Article

A Multidimensional Scaling Study of Native and Non-Native Listeners' Perception of Second Language Speech

Perceptual and Motor Skills
2016, Vol. 122(2) 470–489

DOI: 10.1177/0031512516636528
pms.sagepub.com

Jennifer A. Foote
University of Alberta, Edmonton, AB, Canada

Pavel Trofimovich
Concordia University, Montreal, QC, Canada

Abstract
Second language speech learning is predicated on learners' ability to notice differences between their own language output and that of their interlocutors. Because many learners interact primarily with other second language users, it is crucial to understand which dimensions underlie the perception of second language speech by learners, compared to native speakers. For this study, 15 non-native and 10 native English speakers rated 30-s language audio-recordings from controlled reading and interview tasks for dissimilarity, using all pairwise combinations of recordings. PROXSCAL multidimensional scaling analyses revealed fluency and aspects of speakers' pronunciation as components underlying listener judgments but showed little agreement across listeners. Results contribute to an understanding of why second language speech learning is difficult and provide implications for language training.

Keywords
second language speech, speech perception, multidimensional scaling

Corresponding Author:
Jennifer A. Foote, English Language School, Faculty of Extension, University of Alberta.1-024 Enterprise Square, 10230 Jasper Ave, Edmonton, AB, Canada, T5J 4P6.
Email: jfoote@ualberta.ca

Introduction

Second language speech development is predicated on the idea that learners notice differences between their own speech and that of their interlocutors. Simply put, learners need to attend to language features in some way for input to be processed and developed (Schmidt, 2001). Language teachers and researchers are, therefore, often looking for ways to help learners notice the "gap" between the learners' own output and the language that they are exposed to. Traditionally, the idea of noticing the gap has pointed to differences between the speech produced by learners and the target-language speech as spoken by native speakers. However, the reality for many learners is that frequently their interlocutors are other non-native speakers. It is important to consider not only how learners hear their own speech in relation to native-speaker models but also how learners compare their speech to the speech of other non-native speakers. Therefore, this study used multidimensional scaling (MDS) to understand which dimensions underlie the perception of second language speech by learners as compared to native speakers.

Background

One framework which can accommodate the development of second language speech is Flege's (1995) speech learning model. This model holds that learning second language speech involves creating and using long-term memory representations (categories) for various aspects of speech, including individual segments (sounds) and prosody (Flege, 2003; Mennen, 2015). Yet, how well and how quickly learners establish such categories depends on their ability to detect differences between what they already know, which includes the knowledge of their previously learned language(s), and what they can detect in the ambient linguistic input. For example, if a particular aspect of second language speech, such as an individual consonant, is sufficiently distinct from perceptually similar consonants that learners already know and use, and if learners are sensitive to such differences, it is likely that a separate sound category will be established for this consonant. In contrast, when perceptual similarity is high, so that the target consonant is entirely assimilated, or subsumed within, a learner's existing sound category, it is less likely that a separate second language representation will be created. Most importantly, the model assumes that the capacity for learners to learn second language speech remains intact across an individual's lifespan, and ascribes an important role to input, in terms of its quantity and quality, in enabling learners to perceive crucial perceptual differences which could trigger the process of creating speech categories (Flege, 2009). In essence, this view is premised on the idea that a certain amount of perceptual sensitivity is required for learners to engage or re-engage second language speech learning.

If learners are exposed to the speech of native-speaking interlocutors, learners would have at least some opportunities to notice how their speech differs from

native-speaker output. Some perceived differences might include various linguistic dimensions, such as individual segments or specific aspects of prosody, dysfluencies, poor word choice, and grammar errors, all of which have been linked to listener perception of second language speech (Derwing, Rossiter, Munro, & Thomson, 2004; Isaacs & Trofimovich, 2012; Kang, Rubin, & Pickering, 2010). Thus, exposure to native-speaker input might gradually help learners align their speech—through sustained input and output practice—with the speech of their interlocutors (e.g. Derwing & Munro, 2013; Saito & Lyster, 2012). However, many learners do not frequently speak with native speakers, but instead interact with other second language users, largely in the absence of the types of feedback, interactional modifications, and focus on language typical of language classrooms. This raises the question of whether learners notice how their speech differs from the production of other non-native speakers and, if they do, which speech dimensions underlie these perceived differences.

Unfortunately, current literature offers little to address these issues because previous perception research has focused primarily on native rather than non-native listeners. For instance, research has targeted the role of various linguistic dimensions, such as individual segments or aspects of prosody and fluency, in listeners' perception of speech (e.g. Derwing & Munro, 2001; Kang et al., 2010). Research has also focused on contextual factors, such as the availability of written transcripts (Levi, Winters, & Pisoni, 2007) or the inclusion of native-speaker data among non-native speech samples (Flege & Fletcher, 1992), and examined listener-internal factors, including individual differences in listeners' cognitive variables, training/experience, and familiarity with second language speech (Isaacs & Trofimovich, 2011; Isaacs & Thomson, 2013; Winke, Gass, & Myford, 2013).

While there is less research focusing on non-native listeners, interest in this area is increasing. To date, many of the studies that have involved non-native listeners evaluating second language speech have focused on whether listeners can better understand accented speech if they share the native language of the speaker, examining what is known as the interlanguage speech intelligibility benefit. Results of these studies have been mixed, with some finding evidence that sharing a first language with a speaker does help with understanding second language speech (e.g. Bent & Bradlow, 2003; Munro, Derwing, & Morton, 2006) and others casting doubts over such a benefit (e.g. Stibbard & Lee, 2006).

Other perception studies targeting non-native listeners have investigated the perception of speech rate and fluency (e.g. Derwing & Munro, 2001; Rossiter, 2009), phonetic features (e.g. Riney, Takagi, & Inutsuka, 2005), and the role of segments versus prosody in judgments of speech (e.g. Winters & O'Brien, 2013). For instance, Field (2005) investigated the role of lexical stress in intelligibility, both for native and non-native listeners. Both listener groups, who evaluated words with correct or misplaced stress, found stress to be important for word intelligibility, with both groups performing in essentially similar ways. However,

there is still much that is unknown about which aspects of second language speech are most salient to non-native listeners, especially when listeners are not directed (through research design or explicit instructions) to respond to or reflect on particular dimensions of speech.

To summarize, previous second language perception research has chiefly focused on native speakers as listeners or compared non-native listeners whose linguistic background either matched or mismatched the background of the speaker. As argued previously, a focus on second language listeners' perception of non-native speech is crucial because, in the majority of contexts, second language users tend to interact with other second language users, with the consequence that non-native speech often represents the *only* input that learners receive. To address this issue, 15 non-native university-level students and a comparison group of 10 native speakers of English were asked to listen to short excerpts of second language speech recorded as part of two tasks (reading, interview), with each excerpt presented for comparison against all other excerpts. The listeners rated how dissimilar each pair of speakers sounded, and these judgments were subsequently analyzed using MDS, a procedure which uses similarity or dissimilarity responses to plot stimuli (non-native speakers, in this case) in an *n*-dimensional space. To interpret the dimensions underlying listeners' judgments, the dimensional coordinates for each speaker were compared against background characteristics of the speakers as well as several coded measures of pronunciation, fluency, lexis, and grammar, based on each speaker's excerpt. The research question asked, "Which dimensions underlie second language listeners' perception of non-native speech in controlled reading and extemporaneous interview tasks?"

Method

Participants

The non-native participants were 15 second language speakers of English (M age = 24.8 years, SD = 3.2, range = 20–30), recruited from an English-medium university in Montreal, Canada. The speakers were enrolled in various undergraduate (3) and graduate (12) degree programs and represented a range of backgrounds, including Farsi (5), Telugu, Chinese, French (2 each), Akan/Twi, Arabic, Bengali, and Kinyarwanda (1 each). The speakers had studied English for a mean of 12.2 years (SD = 4.6, range = 2–19) and had resided in Canada for a mean of 0.7 years (SD = 0.6, range = 0.2–2.5). No speaker reported any hearing difficulty or hearing-related disorder. All speakers were men, to ensure that the most obvious potential difference between speech samples (gender) did not factor into listener judgments and mask more interesting differences. The speakers, recruited during the first semester of their studies, had reported recent TOEFL iBT standardized English proficiency scores, with a mean of 89.33 (SD = 6.85,

range = 79–104) and individual subscores of 21.08 (SD = 2.84, range = 17–26) for speaking, 21.50 (SD = 3.03, range = 17–25) for writing, 22.75 (SD = 4.05, range = 16–30) for reading, and 24.00 (SD = 3.54, range =20–30) for listening. All participants were volunteers recruited via emails sent out through student listservs and were paid an honorarium for participating.

Materials

The materials included 30 audio samples recorded by the second language speakers as part of two speaking tasks, which differed in formality (controlled reading vs. spontaneous speaking in response to an interview question). The task variable was manipulated because non-native speakers differ in accuracy and fluency of second language output by task, such that read-aloud tasks often elicit more accurate production of segments and prosody than more spontaneous tasks, such as storytelling and interviews (Rau, Chang, & Tarone, 2009). All recordings were created with high-quality microphones (Plantronics DSP-300) in a quiet, windowless testing room which, although not fully soundproof, was shielded from ambient noise and provided little distraction. The order of tasks was counterbalanced across speakers. The reading task, based on a short paragraph from an English as a second language textbook (Grant, 2001), elicited speech samples that were identical and, therefore, maximally comparable in content. After removing initial hesitations and dysfluencies, the first two sentences ("Have you noticed that some people interrupt conversations more than other people? All cultures do not have the same rules governing these areas of communication") were extracted from each recording and saved as separate files. The resulting audio samples (25 words), which were on average 9.78 s in duration (SD = 1.30, range = 8.33–12.17), were used as target audio samples from the reading task. While listeners can be sensitive to differences in speech even when evaluating very short speech samples (e.g. Flege, 1984), a longer utterance was used to include enough speech to capture a full thought. The free-response task was based on an interview question from the IELTS English proficiency test, *Describe a job you would like to do in the future* (Jakeman & McDowell, 2008). Unlike the reading task, the interview task elicited spontaneous speech, allowing speakers to have control over their linguistic output while keeping thematic content (future employment) constant. After removing initial dysfluencies, the first few complete ideas from each speaker's response (15–36 words), with a mean duration of 9.61 s (SD = 2.40, range = 6.52–14.48), were excised from the recordings and used as target samples from the interview task.

Similarity rating

Roughly three months after recording the audio samples, the same 15 non-native participants returned to participate in individual listening sessions (about 2 h in

total) to evaluate the recordings from all 15 participants using a paired comparison method. This was followed by similar sessions completed by 10 additional native English listeners, who served as a native-speaker baseline group. The 10 native English listeners, with a mean age of 25.2 years (SD = 4.5, range = 20–36) were recruited from the same university. These participants (four females, six males) were born and raised in English-speaking homes and were exposed to English from birth, with one (6) or both (4) parents being native English speakers. Because the native listeners resided in Montreal (a multicultural, bilingual French-English city with a large population of immigrants), they were familiar with accented English as spoken by speakers from diverse language backgrounds.

The listening sessions were completed in the same location as the recording sessions. Listeners used high-quality headsets (Koss R/80) to listen to the audio files and were able to control the volume. Listeners were informed that their task was to help researchers evaluate audio recordings by judging how similar or dissimilar each pair of recordings sounded to them. They received no guidance as to how they should judge the recordings nor what they should attend to. The rating of '1' was reserved for audio recordings that sounded very similar, while the rating of '9' designated very dissimilar recordings. Listeners then performed a brief practice task, judging the similarity of three recordings (containing the same sentence, "Knowing when to take turns in a conversation in another language can sometimes cause difficulty") spoken by three additional non-native speakers, with each pairwise combination of the recordings presented in a unique randomized order. However, no feedback was given to the participants as to how well or poorly they rated the recordings. Listeners then proceeded to evaluate the target audio samples, organized in two separate blocks by task (reading, interview), with the order of tasks counterbalanced across listeners. Listeners would rate all of the audio samples from one of the tasks; they would then be given a break before listening to all of the audio samples from the other task. At the beginning of each block, listeners were reminded about the directionality and the endpoints of the scale and were encouraged to use the scale's entire range. For the reading task, listeners were told that each recording featured the same two sentences, which were then shown to them. For the interview task, they were informed that the content of each recording was different but that all speakers described their future job.

Within each block, the 15 audio samples from the reading task and the 15 samples from the interview task (22.5 kHz, 16-bit resolution) were presented to each listener in all possible pairwise combinations (for a total of 105 pairs per block). The experiment was controlled by E-Prime, a software application for running psycholinguistic experiments (Schneider, Eschman, & Zuccolotto, 2002), and each listener used a headset and clearly labeled 1–9 keys on a computer keyboard to record their judgment. Each trial started with a warning which read "Next pair..." and stayed on the screen for 1.5 s, followed by two

audio samples played in sequence with a 0.25-s interval. All pairs were presented in a unique random order, which included random designation of each recording as the first or second in each pair, and each trial terminated when a response was logged, which initiated a subsequent trial.

Speech analysis

To relate psychological dimensions underlying similarity judgments to specific properties of speech, the 30 target recordings from the reading and interview tasks were analyzed for several pronunciation, fluency, lexis, and grammar variables: (1) segmental errors: number of phonemic substitutions (e.g. *the* spoken as *da*); (2) syllable structure errors: number of phonemic insertion/deletion errors (e.g. *would* without the final/d/), with both error counts divided by the total number of words; (3) word stress errors: number of misplaced or missing stresses in polysyllabic words (e.g. *com-PU-ter* spoken as *COM-pu-ter*) over the total number of polysyllabic words; (4) total sample duration (second) as a coarse measure of fluency; (5) unfilled pauses: total number of silent pauses lasting longer than 0.4 s (e.g. *I think in the future I will still I will still be* [unfilled pause] *a software engineer*); (5) filled pauses: total number of nonlexical pauses such as uh and um (e.g. *In the future I'd like to work in uh* [one filled pause] *corporate finance*), with both filled and unfilled pause measures normalized by dividing pause frequency by the total duration of the recording (yielding pause frequency per second of speaking time); (6) speech rate: total number of syllables produced (including pauses and dysfluencies) over the total duration of the recording (syllables per second); (7) repetitions/self-corrections: all immediately repeated and self-corrected words (e.g. *I I* [repeated] *worked in China for for* [repeated] *about seven years as a softwa ware* [corrected] *engineer*); (8) grammar errors: number of words with at least one error in sentence structure, morphology, or syntax (e.g. *The first time I touched the computer is in my primary school* spoken with a definite article before *computer* and the wrong tense of the verb *to be*) divided by the total number of words; (9) lexical errors: number of incorrectly used or inappropriate lexical expressions (e.g. *desired job* instead of *dream job*) over the total number of words; (10) token frequency or the total number of words spoken; (11) type frequency or the total number of unique words produced, with both token and type frequency corrected for differences in sample length by dividing the raw counts by the total sample duration (yielding token and type rates per second of speaking time).

All measures were first coded by a trained coder then recoded by another trained coder. Although all coding decisions involve a certain degree of subjectivity and may not reflect the variability found in spoken language, only 33 (9%) of all coded data cells involved disagreement (Cronbach's $\alpha = .98$), which was resolved through discussion. Table 1 summarizes the 12 coded linguistic variables from the second language speakers' recordings in the reading and interview tasks.

Table 1. Summary of coded linguistic variables in non-native speakers' speech in reading and interview tasks.

	Reading task		Interview task	
Variable	M (SD)	Range	M (SD)	Range
Segmental errors	0.20 (0.13)	0.08–0.48	0.11 (0.08)	0.00–0.29
Syllable structure errors	0.07 (0.06)	0.00–0.23	0.06 (0.06)	0.00–0.23
Word stress errors	0.16 (0.13)	0.00–0.40	0.20 (0.23)	0.00–0.23
Sample duration	9.78 (1.30)	8.33–12.17	9.61 (2.36)	6.52–14.48
Unfilled pauses	0.11 (0.04)	0.00–0.19	0.20 (0.13)	0.00–0.40
Filled pauses	0.01 (0.03)	0.00–0.10	0.21 (0.20)	0.00–0.72
Speech rate	4.30 (0.55)	3.32–4.98	3.49 (0.87)	2.07–5.53
Repetitions/self-corrections	0.02 (0.02)	0.00–0.07	0.05 (0.07)	0.00–0.20
Grammar errors	0.01 (0.01)	0.00–0.04	0.05 (0.01)	0.00–0.22
Lexical errors	0.00 (0.00)	0.00–0.00	0.02 (0.03)	0.00–0.09
Token frequency	2.64 (0.30)	2.06–3.00	2.54 (0.51)	1.64–3.53
Type frequency	2.42 (0.29)	1.89–2.76	1.95 (0.35)	1.35–2.66

Because the coded data did not conform to a normal distribution, they were analyzed using nonparametric tests (Spearman correlations), as discussed below.

Analysis

The data from the similarity rating task were analyzed using MDS, an exploratory procedure which uses similarity or dissimilarity matrices to generate a representation of stimuli in geometric (Euclidian) space, with each stimulus (e.g. an item or a person) plotted as a point and inter-stimulus distances showing similarity or "psychological distances" between them. The spatial map yielded by MDS represents a visual depiction of underlying dimensions governing a stimulus set, and a researcher's challenge is to identify and interpret these dimensions (Borg & Groenen, 1997). In this study, all MDS outputs were generated through SPSS 21.0 using the PROXSCAL algorithm (Busing, Commandeur, & Heiser, 1997), with 100 random iterations, and all similarity data (based on 9-point Likert scales) treated as ordinal. Because non-native listeners' ratings of their own speech, relative to the speech of their peers, may have affected their judgments (e.g. with own speech rated more favorably, compared to the speech of others), the final similarity matrices for non-native listeners excluded the 14 data points involving listeners' own speech. The final similarity matrices were thus based on a total of 91 pairwise comparisons for each non-native listener and 105 pairwise comparisons for each native listener.

Results

Reading task

Scree plots, which depict stress (a measure of goodness of fit between estimated interstimulus distances and the original listener-based similarity matrices), were inspected first to determine the optimal dimensionality of MDS outputs for non-native listener and native listener data in the reading task. For both outputs, a two-dimensional solution was chosen because adding subsequent dimensions failed to substantially increase fit. The final two-dimensional models featured low stress functions ($S_{\text{Non-native}} = .10$; $S_{\text{Native}} = .10$), which were considered excellent (Jaworska & Chupetlovska-Anastasova, 2009), and high dispersion indexes ($DAF_{\text{Non-native}} = .97$; $DAF_{\text{Native}} = .96$), which exceeded the minimum acceptable value of .60 (Meyer, Heath, Eaves, & Chakravarti, 2005), with each model accounting for over 96% of the variance in the input data. Therefore, both MDS outputs were plotted in two-dimensional Euclidian space, using MDS dimensional coordinates, with the first dimension plotted along the x axis and the second dimension along the y axis (Figure 1).

To interpret MDS output dimensions for both the non-native listener and native listener outputs, two-tailed Spearman correlations were computed between the dimensional coordinates from each MDS output and all relevant non-native speaker background characteristics (e.g. first language group, TOEFL score) and speech variables (e.g. speech rate, lexical errors). The results of these analyses are summarized in Table 2. For non-native listeners,

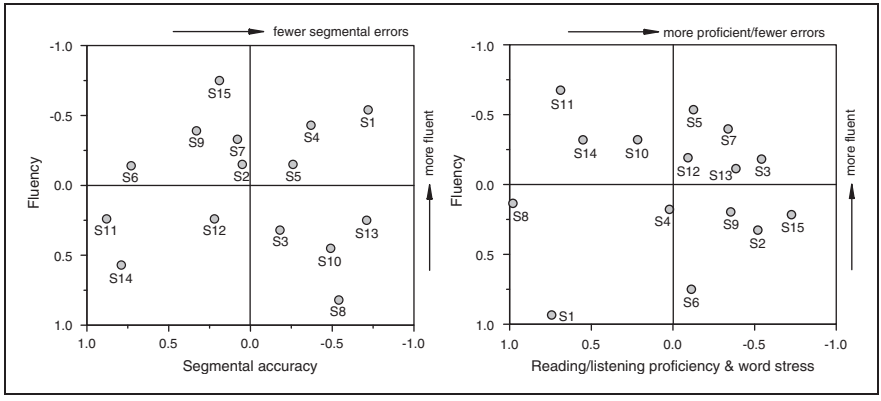

Figure 1. Two-dimensional plots representing MDS outputs for non-native listeners (left) and native listeners (right) in the reading task. The plots depict two dimensions best explaining listeners' dissimilarity ratings for 15 second language speakers (S1-S15), with each speaker compared against all other speakers.

Table 2. Spearman correlation coefficients (two-tailed) between MDS dimensional coordinates and various speaker background and speech characteristics from the reading task.

	Non-native listeners		Native listeners	
Variable	Dimension 1	Dimension 2	Dimension 1	Dimension 2
Language background	.78***	−.08	−.02	−.36
TOEFL score	−.14	−.15	−.86***	−.31
TOEFL reading subscore	.46	−.19	−.63*	−.23
TOEFL listening subscore	−.02	−.17	−.62*	.20
TOEFL speaking subscore	−.40	.31	−.11	−.51
TOEFL writing subscore	.20	−.11	−.06	−.30
Segmental errors	−.66**	.08	−.15	−.23
Syllable structure errors	.18	.17	.29	−.31
Word stress errors	.36	−.38	−.55*	.30
Sample duration	.47	.56	.03	−.19
Unfilled pauses	−.24	−.25	.14	.59*
Filled pauses	.10	.28	.36	−.47
Speech rate	−.45	−.60*	−.09	.23
Repetitions/self-corrections	.02	.28	−.19	−.29
Grammar errors	−.14	.23	−.14	−.09
Token frequency	−.40	−.61*	−.16	.20
Type frequency	.41	−.61*	−.13	.19

Note. Directionality of associations is uninformative because the MDS solution was rotated to achieve the best speaker clustering in two-dimensional space. MDS: multidimensional scaling.
*$p < .05$.
**$p < .01$.
***$p < .001$.

Dimension 1 could best be interpreted in terms of a combination of speakers' first language background and segmental errors in their speech, reflecting a common observation that segmental errors are specific to speaker background. Dimension 2 was associated with speech rate and with token and type production ratios, expressed as the number of word tokens and types uttered per second of speaking time. Based on these data, the dimensions underlying non-native listener perception of non-native speech in the reading task could be labeled as *segmentals* (Dimension 1) and *fluency* (Dimension 2). For native listeners, Dimension 1 strongly patterned with the speakers' overall TOEFL iBT performance, especially reading and listening subscores, which likely reflected orthography-mediated links between reading and listening, as well as with speakers' word stress errors. Thus, the dimensions underlying native listener perception of second language speech in the reading task could be labeled as second language

reading/listening proficiency & word stress (Dimension 1) and fluency (Dimension 2).

To quantify possible differences between the non-native listener and native listener MDS outputs from the reading task, inter-speaker distances from the two MDS outputs were correlated using a Spearman correlation test. The assumption was that perceptual distances between each speaker in the non-native listeners' two-dimensional space should match closely the corresponding distances in the two-dimensional space generated by native listeners if both groups approached the task in a similar way (Hout, Papesh, & Goldinger, 2013). This analysis yielded a weak correlation, $r(103) = .36$, $p < .0001$, suggesting that only about 13% of variance in inter-speaker distance was common between the MDS outputs for the two listener groups. Put differently, the speaker pairs perceived as being similar by non-native listeners were often not the same pairs perceived as similar by native listeners. In sum, while both MDS outputs included a similar number of dimensions, the two groups approached speech rating in the reading task in different ways.

Interview task

As in the previous analyses, scree plots were consulted first to determine most optimal solutions for MDS outputs from the interview task. For both outputs, two-dimensional solutions were deemed most appropriate, with excellent stress functions ($S_{\text{Non-native}} = .12$; $S_{\text{Native}} = .10$) and high-dispersion indexes ($DAF_{\text{Non-native}} = .95$; $DAF_{\text{Native}} = .96$) explaining over 95% of the variance in the input data. Therefore, as with the data from the reading task, both MDS outputs from the interview task were plotted in two-dimensional Euclidian space (Figure 2).

To interpret MDS output dimensions, similar Spearman correlational analyses were conducted for both the non-native listener and native listener outputs, relating MDS dimensional coordinates to several background and speech variables (Table 3). For non-native listeners, Dimension 1 patterned singly with the speakers' TOEFL iBT speaking subscore, while Dimension 2 was linked to the total duration of the speech sample, rate of unfilled pausing, and token frequency, expressed as word tokens spoken per second of speaking time. Thus, the MDS dimensions underlying non-native listener perception of non-native speech in the interview task could be labeled as second language speaking proficiency (Dimension 1) and fluency (Dimension 2). For native listeners, Dimension 1 was not uniquely associated with any single variable investigated here, but the strongest association involved speakers' word stress errors, $r = -.50$, $p = .06$. Dimension 2 was uniquely linked to the total duration of the speech sample, likely reflecting speakers' verbal fluency within total amount of speaking time. Therefore, the MDS dimensions underlying native listeners' perception of non-native speech in the interview task could be tentatively referred to as word stress (Dimension 1) and fluency (Dimension 2).

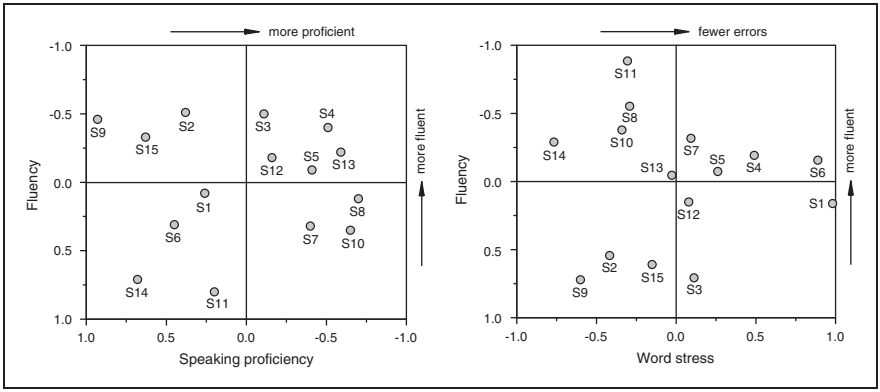

Figure 2. Two-dimensional plots representing MDS outputs for non-native listeners (left) and native listeners (right) in the interview task. The plots depict two dimensions best explaining listeners' dissimilarity ratings for 15 second language speakers (S1-S15), with each speaker compared against all other speakers.

Again, to quantify differences between the two MDS outputs in the interview task, a Spearman correlation was carried out to compare inter-speaker distances from the two outputs. This analysis revealed a moderate correlation, $r(103) = .50$, $p < .0001$, indicating that the two outputs shared about 25% of variance in terms of how closely the same non-native speakers were positioned in the perceptual spaces generated by the two listener groups. Thus, while there was some correspondence in how both sets of listeners approached the task of rating non-native speech in the interview task, the perceptual spaces generated by the two groups were largely different.

Discussion

The research question asked which dimensions underlie native and non-native listeners' perceptions of second language speech in controlled reading and extemporaneous speaking tasks, in the absence of directions for listeners to attend to any specific speech elements. For non-native listeners, segmental accuracy and fluency appeared to underlie listeners' perceptions of second language speech in the reading task, while second language speaking proficiency and fluency reflected their judgments in the interview task. For native listeners, word stress accuracy, along with second language reading/listening proficiency and fluency, characterized listeners' ratings of second language speech in the reading task, while word stress accuracy and fluency were the two dimensions relevant to listener judgments in the interview task. The most consistent finding was fluency as a common component underlying the perception of non-native speech. Despite these similarities, there was only moderate agreement across the two

Table 3. Spearman correlation coefficients (two-tailed) between MDS dimensional coordinates and various speaker background and speech characteristics from the interview task.

	Non-native listeners		Native listeners	
Variable	Dimension 1	Dimension 2	Dimension 1	Dimension 2
	.47	.38	−.30	−.22
TOEFL score	−.04	−.32	.03	.22
TOEFL reading subscore	.28	−.11	−.21	.13
TOEFL listening subscore	−.17	−.41	−.05	.07
TOEFL speaking subscore	−.67*	.00	.31	−.23
TOEFL writing subscore	.32	.01	.18	.24
Segmental errors	.13	.29	−.09	−.27
Syllable structure errors	.49	−.23	−.28	.43
Word stress errors	.15	.12	−.50	−.07
Sample duration	−.23	.66*	−.14	−.72*
Unfilled pauses	−.17	.55*	.47	−.51
Filled pauses	−.14	−.38	−.18	.36
Speech rate	.36	−.47	−.25	.44
Repetitions/self-corrections	.23	.03	−.08	.23
Grammar errors	.22	−.32	−.16	.41
Lexical errors	−.42	−.04	−.20	−.32
Token frequency	.08	−.54*	−.11	.42
Type frequency	−.13	−.49	−.13	.29

Note. Directionality of associations is uninformative because the MDS solution was rotated to achieve the best speaker clustering in two-dimensional space (Figure 1).
MDS: multidimensional scaling.
*$p < .05$.

groups in the interview task and virtually no agreement in the reading task, suggesting that the two listener groups approached the rating of non-native speech in different ways.

Differences among listeners

The first of the two dimensions underlying listeners' perception of second language speech varied across listeners and tasks. Non-native listeners' perception was related to speakers' segmental accuracy in the reading task and to speakers' speaking ability (as measured by TOEFL speaking subscores) in the interview task. The reading task is a formal, controlled speaking activity with identical lexical content across speakers and orthographic support guiding oral production, so it was expected that the speakers would make few grammatical and

lexical errors. Therefore, it is unsurprising that segmental substitutions, which are typically specific to speakers' language background, emerged as a dimension underlying non-native listeners' judgments. For instance, segmental errors contribute to both trained and inexperienced raters' judgments of non-native accent (Kennedy & Trofimovich, 2008). Listeners' use of segmental errors to distinguish other second language speakers from one another is also consistent with a typical instructional emphasis on segmentals in language classrooms (Foote, Trofimovich, Collins, & Soler Urza, 2013) and with learners' beliefs that segmental errors constitute the greatest challenge to their pronunciation (Derwing, 2003).

Compared to the reading task, the interview task was an extemporaneous speaking activity, which allows speakers more linguistic freedom to express themselves using a vernacular speaking style. In this task, the first of the two dimensions that patterned with non-native listener judgments was speakers' global speaking ability, as measured through TOEFL speaking subscores. The TOEFL speaking section includes six tasks, of which two require test takers to speak on familiar topics while the remaining four involve either listening to or both reading and listening to relevant information before integrating it into the response. The speaking subscore appears to reflect test-takers' speaking ability, with integrated components contributing minimally to the reading and listening constructs (Sawaki, Stricker, & Oranje, 2009). The speaking subscore also seems to be distinct, compared to other modalities such as listening or writing, providing an added value to the total test score (Sawaki & Sinharay, 2013). It appears, then, that non-native listeners were sensitive to more than segmental errors or features of accent, as they tried to distinguish one second language user from another in the interview task. Yet, because only global aspects of proficiency, rather than any of the specific linguistic properties of speech (from among those targeted here), patterned with listener ratings, there is no evidence in these data to suggest which linguistic aspects of speech were relevant to making up the speaking ability construct.

Unlike non-native listeners, native listeners appeared to rely on the dimension of word stress accuracy in speakers' output in both reading and interview tasks, with an additional contribution of speakers' reading/listening proficiency in the reading task. Because reading aloud is an activity with fixed lexical and grammatical expression, it is reasonable that native listeners relied on speakers' reading and listening proficiency (as measured through TOEFL reading and listening subscores) in judging how similar or different second language speakers sounded. Indeed, reading aloud involves both reading and listening skills, as it requires speakers to code orthography into phonology, then to articulate the prepared speech plan and to perceptually monitor the speech output for accuracy (e.g. Levelt, 1989). However, the finding that word stress was linked to native listeners' judgments of second language speech in both reading and interview tasks is noteworthy in light of the importance of prosody, which includes

such linguistic categories as intonation, stress, and rhythm, for second language speech learning and teaching. For example, word stress and rhythm (along with other prosodic features) may account for up to 50% of the variance in accent judgments for non-native speakers from varied linguistic backgrounds (Isaacs & Trofimovich, 2012 Kang et al., 2010). Word stress also contributes to listeners' perceptions of comprehensibility for speakers from multiple first language groups (Crowther, Trofimovich, Saito, & Isaacs, 2015) and to intelligibility for both native and non-native listeners (Field, 2005). Similarly, speech training focusing on word stress, along with other prosody and fluency characteristics of speech, can lead to measurable gains in learners' comprehensibility in extemporaneous speaking tasks, as compared to an equivalent amount of instruction targeting only individual sounds (Derwing, Munro, & Wiebe, 1998). In fact, the role of word stress in native listeners' perception of speech might be related to stress being one of the most structural and hierarchical aspects of phonology (e.g. in metrical phonology), representing the core element of native speakers' linguistic competence (de la Mora, Nespor, & Toro, 2013).

Similarities among listeners

Fluency emerged as the most pervasive characteristic underlying listeners' judgments of second language speech, emerging as a dimension across both tasks and both listener groups (cf. Tables 1 and 2). In the reading task, the dimension of fluency encompassed the rate of word type and token production as well as speech rate (for non-native listeners) or pausing frequency (for native listeners). In the interview task, the dimension of fluency involved total sample duration, pausing frequency, and total number of lexical items produced (for non-native listeners) or total sample duration (for native listeners). Although the precise measures of fluency varied across listeners and tasks, these measures share one characteristic, namely, they all reflect temporal dimensions of speech output, such as frequency of pausing, duration of speaking, or lexical fluency expressed as total number of words uttered. This implies that fluency plays a prominent role in the perception of non-native speech by both native and non-native listeners. While it is impossible to know how fluency measures would affect judgments of native speakers' speech, one reason for this finding might be that temporal dimensions of fluency, as perceived by the listener, may mark second language speakers as non-native language users. For instance, Munro and Derwing (2001) showed that a 10% increase in speaking speed resulted in second language utterances being rated as less accented by native listeners. Similarly, overall utterance duration is an indicator of how native-like second language speakers sound, contributing to the perception of accent (MacKay & Flege, 2004). Compared to native speakers, second language users often have a different distribution of pauses in their speech, even though the overall number of pauses might be the same, suggesting that measures of

pausing can also act as salient markers of native-likeness (Bosker, Quené, Sanders, & de Jong, 2014). In essence, temporal elements of speech may act as salient cues distinguishing second language users from one another for both native and non-native listeners. Yet, as a comparison of MDS outputs for native and non-native listeners suggests, these temporal fluency cues differed across the two listener groups and were likely weighed by them in different ways, with the consequence that the second language speakers considered fluent by native listeners were not the same as those judged fluent by non-native listeners (Figures 1 and 2).

Implications

Following the assumption that detecting perceptual differences between learners' existing linguistic repertoire and their target-language input can trigger the process of second language speech learning (Flege, 1995, 2003), it is not at all surprising that learning pronunciation is such a complex task, especially for learners in contexts where the majority of language users are second language speakers. As shown through MDS analyses, both sets of listeners were sensitive to global aspects of second language speech, such as speaking proficiency (non-native listeners) and reading/listening proficiency (native listeners), as well as to some of specific characteristics of speech, including segmental errors (non-native listeners), word stress (native listeners), and temporal aspects of fluency, such as speech rate and pausing (both listener groups). However, outside the domain of fluency, the specific linguistic variables underlying listeners' judgments of second language speech were limited. For instance, word stress accuracy—one aspect of speech prosody—seemed to underlie native speakers' perception in both controlled and extemporaneous tasks. In contrast, non-native listeners appeared to attend only to segmental errors and only in the controlled reading task. This implies that while learners might be aware that their speech is different from an interlocutor, they may not have a clear understanding of exactly how and why it is different. If learners are unable to distinguish differences between their own linguistic performance and the language produced by their interlocutors or between the linguistic output of speakers in their environment, speech learning may not be as efficient, as it could be or may be focussed on aspects of speech which may not be as crucial to communicative success as others (e.g. segmentals vs. prosody). The current findings point to this possibility, particularly in contexts where learners are primarily exposed to non-native input.

Limitations and conclusions

Needless to say, the results of this study must be interpreted with caution, considering the small sample size targeted, the lack of a random sample of participants, and the experimental, lab-based approach employed. The non-native speech analyzed also involved only male speakers from mixed language

backgrounds. Future studies could, therefore, look at differences both within and across different language groups, in terms of speakers and listeners. As university-level users of English, the speakers also represented the range of second language ability considered sufficient for them to pursue academic studies. It would, therefore, be interesting to see how differences in listener and speaker proficiency could contribute to non-native perception. Further, as listeners' perceptions of second language speech depended on speaking task, future investigations of non-native speech perception should target different task types. Future studies could also control speaker variables beyond gender which may have played a role in similarity judgments but which are not particularly interesting in terms of learning or teaching (e.g. nasality, pitch height).

One positive aspect of these findings concerns their instructional implications. Overall, the native and non-native listeners were more similar in the interview task (with 25% shared variance) than in the reading task (with 13% shared variance). It is promising that there is more similarity with the interview task, which more closely mirrors language production in naturalistic settings. However, for the non-native listeners, the more constrained reading task led to more "focused" perception behavior, as it involved some sensitivity to segmental errors, suggesting that the controlled content of the reading task may have enabled non-native listeners to attend to specific linguistic elements in speech. In contrast, in the interview task, non-native listeners mainly attended to global aspects of second language ability, not necessarily the specific linguistic features that distinguish speakers from one another. It may, therefore, help learners to have access to controlled input enabling them to attend to linguistic features rather than content.

In fact, it might be highly difficult for second language learners interacting with other second language users to notice and acquire new features of pronunciation incidentally. For instance, in a survey of 100 learners of English in Canada, Derwing (2003) found that many were unable to describe their own pronunciation difficulties, and for those that could, small sets of segmental targets were most commonly cited, notably, those that were unlikely to cause communication problems (such as English 'th'). Arguably, the task of figuring out pronunciation difficulties from input alone will be even more challenging in contexts where most language users are non-native speakers. For this reason, explicit instruction, which can draw learners' attention to features that they do not notice through exposure alone, will be important for helping them improve their pronunciation. Notwithstanding the benefits of genuine communication for language learning, explicit pronunciation instruction is unquestionably needed to help learners develop their pronunciation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC) and Fonds de recherche sur la société et la culture (FRQSC) awarded to the second author as well as a SSHRC doctoral fellowship awarded to the first author.

References

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, *114*, 1600–1610.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York, NY: Springer.

Bosker, H. R., Quené, H., Sanders, T., & Jong, N. H. (2014). The perception of fluency in native and non-native speech. *Language Learning*, *64*, 579–614.

Busing, F. M. T. A., Commandeur, J. J., & Heiser, W. J. (1997). *PROXSCAL: A multidimensional scaling program for individual differences scaling with constraints*. In W. Bandilla & F. Faulbaun (Eds.), *Softstat 97: Advances in statistical software* (pp. 237–258). Stuttgart, Germany: Lucius.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*(4), 814–837.

de la Mora, D. M., Nespor, M., & Toro, J. M. (2013). Do humans and nonhuman animals share the grouping principles of the iambic-trochaic law? *Attention, Perception, and Psychophysics*, *75*, 92–100.

Derwing, T. M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, *59*, 547–567.

Derwing, T., & Munro, M. J. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics*, *22*, 324–337.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A seven-year study. *Language Learning*, *63*, 163–185.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*, 393–410.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 665–679.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*, 399–423.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, *76*, 692–707.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 229–273). Timonium, MD: York Press.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 319–355). Berlin, Germany: Mouton de Gruyter.

Flege, J. E. (2009). Give input a chance! In T. Piske & M. Young-Scholten (Eds.), *Input matters in SLA* (pp. 175–190). Bristol, England: Multilingual Matters.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America*, *91*, 370–389.

Foote, J. A., Trofimovich, P., Collins, L., & Soler-Urza, F. (2013). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal*. Advance online publication.

Grant, L. (2001). *Well said: Pronunciation for clear communication* (2nd ed.). Boston, MA: Heinle & Heinle.

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 93–103.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*(01), 113–140.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, *34*(3), 475–505.

Jakeman, V., & McDowell, C. (2008). *New insight into IELTS: Student's book with answers*. Cambridge, England: Cambridge University Press.

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, *5*(1), 1–10.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal*, *94*, 554–566.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *64*(3), 459–489.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Boston, MA: MIT Press.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2007). Speaker-independent factors affecting the perception of foreign accent in a second language. *Journal of the Acoustical Society of America*, *121*, 2327–2338.

MacKay, I., & Flege, J. (2004). Effects of the age of second-language (L2) learning on the duration of L1 and L2 sentences: The role of suppression. *Applied Psycholinguistics*, *25*, 373–396.

Mennen, I. (2015). Beyond segments: Towards a L2 intonation learning theory. In E. Delais-Roussarie, M. Avanzi, & S. Herment (Eds.), *Prosody and language in contact: L2 acquisition, attrition and languages in multilingual situations* (pp. 171–188). Berlin, Germany: Springer.

Meyer, J. M., Heath, A. C., Eaves, L. J., & Chakravarti, A. (2005). Using multidimensional scaling on data from pairs of relatives to explore the dimensionality of categorical multifactorial traits. *Genetic Epidemiology*, *9*, 87–107.

Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, *23*, 451–468.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111–131.

Rau, D., Chang, H. H. A., & Tarone, E. E. (2009). Think or sink: Chinese learners' acquisition of the English voiceless interdental fricative. *Language Learning*, *59*, 581–621.

Riney, T. J., Takagi, N., & Inutsuka, K. (2005). Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly*, *39*, 441–466.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*, 395–412.

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction on L2 pronunciation development of /ɹ/ by Japanese Learners of English. *Language Learning*, *62*, 595–633.

Sawaki, Y., & Sinharay, S. (2013). Investigating the value of section scores for the TOEFL iBT® Test. *TOEFL iBT-21*. Princeton, NJ: Educational Testing Service.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*, 005–30.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, England: Cambridge University Press.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

Stibbard, R. M., & Lee, J. I. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *Journal of the Acoustical Society of America*, *120*, 433–442.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252.

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, *55*, 486–507.

Author Biographies

Jennifer A. Foote is an assistant professor with the English Language School in the Faculty of Extension at the University of Alberta. Her research focuses on second language pronunciation teaching, second language speech perception, and comprehensibility.

Pavel Trofimovich is a professor of applied linguistics in the Department of Education at Concordia University, Montreal, Canada. His research focuses on cognitive aspects of second language processing, second language speech learning, sociolinguistic aspects of second language acquisition, and the teaching of second language pronunciation.