Measuring foreign accent strength in English: Validating Levenshtein Distance as a Measure

Wieling, M.; Bloem, J.; Mignella, K.; Timmermeister, M.; Nerbonne, J.

DOI
10.1163/22105832-00402001

Publication date
2014

Document Version
Submitted manuscript

Published in
Language Dynamics and Change

Measuring Foreign Accent Strength in English. Validating Levenshtein Distance as a Measure

Abstract

With an eye toward measuring the strengths of foreign accents in American English, we evaluate the suitability of a modified version of the Levenshtein distance (LD) for comparing (the phonetic transcriptions of) accented pronunciations. Although this measure has been used successfully *inter alia* to study the differences among dialect pronunciations, it has not been applied to study foreign accents. Here, we use it to compare the pronunciation of non-native English speakers to native American English speech. Our results indicate that the Levenshtein distance is a valid native-likeness measurement, as it correlates strongly with the average "native-like" judgments given by more than 1000 native American English raters ($r = -0.8$, $p < 0.001$).

Foreign accent, Levenshtein distance, edit distance, pronunciation, validation

1. Introduction

Most speakers of a foreign language speak that language with an accent, particularly if they have learned the language after childhood. McCullough (2013) emphasizes perception when she defines foreign accent as "the percept of deviations from a pronunciation norm that a listener attributes to the talker not speaking the target language natively" (p. 3). We are interested in the phonetics of the accents and how well they correlate with the perception of "non-native-likeness."

There are many reasons for developing a measure of the strength of foreign accents. Foreign accents have attracted a good deal of attention from specialists in second-language (hence: L2) learning (e.g., Derwing and Munro, 2009), but also from researchers investigating whether there is a critical period within which native-like language acquisition must occur (see DeKeyser, 2012 for a recent survey). There is a substantial amount of research investigating the factors contributing to native-like abilities. Piske, MacKay and Flege (2001) review a large body of literature noting that the age at which one begins learning (age of onset), the time spent in a country where the language is dominant (length of residence), and its amount of use may be shown to affect how native-like pronunciations ultimately become. Investigations seeking to explain the strength of foreign accents may be motivated practically (with an aim to improve second-language learning methods), but also theoretically (with an aim to understand how language is learned). Finally, accented speech is associated so often with lower socio-economic status that accent has occasionally been involved in employment litigation (Kalin et al., 1980).

The purpose of this study is to evaluate the suitability of a computational pronunciation comparison method, the Levenshtein distance (LD; Levenshtein, 1965), for comparing accented pronunciations. The LD was originally developed to measure string similarity (Kruskal, 1983). Kessler (1995) was the first to use it for comparing dialectal pronunciations on the basis of phonetic transcriptions, and the method has been used frequently in dialectology since then (e.g., Heeringa, 2004; Nerbonne and Heeringa, 1997; Wieling, 2012). LD measures the minimum number of insertions, deletions and substitutions needed to transform one pronunciation into another. Below we review other uses to which LD has been put linguistically. When calculating LD, one

automatically obtains an ALIGNMENT of corresponding segments. For example, the LD between the pronunciations [wɛnzdeɪ] and [wɛnəsde] is 3 as can be seen from the following alignment (Kruskal, 1983):

w	ɛ	n		z	d	e	ɪ
w	ɛ	n	ə	s	d	e	
			1	1			1

We introduce several refinements of this basic procedure below. Having a suitable computational method to compare accented pronunciations is attractive, as it can be easily applied to large datasets of transcribed (accented) speech, is easily replicable, and always yields consistent pronunciation distances. Because the procedure is computationally implemented, it is explicit and may therefore be analyzed by researchers seeking more detailed sources for the perception of non-native-likeness, e.g., whether novel phoneme distinctions contribute significantly to the perception that a pronunciation is non-native-like. The pronunciation distances can also be used for quantitative studies investigating foreign accents and, for example, the possible presence of a critical period in L2 learning (Piske et al., 2001).

Of course, in measuring foreign accents based on phonetic transcriptions (in fact, broad transcriptions, see below) we are obviously dependent on transcription quality, which, moreover, may vary. In addition, accents may be characterized by fine pronunciation differences, e.g., in vowel quality and diphthongization, which would be difficult to represent in any transcription system and definitely are not represented in most transcribed data, in particular, not in the data we analyze. Given Flege's (1984) observation that some American listeners were able to detect French accents in English after only 30 ms. (in the word *two*), it is clear that even a part of a segment may signal accents (assuming about 200 wd./min. are pronounced, 4.5 segments/wd., then 66 ms. is roughly one segment). But we shall pursue a line of research in which crucial indicators of accent are not identified beforehand, and in which sporadic modifications of segment pronunciations, also frequently found in accented speech, may also play an evidential role.

1.1. Related Work

Brennan and Brennan (1981) study Mexican-American speech, collecting the relative frequencies of eighteen hand-chosen variables and combining these into a single ACCENTEDNESS INDEX (using the mean). They show that this index correlates very strongly with both the expert judgments of three linguists (on accentedness) and the average lay judgments of adolescent subjects. Our work continues this tradition, but examines a range of accent sources in English and obviates the need to identify variables in an *a priori* fashion.

Magen (1998) studies two native speakers of Latin American Spanish (San Salvador and Chile) on the basis of ten features, two of which influence syllable structure (epenthetic schwa and the allomorphy of the past tense /-ed/ marker in English), and two which involve stress (lexical and phrasal). She compares the features as they occurred naturally in accented speech with alternate versions that were edited acoustically. The stress features are significant, as are fairly subtle distinctions involving reduced vowels. We return to these issues in the discussion. Our work differs in using

only unedited speech, in examining accents from a range of languages, and in not restricting attention to a small number of features.

McMahon et al. (2007) explicitly aim to measure the degree of accentedness in various forms of English world-wide, but they use an algorithm that is not fully specified and therefore not easily applicable to other datasets:

> "For now our model is primarily articulatory, particularly for consonants, though it does also include a number of *ad hoc* mechanisms to balance cases where acoustic similarity departs significantly from articulatory similarity, as for example with [f] and [x] or bunched vs retroflex /r/" (McMahon et al., 2007: 119).

There are also several acoustically based studies, beginning with Major (1987), who measures the degree of aspiration (voice onset time) in native and non-native speakers. We see the benefit of this and other acoustic studies in obviating the need for transcriptions, but on the other hand these studies are all limited in their need to focus on a small number of acoustic features (such as aspiration). McCullough (2013: Sec. 1.1) provides an excellent recent overview.

We turn to attempts to characterize the overall differences in pronunciation from one sample (variety) to another. Nerbonne and Heeringa (2010) review a good deal of literature on the use of pronunciation distance measures focusing on measuring the similarity of pronunciation in the various dialects of a language. They report on applications in more than a dozen languages, and note that Gooskens and Heeringa (2004) show that Levenshtein distances correlate well ($r \approx 0.7$) with naïve speakers' judgments of the degree of dialect differences among Norwegian dialects. Although there have been many dialect studies comparing results from LD-based analyses with others, unfortunately no other direct validation experiments have been conducted with dialectal data from other languages (but see Heeringa et al., 2006) . This means that the current paper will importantly supplement the Norwegian research as a validation study.

Wieling, Prokić and Nerbonne (2009) use the alignments underlying the LD to induce a measure of phonetic similarity between segments. Their approach is based on calculating the information-theoretic pointwise mutual information (PMI, introduced by Church and Hanks, 1990) and assigns a small distance to sound segments which align relatively frequently, whereas a larger distance is assigned to sounds co-occurring relatively infrequently. Incorporating these automatically obtained phonetic distances in the LD algorithm (rather than assigning a cost of 1 for every operation, as shown in the alignment above) also results in improved alignments. Given the intimate relation between distance and alignment, we interpret this result to indicate that the LD is also assaying pronunciation distance validly. Wieling, Margaretha and Nerbonne (2012) show that the phonetic distances obtained using the PMI technique correlate strongly with acoustic vowel distances in formant space ($0.61 < r < 0.76$) for six independent dialect datasets.

Variants of LD have been applied to questions of historical linguistics, as readers of *Language Dynamics and Change* will likely know. Kondrak (2001) uses a modified LD measure to detect cognates, as do Schepens, Dijkstra and Grootjen (2012) on a larger scale, and using only standard orthography. They report a classification performance of over 90%. The Automated Similarity Judgment Program (ASJP) is also based on a LD measure and has been put to a number of uses in historical linguistics (Wichmann,

2008; Holman et al., 2008; Müller et al., 2009). Jäger (2013) uses a variant of LD very similar to the one used here as a basis for phylogenetic inference.

Several non-dialectological and non-diachronic studies have also successfully relied on the LD to measure pronunciation differences. Kondrak and Dorr (2004) use LD to measure the pronunciation similarity of the names of proposed new drugs to existing ones. The goal is to avoid proposing names that patients and health personnel might easily confuse. Sanders and Chin (2009) use a version of the LD to measure the atypicalness of the speech of users of cochlear implants. In a study with aims similar to the present one, Gooskens, Beijering and Heeringa (2008) show that a LD based on segment distances derived from canonical spectrograms and normalized for length correlates highly with intelligibility ($r = -0.86$). We interpret their results, too, as a validation of Levenshtein distance as a measure of pronunciation dissimilarity. So we have every reason to be optimistic in proposing that LD will be suitable to measure the strength of foreign accents in English pronunciations.

Naturally, a measurement technique must be validated before its results may be relied on, which is why we compare the results of our measurements to human judgments of accent strength in this paper. As Derwing and Munro (2009: 478) insist, "listeners' judgments are the only meaningful window into accentedness."

2. Data

In this study, we use data from the Speech Accent Archive (Weinberger and Kunath, 2011). The Speech Accent Archive is digitally available at http://accent.gmu.edu and contains more than 1000 transcribed speech samples in English from people with various language backgrounds. Each speaker reads the same paragraph of 69 words in English:

> *Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*

Besides the transcriptions and the associated audio files, the Speech Accent Archive contains the following speaker-related information: birth place, native language, other language(s) spoken, age, gender, age of English onset, English learning method, current English residence (if applicable), length of English residence (if applicable).

Below we provide (the first lines of) the IPA transcriptions of (i) a German woman who lived in the U.S. for twenty-five years (german1 on the website); (ii) a French woman who lived in the U.S. for only two months (french3); (iii) an Italian man who lived in the U.S. for 3-4 months (italian2); and a Chinese woman who lived in the U.S. for 1 year (cantonese1).

German: [pʰliːs kʰal stela æsk ɚ tu bʁĩŋ d̪iːs sĩŋks wɪθ hɜ fʁɑm d̪ə stoə siks]
French: [pʰliːz kɔl stɛlɐ æsk ɛɹ tŭ bɹĩŋ d̪iz sɻ̩ŋs wɪð ɛɹ fɹʌm ðə stɔə sɪks]
Italian: [pliz kʰɔːl stɛlːa æsk hɜɹ tu bɹĩŋk d̪izə̆ t̪ĩŋ̟ wɛd hɜɹ fɹɔ̃m də stɔɹi sɪks]
Cantonese: [pʰɾis kʰɑl stɛlʌ as hɜɹ tŭ bɹĩŋ dis fɻ̩ɪŋs wɪf hɜɹ fɹɔ̃m də stoɹ sɪs]

We provide these examples to illustrate that the accent database contains a wealth of interesting data. The stereotypical elements of accents are present: every speaker has trouble with the interdental fricatives, but the substitutions are different (compare the Italian's pronunciation of 'things' to the others); the German devoices final obstruents, e.g., 'please'; the French speaker drops initial /h/ in 'her'; the Italian speaker adds a vowel to 'these' to create a second CV syllable; and the Cantonese speaker simplifies consonant clusters in words such as 'ask.' Also note that other stereotypical accent modifications and substitutions are missing or are only inconsistently found. The German, French and Italian speakers all manage the low, front vowel /æ/ although it is missing from these languages (no /ɛ/:/æ/ distinction). The French speaker devoices the final sound in 'things,' but she pronounces it in 'please,' and she uses the English approximant [ɹ] even though the French stereotypically pronounce /r/ as a uvular trill [ʀ] or uvular fricative [ʁ]. We find variation not only in the various groups of speakers but also in the speech of individual speakers.

We are aware that reading a paragraph of text may not be the best method to tap into pronunciation ability, as differences in reading ability may also affect the foreign accent (Piske et al., 2001). However, the advantage of this approach is that a set of comparable text is obtained for every speaker, enabling a straightforward comparison.

It is not surprising that individual non-native speakers vary in the degree to which they conform to stereotypes, i.e. in the strength of their accents. But since accents vary and a wide range of differences with respect to English all fall under the category of 'foreign accent,' we need a measure that takes many differences into account in assessing the strength of the foreign accent. We claim that the Levenshtein distance, introduced above, is appropriate in this respect because it yields a numerical measure representing the pronunciation difference per word, which may then be averaged over multiple words to obtain an aggregate measure of pronunciation difference.

For the validation study we extracted 395 speech samples from the Speech Accent Archive. A subset of 115 speech samples belonged to native English speakers who were born in the United States, whereas the remaining 280 speech samples belonged to speakers with a different native language (other than English) or who were born outside of the United States. In total, there were 99 different native languages represented in this sample, with the most frequent ones being Spanish (17 speakers), French (13 speakers), and Arabic (12 speakers). The remaining samples had all fewer than 10 speakers, and a total of 46 languages were only spoken by a single speaker. In most cases, the transcriptions consisted of 69 separate words. Where this was not the case (e.g., some speakers pronounced a word twice, inserted an incorrect word, merged two words into one, or forgot to pronounce a word), we manually corrected the transcription by removing the superfluous words, splitting up merged words, or marking the word as absent. We note that this procedure respects the sandhi effects in pronunciation since we kept each word transcription exactly as it appeared, including whatever sandhi effects might be present. The procedure merely ensured that each word pronounced by a single speaker was compared to the matching word (i.e. at the same position) from every other speaker in order to obtain the LDs per word (absent words did not yield an LD measure).

3. Methods

3.1. Automatically Obtaining Measurements of Foreignness

As indicated above, we employ the LD algorithm to obtain the pronunciation distances between two transcribed strings. We note that LD is restricted to measuring differences in sequences of phonetic (or phonological) segments. Suprasegmental information, including pitch, duration and intensity, is not taken into account at all. So LD is positioned to measure accent differences that are expressed segmentally, but not those that are reflected only suprasegmentally.

Obviously, the standard LD algorithm is quite crude as it simply measures the minimum number of insertions, deletions and substitutions to transform one pronunciation into the other, and consequently only distinguishes same from different (i.e. substituting completely different sounds, such as [a] for [i], is not distinguished from substituting more similar sounds, such as [ɪ] for [i]). To make the comparison more linguistically sensible, we incorporate the automatically obtained sensitive sound distances obtained using the PMI-based procedure (Wieling et al., 2009).[1] This procedure works by counting how often two segments correspond in alignment, compared to how often they would correspond by chance. Pointwise mutual information is just the logarithm of that ratio of two relative frequencies, and we use this to obtain a cost in the LD algorithm. Segments which frequently correspond are associated with low substitution costs, while infrequently corresponding elements may only substitute for one another at a substantial cost. Jäger (2013) uses a similar procedure in alignment for adducing genetic relations. Our algorithm then proceeds as follows:

1. Use the standard LD algorithm to obtain the initial alignments;
2. Count the frequency of each sound segment x involved in (non-identical) substitutions, insertions and deletions and divide this frequency by the summed frequency of all these sound segments to yield the probability of sound segment x: $p(x)$;
3. Count the frequency of each distinct pair of aligned sound segments x and y in (non-identical) substitutions, insertions and deletions (in the final two cases, one segment is empty) and divide this frequency by the summed frequency of all these sound segment pairs to yield the probability of the aligned sound segments x and y: $p(x,y)$;
4. Calculate the PMI score (Church and Hanks, 1990) for each pair of sound segments x and y using the following formula: $\mathrm{PMI}(x,y) = \log_2(p(x,y) / (p(x) p(y)))$. Higher scores are thereby assigned to segment pairs which co-occur more frequently;
5. Obtain the PMI-based segment distances by inverting (i.e. subtracting from 0) and normalizing the PMI scores to range between 0 and 1. In this way lower values indicate segment pairs which co-occur more frequently;
6. Use the PMI-based sound distances in the LD algorithm to re-align and repeat steps 2 to 5 until the alignments (and, consequently, the PMI-based sound distances) remain constant. The final PMI-based sound distances are then used in the LD algorithm to determine the PMI-based LDs.

Applying this method to our example alignment discussed in the introduction yields the following associated (sensitive) costs:

[1] Other segment distances might be used, but as Laver (1994) notes, phonetics has not succeeded in providing general methods for measuring segment differences, except in the case of vowels. See Heeringa (2004; Ch. 3, 4 and 7) for a range of attempts.

w	ɛ	n		z	d	e	ɪ	
w	ɛ	n	ə	s	d	e		
			.031	.020			.030	

In order to apply the PMI technique effectively, it is best that each segment occurs frequently (i.e. many words and speakers should be included). This also means that it is advantageous to reduce the number of different segments, which we do by ignoring all diacritics, i.e. effectively treating [t], [tʰ], [tʲ], etc. as the same segment. Naturally, this sacrifices some sensitivity in the measure, but otherwise the frequencies of correspondences in alignment are too low to reliably obtain sensible segment distances.

We obtain pronunciation distances per word using this (linguistically sensible) adaptation of the Levenshtein algorithm. As longer words are likely to vary more than shorter words, we divide the pronunciation distances by the alignment length (normalization is also employed by Schepens et al., 2012). Pronunciation distances between two speakers can then simply be obtained by calculating the word pronunciation distances for all words and averaging these. To obtain the most reliable sound segment distances, we used all 989 samples from the Speech Accent Archive for which we had transcriptions available. The most frequent languages in this set were English (181 samples), Spanish (64 samples) and French (34 samples).

Jäger (2013) also uses pointwise mutual information, but without the iteration we include. In our experience, the results are quite similar. List (2012), citing Kessler (2001) and focusing on the problem of identifying cognates in historical linguistics, suggests using correspondences from both related and unrelated words to sharpen the distinction between likely (low-cost) alignments and unlikely ones. This might well improve the induction of segment distances.

To determine the foreignness score of a speaker (with respect to American English), we calculated the mean pronunciation distance between the transcribed speech sample of the foreign speaker and the speech samples of all 115 native American English speakers in our dataset. Conceptually, this can be interpreted as comparing the foreign pronunciation to the speech of the average American English speaker.

We note two ways in which the LD-based procedure we use is superior to other approaches. First, LD is based on the entire phonemic inventory of English (and other languages) and increases based on substitutions and deletions of all sorts, rather than focusing on deviant pronunciations of single segments. Second, by being based on a sample, our procedure potentially notes cases in which phonetic (or phonemic) changes are applied sometimes, but not categorically. This is evident e.g. in the Italian speech above, where [s] is used for [θ], but not all the time. We turn now to the validation of pronunciation distances using judgments of native-likeness.

3.2. Experimental Procedure

To validate the computed Levenshtein distances, we compare them to human native-likeness ratings, as we noted above. To this end we developed an online survey in which native American English-speaking participants were presented with a randomly ordered subset of 50 speech samples from the Speech Accent Archive. Each speech sample consisted of the entire 69-word passage (for a single speaker). Participants could stop playback of a sample at any time and were also not required to rate all 50 speech samples. The samples presented contained no duplicates, so every participant only gave a single rating per sample. Of course, each sample was rated by multiple

participants to increase reliability. For each speech sample, participants had to indicate how native-like each speech sample was. This question was answered using a 7-point Likert scale (ranging from 1: very foreign sounding to 7: native American English speaker).

Via e-mail and social media we asked colleagues and friends to forward the link to the online survey to people they knew to be native American English speakers. In addition, the online survey was linked to in a post on *Language Log* by Mark Liberman.[2] Especially the latter announcement led to an enormous number of responses. As a consequence, we replaced the initial set of 50 speech samples five times with a new set to increase the number of speech samples for which we could obtain native-likeness ratings. As there was some overlap in the native American English speech samples present in each set (used to anchor the ratings), the total number of unique samples was not 300 (i.e. 6 times 50), but 286. Of these 286 samples, six samples were from native American English speakers (i.e. also included in the 115 samples discussed in Section 2.2). We think that the native speakers' function as "anchors" was sensible as an attempt to provide commensurability, and using only a few speakers made sure the anchors were relatively constant.

A total of 1143 native speakers of American English participated in the survey (658 men: 57.6%, and 485 women: 42.4%).[3] Participants were born all over the United States, with the exception of the state of Nevada. Most people came from California (151: 13.2%), New York (115: 10.1%), Massachusetts (68: 5.9%), Ohio (66: 5.8%), Illinois (64: 5.6%), Texas (55: 4.8%), and Pennsylvania (54: 4.7%). The participants were 36.2 years (SD: 13.9) on average and rated 41 samples on average (SD: 14.0). Despite the demographic spread, the sample of participants will be somewhat biased, as it consists of linguistically interested people (i.e. those who read *Language Log*). We doubt, however, that this affects the judgments much.

The analyses and results described in the following section may be reproduced with the paper package (including data, R analysis code, the graph and numerical results) belonging to this manuscript. The paper package can be obtained from the Mind Research Repository (http://openscience.uni-leipzig.de) or via the first author's website (http://www.martijnwieling.nl).

4. Results

In order to assess the consistency of the native-likeness judgments, we calculated Cronbach's alpha (Cronbach, 1951) on the basis of all individual ratings from all participants (i.e. a matrix with 1143 rows for all raters and 286 columns for all speech samples). The internal consistency was good with Cronbach's alpha equal to 0.853. This means that the native-likeness ratings given by the individual participants are generally highly correlated.

To find out how well the PMI-based LD matched with the native-likeness ratings, we calculated the Pearson correlation r between the averaged ratings and the Levenshtein distances. For the 286 speech samples we found a correlation of $r = -0.77$, $p < 0.0001$. When using the log-transformed LDs, the correlation was even stronger: $r = -0.81$, $p < 0.0001$. The direction is negative as the participants indicated how *native-like* each sample was, while the LD indicates how *foreign* a sample is. Figure 1 shows the

[2] http://languagelog.ldc.upenn.edu, May 19, 2012, "Rating American English Accents."
[3] Some participants may have participated multiple times. However, this number is likely limited, as the number of unique IP addresses (1135) was close to the number of participants (1143).

scatterplot (including the trend line) of native-likeness as a function of the logarithm of the Levenshtein distance. Note that the influence of the distribution of native languages used to determine the PMI-based sound segment distances was rather limited. When excluding the pronunciations of Spanish and French speakers (about 10% of the complete dataset) and determining the PMI-based pronunciation distances anew, these correlated $r = 0.99998$ with the original PMI-based pronunciation distances. The reason for this small influence is that the differences at the segment level are being smoothed out by averaging over all segments in a word and all 69 words in the paragraph. Obviously, when evaluating differences at the segment or word level, one may expect larger differences.

It is important to note that the high correlation between LD and the native-likeness ratings is close to how well individual raters agree with the average native-likeness ratings (on average: $r = 0.84$, $p < 0.0001$; the ratings for the individual rater are excluded from the average when calculating the correlation). Consequently, the LD-based method is not very different from a human rater, despite ignoring *inter alia* suprasegmental pronunciation differences (such as intonation).

Given these results, we claim that the automatically obtained LDs are a valid means to assess foreign accent strength in pronunciation.

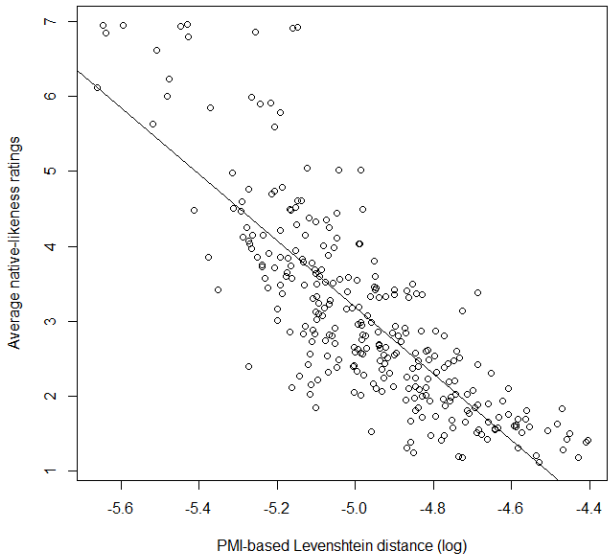

Figure 1. Logarithmically-corrected PMI-based Levenshtein distance as a predictor of mean native-likeness ($r = -0.81$). See text for discussion.

Each point in Fig. 1 pairs the LD measure of non-native-likeness with the mean judgment of the respondents in our survey. Points far to the left represent very low LDs, which increase as one moves to the right on the *x*-axis. Vertically low points were judged to be very unlike native speech, and the similarity to native speech increases as one moves up the *y*-axis. Examining the scatterplot more closely, we note that the cloud of points in the upper left of the graph deviates from the trend line; for these points in the upper left, the LD tends to underestimate how native-like speech samples are when the differences to native pronunciation are judged to be small. As the number of native speakers in the dataset is much lower than the number of non-native speakers of English, the sound correspondences among native speakers will be relatively infrequent, resulting in higher PMI-based substitution costs and greater LDs, which may

explain the greater deviations from the regression line. An alternative explanation might be that native-like suprasegmental qualities might counteract (small) segmental differences where these are present, leading to judgments of more native-likeness even where segmental phonology might be (slightly) non-native-like.

If the measure correlates with human judgments at the absolute level of $r = 0.8$, then it accounts for a good deal, but not all of the variance in the comparison ($r^2 = 0.64$). There are two important candidates to explain the remaining 36%. The first is the suprasegmental information, which we have systematically ignored (see above). The second is the transcription process. While the transcription quality of the Speech Accent Archive seems excellent, we do not know how high its transcriber agreement is, and the fact remains that transcription is always a difficult and error-prone task.

5. Conclusions and Discussion

We used a large set of transcribed data from non-native speakers of English who read the same paragraph aloud (Weinberger and Kunath, 2011), and used the Levenshtein distance to measure how much the transcriptions of non-native speech differed from those of native American English speech. In particular, we used a version of LD which employs automatically induced segment distances, introduced by Wieling et al. (2009), and normalized for alignment length. We collected judgments of native-likeness from over 1,100 native American English speakers and showed that their mean judgments correlated strongly with the logarithmically corrected computational pronunciation distance measure ($r = -0.81$). Given that this correlation is close to the agreement of individual human raters with the mean human judgments, the Levenshtein measure may serve as a proxy for human judgments of non-native-likeness.

As indicated before, the sample of raters is somewhat biased toward linguistically interested participants. However, given that a previous LD validation study with respect to dialectal pronunciation variation also revealed a high correlation (Gooskens and Heeringa, 2004), we do not believe this to be problematic.

One further task is clear, following work such as Brennan and Brennan (1981) and Magen (1998), namely to investigate what sorts of factors predict non-native-likeness while taking into account a large group of non-native speakers with various language backgrounds. A second task would be to investigate refinements of the Levenshtein distance in order to develop a technique even better able to gauge pronunciation differences, perhaps focusing on ways to include both subsegmental and suprasegmental information, or on ways of incorporating the fine-grained information present in the diacritics.

Acknowledgments

In addition to Mark Liberman and the *Language Log* participants mentioned above in Section 2.3, we benefited from astute and generous criticism by two anonymous LDC referees. We also thank Mark for his helpful comments during a presentation of this work in July 2013 in Groningen.

References

Brennan, Eileen M. and John S. Brennan. 1981. Measurements of accent and attitude toward Mexican-American speech. *Journal of Psycholinguistic Research* 10(5): 487-501.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22-29.

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3): 297-334.

DeKeyser, Robert M. 2012. Age effects in second language learning. In Susan Gass and Alison Mackey (eds.), *Handbook of Second Language Acquisition,* 442-460. London: Routledge.

Derwing, Tracey M. and Murray J. Munro. 2009. Putting accent in its place: Rethinking obstacles to communication. *Language Teaching* 42(4): 476-490.

Flege, James E. 1984. The detection of French accent by American listeners. *The Journal of the Acoustical Society of America* 76: 692-707.

Gooskens, Charlotte and Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16(3): 189-207.

Gooskens, Charlotte, Karin Beijering, and Wilbert Heeringa. (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing* 2(1-2): 63-81.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In John Nerbonne and Erhard Hinrichs (eds.), *Proceedings of the Workshop on Linguistic Distances,* 51-62. Sydney: Association for Computational Linguistics.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42(3-4): 331-354.

Jäger, Gerhard. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2): 245-291.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics,* 60-66. Dublin: Association for Computational Linguistics.

——— 2001. *The Significance of Word Lists*. Stanford: CSLI Publications.

Kalin, Rudolf, Donald S. Rayko, and Norah Love. 1980. The perception and evaluation of job candidates with four different ethnic accents. In Howard Giles, W. Peter Robinson, and Philip M. Smith (eds.), *Language: Social Psychological Perspectives,* 197-202. Oxford: Pergamon Press.

Kondrak, Grzegorz. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of NAACL 2: Second Meeting of the North American Chapter of the Association for Computational Linguistics,* 1-8. Stroudsburg, PA: Association for Computational Linguistics.

Kondrak, Grzegorz and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *COLING 04: Proceedings of the 20th International Conference on Computational Linguistics,* 952-958. Stroudsburg, PA: Association for Computational Linguistics.

Kruskal, Joseph. 1983 [1999]. An overview of sequence comparison. In David Sankoff and Joseph Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The*

Theory and Practice of Sequence Comparison, 1-44. Reprinted with a foreword by John Nerbonne. Stanford, CA: CSLI Publications.
Laver, John. 1994. *Principles of Phonetics*. Cambridge: Cambridge University Press.
Levenshtein, Vladimir. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR* 163: 845-848. In Russian.
List, Johann-Mattis. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH,* 117-125. Avignon, France: Association for Computational Linguistics.
Magen, Harriet S. 1998. The perception of foreign-accented speech. *Journal of Phonetics* 26: 381-400.
Major, Roy C. 1987. English voiceless stop production by speakers of Brazilian Portuguese. *Journal of Phonetics* 15(2): 197-202.
McCullough, Elizabeth A. 2013. *Acoustic Correlates of Perceived Foreign Accent in Non-Native English.* PhD dissertation, The Ohio State University.
McMahon, April, Paul Heggarty, Robert McMahon, and Warren Maguire. 2007. The sound patterns of Englishes: Representing phonetic similarity. *English Language and Linguistics* 11(1): 113-142 (doi: 10.1017/S1360674306002139).
Müller, André, Viveka Velupillai, Søren Wichmann, Cecil Brown, Pamela Brown, Eric W. Holman, Dik Bakker, Oleg Belyaev, Dmitri Egorov, Robert Mailhammer, Anthony Grant, and Kofi Yakpo. 2009. ASJP World Language Tree: Vers. 1 (April 2009). Available at http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm.
Nerbonne, John and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In John Coleman (ed.), *Workshop on Computational Phonology,* 11-18. Madrid: Association for Computational Linguistics.
——— 2010. Measuring dialect differences. In Peter Auer and Jürgen Erich Schmidt (eds.), *Language and Space. An International Handbook of Linguistic Variation. Vol 1: Theories and Methods,* 550-567. Berlin: Mouton De Gruyter.
Piske, Thorsten, Ian R. A. MacKay, and James E. Flege. 2001. Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics* 29(2): 191-215.
Sanders, Nathan C. and Steven B. Chin. 2009. Phonological distance measures. *Journal of Quantitative Linguistics* 43: 96-114.
Schepens, Job, Ton Dijkstra, and Franc Grootjen. 2012. Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition* 15(1): 157-166.
Weinberger, Steven H. and Stephen A. Kunath. 2011. The Speech Accent Archive: Towards a typology of English accents. In John Newman, R. Harald Baayen, and Sally Rice (eds.), *Corpus-based Studies in Language Use, Language Learning, and Language Documentation,* 265-281. Amsterdam/New York: Rodopi.
Wichmann, Søren. 2008. The emerging field of language dynamics. *Language and Linguistics Compass* 2(3): 442-455.
Wieling, Martijn. 2012. *A Quantitative Approach to Social and Geographical Dialect Variation.* PhD dissertation, University of Groningen.
Wieling, Martijn, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from dialect variation. *Journal of Phonetics* 40(2): 307-314.
Wieling, Martijn, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise alignment of pronunciations. In Lars Borin and Piroska Lendvai (eds.), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities and Education,* 26-34. Athens: EACL.