Long-term convergence of speech rhythm in L1 and L2 English

Hugo Quené[1], *Rosemary Orr*[1 2]

[1]Utrecht inst of Linguistics OTS, Utrecht University, Utrecht, the Netherlands
[2]University College Utrecht, Utrecht, the Netherlands
h.quene@uu.nl, r.orr@uu.nl

Abstract

When talkers from various language backgrounds use L2 English as a lingua franca, their accents of English are expected to converge, and talkers' rhythmical patterns are predicted to converge too. Prosodic convergence was studied among talkers who lived in a community where L2 English is used predominantly. Speech rhythm was operationalized here as the peak frequency in the spectrum of the intensity envelope, normalized to the speaking rate (in syll/s). Results indicate that talkers produced intensity contours with maximum periodicity at frequencies of about 0.32 times their syllable rates, i.e., peaks in intensity tend to occur every 1/0.32 syllables. These results were collected repeatedly, from 5 recordings conducted over 3 years with the same talkers. We found that variance between talkers in their rhythm decreases over time, thus confirming the predicted convergence in speech rhythm in L2 English. These findings show that speech rhythm in L2 English tends to converge, and that this prosodic convergence continues to proceed over several years, as well as over communicative settings.
Index Terms: speech rhythm; phonetic convergence; accommodation; L2;

1. Introduction

When talkers from different varieties or dialects of a language converse with each other, their dialects and accents tend to converge. This convergence has been reported on various time scales for varieties of L1 Dutch [11], L1 English [7, 5] and L1 French [4]. But does convergence also occur among varieties of an L2 language used as a lingua franca by talkers from various L1 backgrounds? It is widely assumed that a talker's accent will weaken or disappear in intensive contact with other talkers of L1 and L2 English. However, we have only limited insight in how and why L2 accents change over time in highly interactive environments. Do non-native speakers become more native-like, and does interference from their L1 decrease over time? And do native speakers also drift away from their native pronunciation standards? The international community of the University College Utrecht (UCU) provides an excellent environment to study phonetic convergence in L2: its students vary in their native languages (with Dutch being the majority L1), English is spoken as the lingua franca on campus, and the UCU is located in a Dutch-speaking country, so that any changes in English accents can be validly attributed to internal processes within the community.

The core hypothesis in this longitudinal research project is that the native and non-native accents of UCU students will gradually converge to a single common UCU English accent, in which properties of L1 British English, L1 American English, L2 Dutch English, and other varieties will be mixed. With regard to speech rhythm, we hypothesize that talkers' rhythmical patterns will converge. A key property of English is its lexical stress, which typically results in one stressed syllable of a word being more prominent (produced with more effort and longer duration) than the other unstressed ones. A salient rhythmical feature of English is that the unstressed syllables tend to be reduced more dramatically (both spectrally and temporally) than they are in other lexical-stress languages such as Dutch or German, or in non-stress languages such as Japanese.

In the present longitudinal study, we focus on long-term convergence in speech rhythm in L1 and L2 English, over a period of 3 years. Do L2 English speakers become more native-like, in that they will show more reduction of unstressed syllables? Or do L1 English speakers become less native-like, in that they will show less reduction of unstressed syllables? Does the variance between talkers in their preferred rhythmical pattern decrease over the years? These questions will be answered by means of the longitudinal speech corpus described below.

2. Corpus

The Longitudinal University College English Accents Corpus (LUCEA) is a corpus of speech recordings, collected at University College Utrecht (UCU) in Utrecht, the Netherlands. Four consecutive annual cohorts of students were or will be recorded, each at five "rounds" or time points over the three years of their undergraduate study, in longitudinal fashion. Talkers are recorded in Sept of year 1 of their programme (month 1), May of year 1 (month 9), Sept of year 2 (month 13), May of year 2 (month 21), and May of year 3 (month 33; most UCU students are absent on exchange visits in the first semester of year 3). Actual recording dates may differ by a few weeks from the nominal month of recording, due to students' availability. Recordings began in September 2010, and are still in progress.

2.1. Talkers

The talkers in the corpus are mostly full-time degree students, with some incoming exchange students and staff included who are native speakers of English. About 60% of the degree students and of the talkers in the corpus have Dutch as their L1. In addition to the Dutch speakers and the native speakers of English, over 30 other native languages are represented in the corpus. UCU students are required to use English in all their interaction with tutors and teachers. Students live on campus during their entire three-year degree programme, and in culturally and linguistically mixed social settings, English is also the language of choice. Thus the talkers use L1 or L2 English very intensively for interacting in the campus community. Of the student population of about 700, approximately one quarter (about 60 per cohort) participates as a talker contributing speech data to the LUCEA corpus.

The present study targets those talkers for which a full sequence of 5 recordings is available ($n = 18$); all these talkers were students of the first cohort (class of 2013) participating in the LUCEA corpus collection. In the first and last recording sessions, talkers also filled in a questionnaire about their language background, language use, study exchange experience, etc. In the entry questionnaire, 15 talkers declared themselves as native speakers of Dutch, and 1 talker each as a native speaker of Russian, Vietnamese, and German. In the exit questionnaire, 3 talkers (1 female, 2 male) also regarded themselves as L1 English speakers. These answers need not be contradictory, as many students at UCU have a complex language history with multiple native languages (e.g. different languages of mother, father, and country of residence).

2.2. Procedure

Each student talker receives €10 for each recording, with a bonus of €10 if all 5 rounds of recordings are completed. Recordings take place in a quiet office room on the UCU campus. The talker's speech is recorded by means of 7 microphones (Sennheiser ME64/K6p) placed around, behind, and above the talker, and also via a close-talking microphone (Sennheiser Headset HSP 2ew). All 8 channels are recorded digitally by means of a Saffire Pro 40 multichannel preamp and AD converter (at 44.1 kHz, 16 bits).

Talkers are asked to perform between 10 and 12 speech tasks, listed in Table 1, including texts to read aloud, monologues in both the native language and English (if English is not the native language), and a dialogue with the facilitator. Each recording session takes about 45 minutes, in which approximately 25 minutes of speech is collected.

Table 1: *Summary of tasks for talkers to be performed in each recording.*

Nr	Description
1	Say your name, and today's date and time
2	Short extract from the Rainbow Passage
3	*Please Call Stella* [13]
4	*The Boy who Cried 'Wolf'*†
5	Sentence sets for intelligibility testing
6	Five sentences for investigating prosody [14]
7	Extract from Declaration of Human Rights in L1 [12]
8	Extract from Declaration of Human Rights in English [12]
9	2 minute monologue in L1 (informal topic)
10	2 minute monologue in English (informal topic)
11	2 minute monologue in English (formal topic)
12	3 minute dialogue with facilitator

†Two different versions of this text have been used.

3. Method

The present study focused on a small part of the recordings, viz. the 5 English sentences taken from and studied by [14] (Table 1, task 6; see Appendix A). These sentences may be regarded as a semi-random sample in their distribution of stressed and unstressed syllables. Jointly the 5 sentences contain 26 stressed syllables out of a total of 81 syllables, a ratio of 0.321.

Talkers were always instructed to read each sentence without pausing or inhaling in mid-sentence, and without any errors. If necessary, a talker repeated a sentence until this was achieved. Only fluent realizations without error and without pausing were selected for subsequent analysis. Each sentence was excised and stored as a separate audio file. Incomplete recordings (due to errors in post-processing) were discarded, with $N = 423$ recordings by 18 talkers remaining.

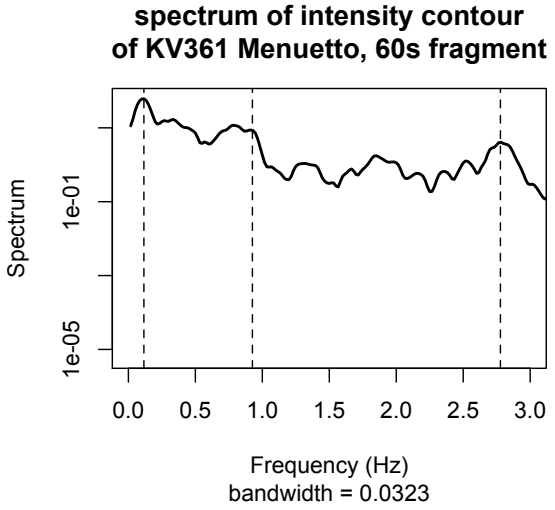

Figure 1: *Spectrum of intensity contour, with marking of salient frequencies.*

The intensity envelope of each sentence was measured by means of Praat [2], using a 25 ms window shift (40 Hz sampling frequency). The resulting intensity contour (in dB units) was exported. Subsequent analysis of the intensity envelope was inspired by [6] and was performed using `R` [10]. The intensity contour was converted to a spectrum, in order to bring out its periodic components. As an example, consider the spectrum of the intensity envelope of a musical fragment (by W.A. Mozart, KV361, Menuetto, in $3/4$ time) shown in Fig. 1. The peak at 2.77 Hz corresponds with the quarter notes (at a tempo of about 167 beats per minute, or 2.77 per s), and the peak at 0.92 Hz corresponds with the full measures (0.92 per s). The strongest peak at 0.116 Hz corresponds with an 8-measure or 24-beat periodicity of the intensity envelope.

From each spectrum, we then identified the frequency bin with the maximum intensity, i.e., the strongest periodic component of the intensity envelope. The windows used during preceding intensity analysis and spectral analysis and smoothing resulted in an eventual frequency resolution of 0.016 Hz (spacing between frequency bins).

The duration of each sentence was assessed from the duration of the intensity envelope, excluding silent intervals before and after the sentence. The articulation rate (sc. tempo, in syll/s) of each spoken sentence was computed using the syllable counts in Appendix A. (Note that sentences were always spoken without pauses.) The observed peak frequency in the spectrum of the intensity envelope was then normalized to this articulation rate. Thus the strong spectral component that corresponds to the syllable repetition frequency is removed, and the resulting measure represents the periodicity of the intensity envelope, expressed *relative* to the syllable-based periodicity of the intensity envelope. For the spectrum of Fig. 1 (with a base tempo of 2.77 Hz) this would yield a single dimensionless value of $0.116/2.77 = 1/24$, indicating a periodic or cyclical pattern in the intensity envelope that spans over 24 quarter-notes. The

relative strength of this normalized peak frequency was not assessed. Data from 2 sentence recordings were excluded from further analysis because of their unusually high articulation rate (> 8 syll/s), with $N = 421$ spoken sentences remaining.

The normalized peak frequencies of the intensity envelope of each spoken sentence were fed into a linear mixed-effects regression model (LMM) [8, 9], with talkers ($n = 18$) and sentences ($n = 5$) as two crossed random effects, using maximum likelihood estimation. LMM was done using package `lme4` [1] in `R` [10]. Fixed predictors were (i) the "round" or time of recording (coded as 4 contrasts between consecutive rounds), (ii) the talker's status as a native speaker of English (0=no, 1=yes), (iii) the interaction of predictors (i) and (ii), and (iv) the articulation rate (in syll/s, centered to its median). The rounds were also included in the random part at the talker level, i.e., we explicitly modeled the between-talker variance for each round separately.

4. Results and discussion

The coefficients and variances estimated by the LMM described above are listed in Table 2. The estimated normalized peak frequencies in the spectrum of the intensity envelope are also illustrated in Fig. 2.

Table 2: *Estimated coefficients of the LMM. For fixed effects,* r2r1 *refers to the contrast between round 2 and round 1, etc.;* native *indicates the talker's status as native speaker of English (yes=1); colons are used for interaction terms; asterisks mark fixed effects with $p < .05$ (based on 2.5% and 97.5% percentiles of bootstrapped estimates over 200 bootstrap replications). For random effects,* u *refers to talkers and* v *to sentences, and* e *to residual error; the 95% confidence intervals are based on percentiles of bootstrapped estimates over 200 replications.*

Fixed effects:	Estimate	Std.Error	t value
(Intercept)	0.311	0.036	8.674
tempo	−0.039	0.015	−2.609 *
r2r1	0.018	0.051	0.035
r3r2	0.032	0.059	0.548
r4r3	0.032	0.047	0.669
r5r4	0.037	0.036	1.025
native	0.082	0.029	2.822 *
r2r1:native	−0.004	0.122	−0.029
r3r2:native	−0.125	0.140	−0.892
r4r3:native	−0.195	0.112	−1.739 *
r5r4:native	−0.167	0.086	−1.939 *
Random effects:	Estimate	95%C.I.	
σ^2_{u1}	0.0082	(0.0020, 0.0167)	($n = 18$)
σ^2_{u2}	0.0048	(0.0009, 0.0107)	
σ^2_{u3}	0.0024	(0.0003, 0.0070)	
σ^2_{u4}	0.0024	(0.0005, 0.0076)	
σ^2_{u5}	0.0015	(0.0003, 0.0058)	
σ^2_{v}	0.0057	(0.0002, 0.0136)	($n = 5$)
σ^2_{e}	0.0165	(0.0137, 0.0182)	($n = 421$)

In the present sample of spoken English sentences, the intensity envelope shows a major periodicity at $0.31\times$ the syllable rate. This value matches with the proportion of stressed syllables (0.321) in the test sentences, suggesting that the peak frequency in the intensity envelope, or "normalized stress rate", corresponds with the rhythm of the spoken sentences. In English, unstressed syllables tend to be weaker in intensity (and

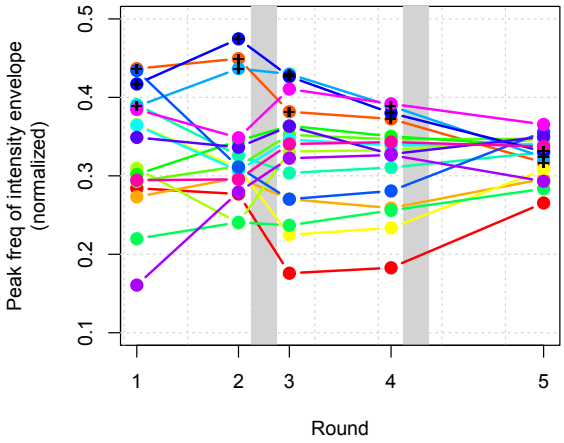

Figure 2: *Estimates of normalized peak frequencies in the spectrum of intensity envelope, broken down by round of recording (along abscissa, on approximate time scale) and by talker (with plussed symbols representing L1 English speakers). Shaded areas represent 2-month summer breaks during which talkers do not live on the UCU campus.*

shorter in duration) than stressed syllables, and this property is indeed captured by the intensity envelope showing periodicity at a rate corresponding with the average inter-stress interval. Thus this stress rate [6], here normalized relative to syllable rate, may be of interest for evaluating speech rhythm.

Indeed, the normalized "stress rate" or peak frequency of the intensity envelope is significantly higher for the 3 native English talkers than for the 15 nonnative talkers in the present sample, as shown in Table 2. In our interpretation, this significant difference may be ascribed to the stronger reduction of unstressed syllables in English as compared to Dutch. This difference in reduction has been reported to extend to L1 English and L2 English spoken by Dutch native speakers [14, 3]. In our sample, 15 of the 18 talkers mentioned Dutch as one of their native languages. If these L2 talkers would reduce their unstressed vowels in their L2 English to a lesser degree than their native counterparts, then the intensity peaks of their stressed and unstressed syllables would also differ to a lesser degree than their native counterparts would produce. This would in turn result in a somewhat *lower* "normalized stress rate" for L2 talkers as compared to what native speakers would produce. For example, the native English speakers all pronounce *chairman* as [ˈtʃɛəmən], whereas the native Dutch speakers of L2 English tend to pronounce this word as [ˈtʃɛəˌmæn] with a less reduced second syllable, yielding a higher intensity peak for the unstressed syllable, and hence a lower "normalized stress rate". Thus the significant main effect of native-speaker status on normalized stress rate reflects this relatively subtle prosodic difference between native and Dutch-accented English in degree of syllable reduction [14, 3].

The significant interaction between native-status and recording round indicates that the 15 nonnative talkers do *not* change over time in their normalized stress rate, whereas the 3 native English talkers tend to converge to the somewhat lower, nonnative stress rate observed for the nonnatives. In other

words, the native speakers of English (who are a minority at UCU) tend to adapt their English speech rhythm to that of their nonnative peers (who are a majority), especially at the 4th and 5th recording session (end of year 2 and end of year 3, respectively). The convergence seems to extend into the students' third year on campus; this finding matches similar reports of long-term phonetic accommodation [11, 5]. Moreover, the phonetic accommodation is extended from conversational settings with peer students, and from classroom situations, to the interview setting of the corpus recordings. These findings supports the core hypothesis of long-term prosodic convergence among students in the UCU community.

The core hypothesis of phonetic convergence also predicts that the variance between talkers decreases over time. LMM allows this prediction to be tested, by modeling separate $\sigma^2_{u.}$ between-talker variances for each round of recordings. The resulting variance estimates in Table 2 seem to confirm this prediction to some extent (cf. Fig. 2), but the 95% confidence intervals of these variance estimates do overlap in bootstrap validation. The LMM reported in Table 2 also does not perform significantly better than a simpler model in which between-talker variances are pooled into a single estimate for all 5 rounds (i.e. in which homoskedasticity is assumed; likelihood ratio test, $\chi^2 = 7.62, \mathrm{df} = 14$, n.s.). Thus the decreasing variance between talkers is adequately captured by the significant interaction between native status and recording round (in the fixed part of the LMM, [9]), and there is no significant deviation from homoskedasticity beyond this interaction.

In conclusion, L2 English talkers do not show longitudinal changes in their rhythmical pattern. The native L1 English talkers however tend to move away from their native rhythmical patterns (observed initially), by decreasing the degree of reduction of unstressed syllables; hence they accommodate to the predominant variety of L2 English in the language community. These longitudinal changes confirm that members of this multilingual community, where English is used as the lingua franca, do converge in their speech rhythm of their L1 and L2 English accents.

5. Acknowledgements

We thank the UCU talkers for lending their voices, Roeland van Beek, Thari Diefenbach, Anne van Leeuwen, Lisa Teunissen, Kate Backhouse, Maria Koutiva and Kim Cruden for their assistance in conducting the recording sessions, and David van Leeuwen for technical assistance.

6. References

[1] Bates, D., Maechler, M., Bolker, B. and Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5, 2013. Online: http://CRAN.R-project.org/package=lme4

[2] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer, version 5.3.51, 2013. Online: http://www.praat.org

[3] Braun, B., Lemhöfer, K. and Mani, N. Perceiving unstressed vowels in foreign-accented English, J. Acoust. Soc. Am., 129(1):376–387, 2011.

[4] Delvaux, V., and Soquet, A. "The influence of ambient speech on adult speech productions through unintentional imitation", Phonetica, 64(2–3):145–173, 2007.

[5] Evans, B.G. and Iverson, P., "Plasticity in vowel perception and production: A study of accent change in young adults", J. Acoust. Soc. Am., 121(6):3814–3826, 2007.

[6] Liberman, M.Y., "Speech rhythms and brain rhythms", Language Log, 2 Dec 2013.
Online: http://languagelog.ldc.upenn.edu/nll/?p=8116

[7] Pardo, J.S., "On phonetic convergence during conversational interaction", J. Acoust. Soc. Am., 119(4):2382–2393, 2006.

[8] Quené, H. and van den Bergh, H., "On Multi-Level Modeling of data from repeated measures designs: A tutorial", Speech Comm., 43(1–2):103–121, 2004.

[9] Quené, H. and Van den Bergh, H., "Examples of mixed-effects modeling with crossed random effects and with binomial data", J. Memory and Language, 59(4):413–425, 2008.

[10] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, version 3.0.2, 2013. Online: http://www.R-project.org/.

[11] Scholtmeijer, H. *Het Nederlands van de IJsselmeerpolders*. Kampen, 1992.

[12] Van Engen, K. J., M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow. "The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles." Language & Speech, 53(4):510–540, 2010.

[13] Weinberger, S. Speech Accent Archive. George Mason University, 2013. Online: http://accent.gmu.edu

[14] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", J. Phonetics, 35(4):501-522, 2007.

A. Test sentences

Each test sentence (from [14]) is followed by its numbers of stressed and unstressed syllables (in typical readings in the present corpus) and its total number of syllables.

1 The supermarket chain shut down because of poor management (5, 10, 15). 2 Much more money must be donated to make this department succeed (6, 11, 17). 3 In this famous coffee shop they serve the best doughnuts in town (5, 10, 15). 4 The chairman decided to pave over the shopping center garden (5, 12, 17). 5 The standards committee met this afternoon in an open meeting (5, 12, 17).