (Anderson-Hsieh et al., 1992; Derwing and Munro, 1997; Smith, 2004).

The finding that segmental errors played a relatively larger role than deviant durational patterns, may have been due in part to the unpredictable (and non-spontaneous) nature of the materials. All sentences were originally read from text, and their content was semantically unpredictable. This may have inflated the relative importance of segmental errors for the listeners in the present study, because listeners could not use any contextual information in repairing any segmental errors. In a spontaneous conversation, where speech segments are less clearly pronounced but richer pragmatic context makes the message content far more predictable, listeners might rely less on segmental phonetic cues for speech intelligibility. Durational patterns might then be relatively more important for intelligibility.

Tajima et al. (1997) have suggested that non-native speakers might benefit from training programs which focus on various temporal aspects of L2 speech. The present findings support this recommendation to some extent, although the L2 learners should obviously also attempt to reduce the segmental errors that contribute to their foreign accent in L2 pronunciation. Again, the communicative importance of durational patterns might be larger in spontaneous conversation.

One drawback of the present study is the relative inefficiency of the SRT method. For each participant, all responses to a list of 13 sentences are combined into a single SRT value. Thus each participant produced only 8 data points. With only 10 lists available (Plomp and Mimpen, 1979), no more than 10 data points per participant can be acquired, which limits the power

of the SRT method. In addition, the SNR step size ( $\pm 2$  dB) is an uneasy compromise. Smaller steps would require even longer lists of sentences (and hence even less efficiency), but they would yield smaller standard deviations in the outcomes. Larger steps would require shorter lists (so the sentence set could in theory be recombined into more lists), but the resulting thresholds may be useless in practice.

A second drawback is that results of the present study may not generalize to other L2 speakers. As mentioned in the Introduction above, our findings may nevertheless be relevant, because the present L2 speaker exhibits a typical combination of segmental and durational errors. The relative contributions of segmental and durational errors to intelligibility may vary between L2 speakers (and their native languages), just like they vary within a single speaker as reported above.

In our manipulations, the superficial segmental durations of the L1 speech were transplanted as a whole into the L2 speech. That is, no attempt was made here to capture the linguistic patterns underlying these durations (e.g. phonological vowel length, voicing effect, phrase-final lengthening, word boundary demarcation, rhythmic patterns, etc). Investigating the underlying linguistic patterns, rather than the resulting segmental durations, would be an obvious next step now that the SRT method has been shown to be useful for quantifying different components in L2 speech intelligibility.

In conclusion, the non-transplanted native and non-native speech differ by about 4 dB in their speech reception thresholds. Transplanting the non-native durational patterns onto the native speech segments decreases speech intelligibility by about 1.1 dB. Conversely, correcting the non-native speech with native-like durational patterns increases the intelligibility threshold by about 0.4 dB (or by about 1.0 dB for the non-native sentences that were du-

rationally most deviant). These durational effects are relatively small compared to segmental effects on intelligibility, but these findings suggest that targeted attention to linguistically important duration patterns (e.g. phonological vowel length, stress and accent, etc.) may well be worthwhile when learning to speak a foreign language.

### **Acknowledgements**

This research was conducted as part of the MSc thesis requirements of the second author. We are very grateful to Sander van Wijngaarden for providing the recordings of Dutch L2 speech, and to Joost Festen for providing the recordings of the matching Dutch L1 audiological test sentences. Our thanks are also due to Theo Veenker for technical assistance, and to Sieb Nootboom, Esther Janse, Liesbeth Mus, Ditta van Dijk and two anonymous reviewers for helpful comments and suggestions.

### **References**

- Adams, C., 1979. English speech rhythm and the foreign learner. Mouton, The Hague.
- Anderson-Hsieh, J., Johnson, R., Koehler, K., 1992. The relationship between native speaker judgement of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning* 42 (4), 529–555.
- Baayen, R. H., Davidson, D. J., Bates, D. M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59 (4), 390–412.

- Bates, D., 2005. Fitting linear models in R: Using the `lme4` package. *R News* 5 (1), 27–30.
- Bent, T., Bradlow, A. R., Smith, B. L., 2008. Production and perception of temporal patterns in native and non-native speech. *Phonetica* 65 (3), 131–147.
- Boersma, P., Weenink, D., 2008. Praat: Doing phonetics by computer. Version 5.0.43. Available: <http://www.praat.org>.
- Cambier-Langeveld, G., 2000. Temporal marking of accents and boundaries. Ph.D. thesis, University of Amsterdam. Available: <http://dare.uva.nl/en/record/86264>.
- Chun, D., 2002. Discourse intonation in L2: From theory and research to practice. Benjamins, Amsterdam.
- Cutler, A., Dahan, D., van Donselaar, W., 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40, 141–201.
- Derwing, T. M., Munro, M., 1997. Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition* 20, 1–16.
- Eefting, W., Nootboom, S., 1993. Accentuation, information value and word duration: effects on speech production, naturalness and sentence processing. In: Van Heuven, V., Pols, L. C. (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech generation*. *Speech Research*; 11. Mouton de Gruyter, Berlin, pp. 225–240.

- Faraway, J. J., 2006. *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall, Boca Raton, FL.
- Flege, J., 1993. Production and perception of a novel, second language phonetic contrast. *Journal of the Acoustical Society of America* 93 (3), 1589–1608.
- Flege, J., Port, R., 1981. Cross-language phonetic interference: Arabic to English. *Language and Speech* 24 (2), 125–146.
- Goetry, V., Kolinsky, R., 2000. The role of rhythmic cues for speech segmentation in monolingual and bilingual listeners. *Psychologica Belgica* 40 (3), 115–152.
- Holm, S., 2008. *Intonational and durational contributions to the perception of foreign-accented Norwegian: An experimental phonetic investigation*. Ph.D. thesis, Norwegian University of Science and Technology. Available: [http://www.hf.ntnu.no/hf/isk/Ansatte/snefrid.holm/PhD\\_thesis.Snefrid\\_Holm.pdf](http://www.hf.ntnu.no/hf/isk/Ansatte/snefrid.holm/PhD_thesis.Snefrid_Holm.pdf).
- Hox, J., 2002. *Multilevel Analysis: Techniques and Applications*. Lawrence Erlbaum Assoc, Mahwah, NJ.
- Klatt, D., 1976. Linguistic uses of segmental durations in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59, 1208–1221.
- Laeuffer, C., 1992. Patterns of voicing-conditioned vowel duration in French and English. *Journal of Phonetics* 20 (4), 411–440.

- Maassen, B., Povel, D.-J., 1984. The effect of correcting temporal structure on the intelligibility of deaf speech. *Speech Communication* 3 (2), 123–133.
- Mareüil, P., Vieru-Dimulescu, B., 2006. The contribution of prosody to the perception of foreign accent. *Phonetica* 63 (4), 247–267.
- Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453–467.
- Munro, M., 2008. Foreign accent and speech intelligibility. In: Hansen Edwards, J., Zampini, M. (Eds.), *Phonology and Second Language Acquisition*. Benjamins, Amsterdam, pp. 193–218.
- Munro, M., Derwing, T. M., 1999. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 49 (S1), 285–310.
- Nooteboom, S., 1997. The prosody of speech: Melody and rhythm. In: Hardcastle, W., Laver, J. (Eds.), *The Handbook of Phonetic Sciences*. Blackwell, Oxford, pp. 640–673.
- Patel, A. D., 2008. *Music, Language, and the Brain*. Oxford University Press, Oxford.
- Peterson, G. E., Lehiste, I., 1960. Duration of syllable nuclei in English. *J. Acoustical Society of America* 32 (6), 693–703.
- Pinheiro, J. C., Bates, D. M., 2000. *Mixed-effects models in S and S-Plus*. Statistics and Computing. Springer, New York.

- Plomp, R., 1986. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech and Hearing Research* 29 (2), 146–154.
- Plomp, R., Mimpen, A. M., 1979. Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology* 18 (1), 43–52.
- Quené, H., 1992. Durational cues for word segmentation in Dutch. *Journal of Phonetics* 20 (3), 331–350.
- Quené, H., Port, R., 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* 62 (1), 1–13.
- Quené, H., Van den Bergh, H., 2004. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication* 43 (1–2), 103–121.
- Quené, H., Van den Bergh, H., 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59 (4), 413–425.
- R Development Core Team, 2008. R: A language and environment for statistical computing. Version 2.8.1. Available: <http://www.r-project.org>.
- Rajadurai, J., 2007. Intelligibility studies: a consideration of empirical and ideological issues. *World Englishes* 26 (1), 87–98.
- Shatzman, K., McQueen, J. M., 2006. Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics* 68 (1), 1–16.

- Slis, I., Cohen, A., 1969. On the complex regulating the voiced–voiceless distinction, I and II. *Language and Speech* 12, 80–102 and 137–155.
- Sluifjter, A. M. C., Van Heuven, V. J., 1995. Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica* 52 (2), 71–89.
- Smith, R., 2004. The role of fine phonetic detail in word segmentation. PhD thesis, University of Cambridge.
- Studebaker, G., 1985. A “rationalized” arcsine transform. *Journal of Speech and Hearing Research* 28 (3), 455–462.
- Tajima, K., Port, R., Dalby, J., 1997. Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics* 25 (1), 1–24.
- Van Santen, J., 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language* 8 (2), 95–128.
- Van Wijngaarden, S. J., 2001. Intelligibility of native and non-native Dutch speech. *Speech Communication* 35 (1-2), 103–113.
- Van Wijngaarden, S. J., Steeneken, H. J. M., Houtgast, T., 2002. Quantifying the intelligibility of speech in noise for non-native talkers. *The Journal of the Acoustical Society of America* 112 (6), 3004–3013.
- White, L., Mattys, S., 2007a. Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35 (4), 501–522.
- White, L., Mattys, S. L., 2007b. Rhythmic typology and variation in first and second languages. In: Prieto, P., Mascaró, J., Solé, M.-J. (Eds.), *Segmental and Prosodic issues in Romance Phonology. Current Issues in Linguistic Theory*. John Benjamins, Amsterdam, pp. 237–257.

Figure 2: Average spectra of masking noise, for the native (S1) and non-native speaker (S2).

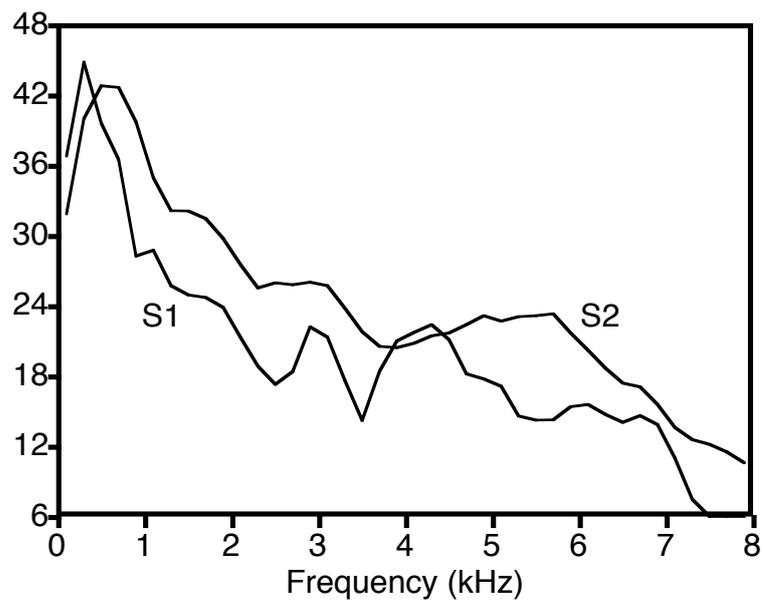


Figure 3: Average Speech Reception Thresholds, expressed as speech-to-noise ratios (in dB), broken down by native speaker (S1: native, S2: non-native) and by durational patterns (D1: native, D2: non-native). Error bars indicate standard errors.

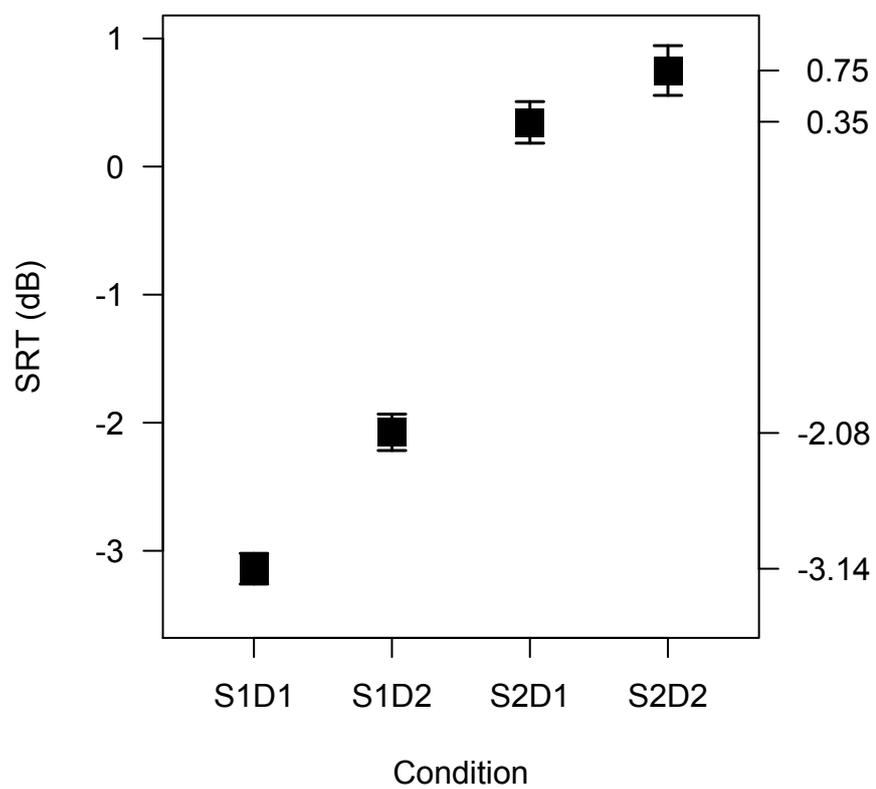


Figure 4: Scattergram of average durational deviance (horizontal axis) and average effect of duration manipulation using S2 speech (vertical axis). Each point represents a list of 13 stimulus sentences, see text.

