INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speaking rate, information density, and information rate in first-language and second-language speech

Ann R. Bradlow

Department of Linguistics, Northwestern University, Evanston, IL, USA
abradlow@northwestern.edu

Abstract

Using a corpus of multilingual recordings of a standard text (the North Wind and the Sun passage, NWS) in 11 languages, speaking rate (SR, syllables/second) and information density (ID, number of syllables for the NWS text) were examined in first-language (L1) and second-language (L2) speech. Replicating prior work, cross-language comparison of L1 speech showed a trade-off between SR and ID such that relatively low density languages (many syllables for the NWS text) tended to be produced at relatively fast rates, and vice versa. Furthermore, L2 English was characterized by both slower rate and lower ID than L1 English. That is, L2 English involved more syllables than L1 English for the same NWS text. A comparison of the number of acoustic syllables (i.e. amplitude peaks) with the number of orthographic syllables (i.e. dictionary-based syllable counts for the NWS text) indicated that L1 speech involved substantial syllable reduction (fewer acoustic than orthographic syllables) while L2 speech involved substantial syllable epenthesis (more acoustic than orthographic syllables). These findings suggest that L2 speech production involves temporal restructuring beyond increased segment, syllable and word durations, and that the resultant information rate (information bits transmitted/second) of L2 speech diverges substantially from that of L1 speech.

Index Terms: speech production, speaking rate, information density, foreign-accented speech, cross-language variation

1. Introduction

It is well-established that second-language (L2) speech is produced at consistently slower rates than first-language (L1) speech [1-3]. This striking difference in speaking rate (SR) between L2 and L1 speech is evident in cross-talker comparisons of L1 and L2 speech of a given language (e.g. L1 English versus L2 English), as well as in cross-language comparisons of L1 and L2 speech within bilinguals. Moreover, this slower rate of L2 versus L1 speech is evident in both read speech [1] and spontaneous speech [3], and there is also evidence that L2 SR may be more variable than L1 SR [2]. Findings such as these establish that L1 versus L2 "state" involves a robust macro-level articulatory setting that controls the overall temporal structure of an utterance and within which timing patterns at the segmental, syllabic, lexical, and phrasal levels are set.

Prior work has also demonstrated that, nested within the temporal variation induced by L1 versus L2 state is another important SR control variable, namely individual talker "trait." Using a corpus of L1 and L2 speech recordings from a large set of bilingual speakers (n=86) from various L1 backgrounds (n=10), prior work demonstrated that within a group of bilinguals, variation in L1 SR was a significant predictor of L2 SR [4]. Talkers who were relatively slow or fast in the L2 (English) tended to also be relatively slow or fast, respectively, in their L1. This finding indicates that a language-independent talker-specific speech articulation trait characteristic combines with the L1 versus L2 state characteristic to determine overall SR in the speech of bilingual individuals.

Finally, prior work has demonstrated cross-language variation in L1 SR, suggesting that language-specific structure also plays a significant role in determining SR. Specifically, a group-wise average of 5.2 to 7.8 syllables per second was observed across a matched set (i.e. direct translations) of sentence productions by 59 native speakers of seven typologically distinct languages [5]. Taking an information theoretic approach [6], it was hypothesized that the observed cross-language variation in L1 SR may be accounted for by a trade-off between information density (ID) and SR where ID is calculated based on the number of speech units (syllables) required to convey a given meaning. According to this trade-off, high density languages (i.e. languages that require relatively few syllables to convey a given meaning) should be produced at relatively slow SRs (few syllables per second) in comparison to relatively low density languages. In principle, this trade-off should result in reduced cross-language variation in terms of information rate (IR), i.e. languages should be fairly consistent in the amount of information transmitted per unit of time. The cross-language comparison [5] showed the expected trade-off between ID and SR. However, the negative correlation between these two variables was not strong enough to result in a constant IR across the languages. Thus, while syllable-based ID can account for some cross-language variation in SR, the linguistic encoding of information at other levels of linguistic structure (morpho-phonology and syntax) is also important for regulating the IR of spoken utterances.

Taken together, these cross-language and cross-talker studies of SR variation establish three confluent sources of influence on the overall temporal structure of speech:

(a) L1 versus L2 state-specificity: L2 speech is consistently slower than L1 speech,

(b) talker-specificity: within bilingual individuals, L1 SR significantly predicts L2 SR, and

(c) language-specificity: SR varies cross-linguistically with an inverse relation to ID.

The present study aims to gain further insight into sources of converging and diverging variation in L1 and L2 SR. In particular, this study investigates the impact of language-specificity and of L1 versus L2 state-specificity ((a) and (c) above) on the relation between SR and ID.

One possibility is that the reduced rate of L2 speech relative to L1 speech involves slower speech production without any effect on ID. Under this scenario, while the number of syllables

http://dx.doi.org/10.21437/Interspeech.2019-1150

per second differs across L1 and L2 speech, the total number of syllables produced for a given meaning/text remains constant. The slower syllable rate of L2 speech would therefore result in a decreased IR (i.e. the longer syllable durations would lead to fewer bits of information conveyed per second).

Alternatively, in addition to involving slower articulation than L1 speech (resulting in longer syllables), L2 speech may also involve substantial temporal restructuring of L2 speech relative to L1 speech resulting in a significant difference in the overall number of syllables produced for a given text (i.e. different ID). This would, in turn, influence the IR of L2 speech. For target languages that allow relatively complex syllable structures (such as, English) L2 speakers may adopt a cluster simplification strategy that involves vowel epenthesis to break up complex consonant clusters. For example, Spanish-accented English frequently involves vowel insertion before /s/+stop consonant clusters in word initial position ('*especial'* for '*special'*). This strategy would result in a greater number of total syllables for a given text/meaning thereby lowering the overall ID of Spanish-accented English. L2 speech may also exhibit fewer phonetic reductions than L1 speech at the phrase level (e.g. function word reduction), particularly for read speech where L2 speakers may adopt a reading strategy with full pronunciation of all (or most) orthographic syllables while L1 speakers may be more likely to elide syllables in prosodically weak positions. For instance, L1 English speakers may reduce the word "and" in phrases such as "salt and pepper" or "run and jump" to the point where these phrases are produced with one fewer acoustic syllables (salient acoustic peaks) than predicted based on the pronunciation norms as reflected in dictionary pronunciation guides for citation speech. In contrast, L2 English speakers may be more likely to produce these phrases with non-reduced "and." Both of these features of L2 speech – vowel insertion for cluster simplification and lack of function word reduction – would raise the number of acoustic syllables for a given text/meaning and concomitantly lower the ID of the utterance relative to the ID of L1 productions of the same text/meaning. This then raises the question of whether and how status-specificity (L1 versus L2 status) affects the trade-off between ID and SR, which may, in turn, influence information processing in communicative situations that involve L2 speech.

Accordingly, the aims of the present study aims are (i) to replicate the cross-language trade-off between ID and SR in L1 speech [5], and (ii) to extend this information theoretic analysis to L2 speech.

2. Method and materials

2.1. The NWS sub-corpus of the ALLSSTAR Corpus

The data for this study were based on recordings taken from the ALLSTAR Corpus [7]. The key feature of this corpus is that it includes recordings from a large group of sequential bilingual speakers (n>120, age range 19-41 years) from a wide range of native language backgrounds (n>20) all of whom provided both scripted and spontaneous speech recordings in both their L1 and L2 (in all cases, English, typically with intermediate proficiency). The scripted speech includes (i) simple sentences taken from a set of sentences used for multi-lingual audiometric hearing in noise testing [8], (ii) complex sentences extracted from two widely translated texts, the novella, *Le Petit Prince* [9] and articles of the Declaration of Human Rights [10], and (iii) a widely translated one-paragraph passage, The North Wind and the Sun [11]. The spontaneous speech recordings include narratives of picture stories and answers to a set of standard prompts (e.g. describe your home-town). Recordings are stored in a web-based archive developed in our laboratory [12]. All of the scripted speech recordings can be downloaded with time aligned textgrids. Transcription and forced-alignment of the spontaneous speech recordings is ongoing.

For the present study, a sub-corpus of the ALLSSTAR Corpus, the NWS sub-corpus, was constructed to include only recordings of the North Wind and the Sun passage (NWS) in those languages for which there are recordings by at least four speakers (see Table 1 below). These restrictions were imposed to ensure that we could adequately compare ID across languages based on speech by enough talkers to ensure some (albeit limited) generalizability and on a discourse with a fixed semantic content (i.e. message communicated) across all languages. The brevity and universally accessible meaning of the NWS fable make it particularly well-suited to cross-language comparison of ID.

Table 1: *Native languages, ages, and English proficiency scores (where available) for talkers in the NWS sub-corpus.*

Language	Number Talkers	Average Age (range)	Average Versant* Score (% available)
Cantonese	14	22 (19-27) yrs.	NA (0/14)
English	25	29 (18-26) yrs.	--
Hebrew	4	30 (29-31) yrs.	74 (3/4)
Hindi	5	24 (22-26) yrs.	79 (3/5)
Korean	11	26 (22-29) yrs.	59 (9/11)
Mandarin	14	23 (21-25) yrs.	59 (14/14)
Portuguese	5	28 (25-34) yrs.	61 (5/5)
Russian	5	25 (22-32) yrs.	72 (3/5)
Spanish	11	27 (22-33) yrs.	68 (6/11)
Turkish	13	24 (21-27) yrs.	70 (12/13)
Vietnamese	4	24 (20-26) yrs.	56 (4/4)

* Versant [13] scores are in the B1-C2 range of The Common European Framework of Reference for Languages (CEFR).

2.2. Speaking rate (SR)

The present analyses were based on a measure of SR calculated as the total number of syllables produced in the NWS recording divided by the duration of the recording, excluding non-speech disfluencies (e.g., coughs) and silent pauses of at least 100 milliseconds in duration. The number of syllables was obtained using an automatic syllable counting algorithm implemented as a Praat script that counts peaks in intensity that are preceded and followed by dips in intensity excluding peaks that are not voiced [14]. This measure of SR thus counts the number of acoustic (as opposed to orthographic/phonological) syllables per second of speech. As such, it is a consistent and objective measure of temporal modulation of the speech signal without regard for language-particular phonetics, phonotactics, or phonology.

2.3. Information Density (ID)

Following [5], we calculated ID for each NWS recording by each speaker against a standard benchmark. The reasoning behind this approach is that the overall semantic content of the NWS reading passage (i.e. the message communicated by the passage) is constant across all languages, as well as across L1

and L2 speech. Thus, we can calculate the average quantity of information, I, per syllable for a given NWS recording, k, as (1).

$$I^k = S^k / \sigma^k \quad (1)$$

where S^k is the language-independent semantic content of the recording (i.e. the meaning of the NWS fable) and σ^k is the number of acoustic syllables in the recording.

We can then compute normalized ID for a given language, L, using English as the benchmark:

$$ID^k_L = I^k_L / I^k_{ENG}$$
$$= S^k / \sigma^k_L \times \sigma^k_{ENG} / S^k$$
$$= \sigma^k_{ENG} / \sigma^k_L \quad (2)$$

σ^k_{ENG} is calculated as the average number of acoustic syllables across all L1 English recordings of the NWS passage (i.e. over all 25 monolingual English talkers included in the NWS sub-corpus). ID values less than one indicate lower information density – speech that expresses the NWS meaning in a greater number of acoustic syllables -- than the English benchmark. Conversely, ID values greater than one indicate higher information density than the English benchmark (same meaning conveyed in fewer acoustic syllables).

3. Results

3.1. Cross-language comparison of information density (ID) and speaking rate (SR) in L1 speech

Figure 1 shows the relationship between L1 SR and L1 ID averaged across talkers within each of the 11 language groups in this study. Error bars show the standard error of the mean for each language group in each dimension. As explained above, all ID values are calculated with respect to the English benchmark. Accordingly, the English group mean ID is equal to 1. These group means are also shown in Table 2. In Figure 1 and Table 2 (and in all subsequent analyses involving L1 ID), one outlier from the L1 Cantonese group with ID far greater than the mean plus twice the standard deviation was removed.

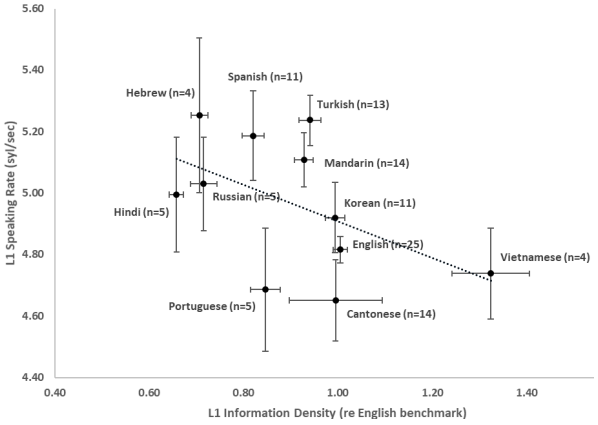

Figure 1: *L1 SR as a function of L1 ID for each language group. Error bars show the standard error of the mean in each dimension.*

As shown in Figure 1, the data exhibit substantial within-group variation in both SR and L1 ID. While the variation in SR is not surprising (some speakers are inherently faster in their L1 speech than others), the within-group variation in ID suggests that even when reading from a standard text, speakers vary in the number of acoustic syllables they produce. In other words, relative to the number of "orthographic" syllables (i.e. the number of expected syllables based on pronunciation norms in a dictionary), individual speakers delete and/or insert syllables even for a fixed reading passage [see also 15-16].

Critically for the present study, the group-wise data in the NWS sub-corpus showed a strong trend towards a trade-off in ID and SR such that languages with relatively high average ID values were spoken at relatively slow average SRs, and conversely those languages with relatively low average ID values were spoken at relatively fast average SRs (Spearman rho = -0.56, p=.08). This trade-off between ID and SR replicates the result first reported in [5] with a different set of languages and speech recordings. Note that the data in [5] were based on a more extensive set of recordings than the present study: 20 sets of 5 sentences in each language for a total of 100 sentences in each language by 6-10 native speakers of 7 languages (English, French, German, Italian, Japanese, Mandarin, and Spanish).

Table 2: *Average L1 and L2 information density (ID wrt English benchmark) and speaking rate (SR, acoustic syllables/second) for all L1 groups.*

	L1 Speech		L2 Speech (English)	
Language Group	**ID (re Eng)**	**SR (syl/sec)**	**ID (re Eng)**	**SR (syl/sec)**
Cantonese	1.00*	4.65	0.84	3.89
English	1.00	4.82	--	--
Hebrew	0.71	5.25	0.72	4.45
Hindi	0.66	4.99	0.86	4.53
Korean	0.99	4.92	0.91	3.97
Mandarin	0.93	5.11	0.85	4.26
Portuguese	0.85	4.69	0.79	4.05
Russian	0.72	5.03	0.83	4.71
Spanish	0.82	5.19	0.86*	4.23
Turkish	0.94	5.24	0.97	4.21
Vietnamese	1.32	4.74	0.90	4.01

* = one outlier excluded

3.2. Information density (ID) in L2 speech

In order to extend the information-driven analysis to L2 speech, we calculated L2 English ID for all of the bilingual speakers (n=86) using the same English benchmark as in the cross-language L1 ID analysis presented above. We observed substantial variation across language groups for average L2 ID, ranging from 0.72 to 0.97 (see Table 2). Despite this variation in L2 ID, L2 English was consistently produced with lower ID than L1 English, indicating that the L2 English NWS passages were produced with more acoustic syllables relative to the L1 English production of the same NWS passage. Figure 2 (left panel) shows the significant difference between the average L2 ID across all L2 English speakers and the L1 English benchmark ($t(100)=-2.09$, $p<.04$). The L2 English ID averages for each L1 group are also shown (right panel). Note that the L2 English average excludes one outlier from the L1 Spanish group whose L2 English ID score was far greater than the mean plus twice the standard deviation.

In contrast to the ID-SR trade-off that we observed in L1 speech, ID was not correlated with SR in L2 speech (Spearman rho = -0.36, p=.31). Thus, while the ID-SR trade-off in L1 speech may reflect a language-general tendency towards a constant IR (bits of information transmitted/unit of time), the break-down of this link in L2 speech suggests less systematicity in L2 temporal structure. Within the L2 group, individual L2 utterances may involve any combination of relative ID and SR resulting in variable IR across the group of L2 English speakers.

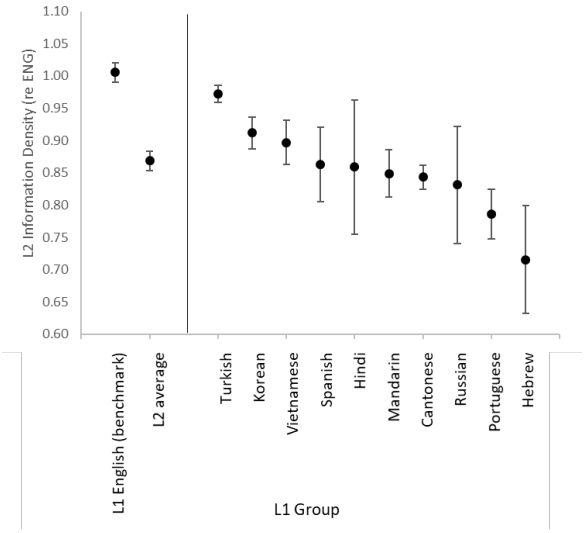

Figure 2: L2 *information density (ID) as compared to the L1 English benchmark. L2 ID average across all 10 language groups and for each L1 group are shown to the left and right of the line; respectively. Error bars show standard error of the mean.*

3.3. Syllable loss/gain in L2 versus L1 speech

In order to more fully understand the difference in ID across L1 and L2 English (specifically, more syllables for L2 than L1 productions of the same English text), we counted the number of orthographic syllables in the NWS text (i.e. we determined the expected number of syllables according to standard American English pronunciation). Based on this number (139) we can define syllable loss/gain as follows:

$$\sigma_{LG} = (\sigma_a - \sigma_o) / \sigma_o \quad (3)$$

where σ_o is the number of orthographic syllables, and σ_a is the number of acoustic syllables. As shown in Table 3, the average number of syllables in the NWS productions by L1 English speakers was 123, indicating elision of 16 syllables on average. This represents an average syllabic loss/gain of -11% (i.e. a loss of 11%). In contrast, the average number of syllables in the English NWS productions by L2 English speakers was 147, indicating that the L2 speakers produced the NWS passage with more syllables than the dictionary pronunciation would prescribe. The L2 average syllabic loss/gain was +6% (i.e. a gain of 6%). A closer examination of the individual L2 speaker productions of the NWS passage is required to determine whether the low L2 syllabic loss is due to epenthesis of vowels to break up some of the typologically anomalous complex consonant clusters of English and/or a tendency for L2 speakers to produce full vowels where L1 English rhythm would typically permit syllable reduction. Cross-language analyses would also establish the extent to which L1 phonotactic transfer accounts for cross-group variation. Critically, these data show that the slow rate of L2 relative to L1 speech is due to temporal restructuring that goes beyond an increase in the duration of segments, syllables, and words, and that this L2 temporal restructuring influences the ID of L2 speech.

Table 3: *Syllable loss/gain (wrt 139 orthographic syllables) and average number of syllables in the English NWS passage.*

Language Group	Syllable Loss/Gain	Avg. # Syllables (SE)
L1 English	-11%	123 (2)
L2 English (combined)	6%	147 (4)
Turkish	-9%	127 (2)
Korean	-2%	136 (4)
Vietnamese	-1%	138 (5)
Cantonese	6%	147 (3)
Mandarin	8%	150 (9)
Spanish	10%	153 (18)
Hindi	13%	158 (30)
Russian	14%	158 (24)
Portuguese	14%	158 (8)
Hebrew	30%	181 (27)

4. Discussion

In their cross-language analysis of speaking rate, Pellegrino and colleagues [5] hypothesized that cross-language variation in speaking rate may be constrained by general characteristics of human information processing of temporally dynamic signals. Specifically, according to this hypothesis, utterances should exhibit a trade-off between information density and rate of production of speech units (such as syllables) as a means of regulating the amount of information transmitted in a unit of time (information rate). This hypothesis has now been supported by cross-language comparisons reported by Pellegrino and colleagues [5] as well as the present study. Inspired by this account for speaking rate variation across L1 speech in various languages, the present study compared speech rate, information density, and rate of syllable loss/gain in productions of an English passage (NWS) by L1 and L2 English speakers. The results demonstrated that L2 English is characterized by slower speaking rate, lower information density, and less syllable reduction than L1 English, a combination that at an extreme might deviate substantially from the optimal information transmission rate for dynamic signals.

We might speculate that some combination of these three information theoretic parameters – speech rate, information density, and syllable count – all of which relate to information rate (bits of information conveyed per unit of time), might eventually be exploited for detecting, evaluating, and enhancing foreign-accented speech by both humans and machines. Next steps towards this goal require analyses of more extensive data sets (more texts, more speakers, and more languages), particularly for L2 speech in various languages with phonotactic and prosodic structures that promote and/or constrain production strategies that may impact the overall temporal structure of spoken utterances in various ways.

5. Acknowledgement

Work supported by Grant R01-DC005794 from NIH-NIDCD.

6. References

[1] S. Guion, J. Flege, S. H. Liu, and G. H. Yeni-Komshian, "Age of learning effects on the duration of sentences produced in a second language," *Applied Psycholinguistics*, vol. 21, no. 2, 205-228, 2000.

[2] M. Baese-Berk, and T.Morrill, "Speaking rate consistency in native and non-native speakers of English," *Journal of the Acoustical Society of America,* vol. 138, no. 3, EL223-EL228, 2015.

[3] M. L. García Lecumberri, M. Cooke, M. Wester, "A bi-directional task-based corpus of learners' conversational speech," *International Journal of Learner Corpus Research*, vol. 3, no. 2, 175-195, 2017.

[4] A. R. Bradlow, M. Kim, and M. Blasingame, "Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, 886–899, 2017.

[5] F. Pellegrino, C. Coupé, and E. Marsico, "Across-language perspective on speech information rate," *Language,* vol. 87, no. 3, 539-558, 2011.

[6] C. E. Shannon, and W. Weaver, *The mathematical theory of communication.* Urbana: University of Illinois Press, 1949.

[7] A. R. Bradlow, *The ALLSSTAR Corpus,* Retrieved from http://groups.linguistics.northwestern.edu/speech_comm_group/allsstar2/#!/.

[8] S. D. Soli and L. L. Wong, "Assessment of speech intelligibility in noise with the Hearing in Noise Test," *International Journal of Audiology*, vol. 47, no. 6, 356-361, 2008.

[9] A. de Saint-Exupéry, *Le Petit Prince*, Harcourt, 1943.

[10] United Nations, Universal Declaration of Human Rights, Retrieved from http://www.un.org/en/universal-declaration-human-rights/index.html.

[11] *The Handbook of the International Phonetic Association*. Cambridge University Press, Cambridge, UK, 1999.

[12] A. R. Bradlow, *OSCAAR: The Online Speech/Corpora Archive and Analysis Resource,* Retrieved from https://oscaar3.ling.northwestern.edu.

[13] The Versant™ English Test: Automatic evaluation of the spoken English skills of non-native English speakers (Pearson Education, Inc., Menlo Park, CA).

[14] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, 385 – 39, 2009.

[15] K. Johnson, K. "Massive reduction in conversational American English," in *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, edited by K. Yoneyama and K. Maekawa, The National International Institute for Japanese Language, Tokyo, Japan, pp. 29–54, 2004.

[16] L. A. Burchfield and A. R. Bradlow, A. R., "Syllabic reduction in Mandarin and English speech," *Journal of the Acoustical Society of America-Express Letters,* vol. 135, no. 6, EL270 –EL276, 2014.