PREDICTING INTELLIGIBILITY FROM PRONUNCIATION DISTANCE METRICS

Tessa Bent and Rachael F. Holt

Indiana University; Ohio State University
tbent@iu.edu; holt.339@osu.edu

ABSTRACT

Unfamiliar native and non-native accents can cause word recognition challenges, particularly in noisy environments, but few studies have incorporated quantitative pronunciation distance metrics to explain intelligibility differences across accents. Here, intelligibility was measured for 18 talkers -- two from each of three native, one bilingual, and five non-native accents -- in three listening conditions (quiet and two noise conditions). Two variations of the Levenshtein pronunciation distance metric, which quantifies phonemic differences from a reference accent, were assessed for their ability to predict intelligibility. An unweighted Levenshtein distance metric was the best intelligibility predictor; talker accent further predicted performance. Accuracy did not fall along a native - non-native divide. Thus, phonemic differences from the listener's home accent primarily determine intelligibility, but other accent-specific pronunciation features, including suprasegmental characteristics, must be quantified to fully explain intelligibility across talkers and listening conditions. These results have implications for pedagogical practices and speech perception theories.

Keywords: Speech perception, intelligibility, non-native accents, regional accents

1. INTRODUCTION

Theories of speech perception must account for how listeners can recognize words amidst the variability present in speech signals [1, 2]. One factor that introduces substantial variability across talkers is a talker's accent, stemming either from regional differences or from influences of the first language when communicating in a second language. When these accent variants are unfamiliar, they can cause word recognition decrements [3, 4]. The challenge for understanding unfamiliar accents can be particularly acute in cases in which communication is occurring in the presence of background noise [5-8]. Although poorer accuracy has been observed for both unfamiliar native and non-native varieties, some research suggests that the pronunciation patterns found in unfamiliar non-native accents are more challenging to overcome than those in unfamiliar native varieties [9, 10]. However, most studies that include both native and non-native variants have been limited by including a small number of unfamiliar accents (e.g., [9, 10]). Therefore, it is difficult to determine the generalizability of the findings. [11] included a much wider range of accents but was limited by including only one talker per accent. Thus, there was a confound between the talker and accent variables. Studies that include both a wider number of accent varieties and multiple talkers representing each variety can provide greater insight into the factors that cause word recognition difficulties across native and non-native accents.

One open question in these studies is what specific talker or accent characteristics are leading to word recognition decrements. There is not a consensus regarding methods for quantifying how the speech differs from the listener's home accent. Researchers have taken different approaches to this problem. One approach is to focus on the impact of very specific production features (e.g., VOT differences [12]). Another approach is to synthetically modify the speech so that specific dimensions are more or less native-like [13, 14]. Work using this approach suggests that while keeping phonemic properties constant, changing the rhythmic properties of non-native speech to be more native-like increases intelligibility [13] but changing intonation patterns to be more native-like does not [14]. These approaches have provided insight into which dimensions of speech impact intelligibility for non-native talkers, but they do not allow for understanding how naturally produced interactions among multiple phonemic differences or between segmental and suprasegmental features may impact intelligibility. Finally, some researchers have incorporated general descriptions of the accents included (e.g., vowel differences across regional variants [8] or descriptions of general phonemic and suprasegmental differences across varieties [9]) but fell short of incorporating quantitative pronunciation metrics into their statistical modelling.

A recent approach to understanding how phonemic variability across accents impacts speech perception has been to incorporate edit-distance metrics into investigations of non-native accent strength ratings [15-17] and intelligibility [11, 18]. All these metrics measure pronunciation distance for

a speaker relative to a reference set (e.g., distance of a non-native talker to a specific native variant) using phonemic transcription. Several variations have been evaluated in relation to non-native accent strength ratings including Levenshtein Distances, Native Discriminative Learning (NDL) and Pointwise Mutual Information (PMI) metrics [15-17]. These metrics have shown substantial promise in accounting for variance in accent strength ratings across talkers. An advantage of the NDL and PMI metrics is that they incorporate the frequency of pronunciations within a corpus such that there are greater penalties for less-frequent pronunciation patterns, for which listeners are likely to assign higher accent strength ratings. However, these metrics require large amounts of data to calculate (e.g., hundreds of talkers), limiting their utility for most researchers.

Few investigations have incorporated these pronunciation distance metrics into studies of intelligibility. One study used the Levenshtein distance metric to characterise the talkers' utterances but did not incorporate the distance scores into the statistical models [18]. [11] incorporated the Levenshtein distance metric to model intelligibility of seven different accents. Levenshtein scores were a significant predictor of intelligibility in both quiet and noise-added conditions. Furthermore, a model including both Levenshtein distances and talker accent was a better fit than the one using only the Levenshtein distances. However, they only used one talker per accent, thus confounding talker- and accent-specific effects. Here, we build on [11] by including two more accents than used previously and, more importantly, including two talkers per accent. Incorporating more than one talker per accent will begin to address whether variability in intelligibility can be traced to pronunciation features that are characteristics of a specific accent or whether effects that have been described as "accent effects" are talker-level effects. Finally, we evaluate two versions of the Levenshtein distance metric to determine whether the weighted version used in [11] and [18] explains more variability in intelligibility than an unweighted version used in [19].

2. METHOD

2.1. Participants

There were 370 American English monolingual listeners between the ages of 18 – 35 years (average = 26). All listeners had self-reported typical speech, language, and hearing abilities. Midland American English was familiar to all participants due to its similarity to Standard American English. Participants who reported daily exposure to one of the other accents in their condition were excluded.

2.2. Stimuli

Sixty sentences from the Hearing in Noise Test for Children [20] were produced by 18 talkers representing 9 different accents: three native (Midland American English, Southern Standard British English, Scottish English), five non-native (French-, Spanish-, German-, Japanese-, and Mandarin-accented English), and one bilingual (Hindi-Indian English) variety. Two speakers (1 female and 1 male) represented each variety.

2.3. Procedure

Listeners were recruited through Prolific (https://www.prolific.co/) and tested online, using Pavlovia [21]. Prior to the intelligibility experiment, participants completed a consent form, a background questionnaire, and a headphone screening [22]. Listeners were presented with sentences from three talkers of the same gender including a Midland American English talker and talkers with two less familiar accents, each contributing 20 sentences. Listeners were randomly assigned to one of three listening conditions: quiet, +4 dB signal-to-noise ratio (SNR), or 0 dB SNR. The noise was an 8-talker babble with talkers matched in gender to the target speech. A randomly selected section of the babble file that was 1 second longer than the sentence was selected as the masker for each item. For each accent / listening condition combination, 15–18 participants were tested.

Listeners were presented with 9 practice trials, including three from each talker in the experimental trials. These were presented in the same listening condition as the experimental trials. Experimental trials were blocked by talker and randomized within a block. Participants were instructed to listen to a sentence and type in what they heard. They were not provided with any feedback.

2.4. Analysis

2.4.1 Levenshtein distances

All sentences were phonemically transcribed by two trained research assistants. These transcriptions were compared, and any discrepancies were resolved by discussion with a third transcriber. The transcriptions for the non-Midland speakers were compared to four talkers representing the familiar Midland American English referent for the calculation of two different versions of the Levenshtein distance metric: weighted and unweighted. In the unweighted version [19], the

target stimulus is compared to the reference and any phonemic difference between the two (i.e., addition, deletion, substitution) is given a one-point penalty. These penalties are then summed and divided by the total possible number of phonemes. This total is determined by the maximum logical alignment of the two strings (Table 1). The weighted Levenshtein distance [18] assigns different penalties depending on the type of difference between the two strings which are then summed at the word level:

- Vowel substituted by a vowel = 0.5
- Consonant substituted by a consonant = 0.75
- Phoneme insertion = 1.0
- Change to word length = 1/log10(max(length(Word1), length(Word2)))
- Other (e.g., deletion, vowel to consonant substitution, consonant to vowel, etc.) = 0.4

These weightings were developed based on theoretical assumptions from prior literature, but this metric has not been compared to other Levenshtein variants to determine whether these weightings add further explanatory power.

Levenshtein scoring	
Target word	*stopped*
Midland American accent	Ø s t ɑ p t
	| | | | | |
Spanish-accented English	ɛ s t ʌ p Ø
Unweighted Levenshtein scores	1 0 0 1 0 1
Final score	3/6 or .5
Weighted Levenshtein scores	1 0 0 .5 0 .4
Final score	1.9

Table 1: Unweighted and weighted Levenshtein scoring examples

2.4.2 Intelligibility scoring

Intelligibility scores were determined at the sentence level using Fuzzy String Matching, specifically the token sort ratio from [23] using the online implementation (https://tokensortratio.netlify.app). These scores range from 0 to 100. Sentences that match the target exactly are given a score of 100 and those without any matching characters are given a 0. For example, for the target sentence "Mother read the instructions", a response of "motherly subscriptions" received a score of 36, "follow the instructions" 63, "read the instructions" 86 and "mother read the instruction" a 98. These scores tend to be higher than percent words correct scores with a strict scoring criterion in which the examples above would have received scores of 9, 50, 75, and 75, because responses are given credit for partial matches. However, [23] showed that scores obtained via Fuzzy String Matching are highly correlated with traditional hand calculated measures of percent words correct.

3. RESULTS

3.1 Weighted vs. unweighted Levenshtein scores

We first evaluated whether the weighted or unweighted Levenshtein distance scores were a better predictor of intelligibility. Intelligibility score for each sentence was entered into a linear mixed effects model with fixed effects of Levenshtein score and SNR and their interaction with random intercepts for participant and item as well as by-participant and by-item varying slopes for talker accent. Two models were built, one with the weighted and one with the unweighted Levenshtein scores. From Type III Wald chi-square tests, both models showed significant effects of the Levenshtein scores, SNR, and their interaction (all p-values < 0.001). Therefore, we computed model comparisons to determine which Levenshtein score resulted in a better fit for the intelligibility scores. The model comparisons showed that the AIC was lower for the model with the unweighted (179969) than the weighted (179972) Levenshtein scores; thus, the unweighted Levenshtein scores were a better fit for predicting intelligibility. In further modelling, we employed the unweighted scores.

3.2 Levenshtein scores and talker accent

We next investigated whether talker accent contributed to intelligibility scores (Figure 1). In this model, we included fixed effects of unweighted Levenshtein scores, SNR, and talker accent, as well as their interactions. The model also included random intercepts for participant and item as well as by-participant and by-item varying slopes for talker accent. The results are shown in Table 2.

	F-value	*p*-value
Levenshtein	35.0	< 0.001
SNR	153.9	< 0.001
Accent	23.2	< 0.001
Levenshtein x SNR	2.1	n.s.
Levenshtein x Accent	11.3	< 0.001
SNR x Accent	29.7	< 0.001
Levenshtein x SNR x Accent	2.9	< 0.001

Table 2: Output of Type III Analysis of Variance Table with Satterthwaite's method for full model of intelligibility

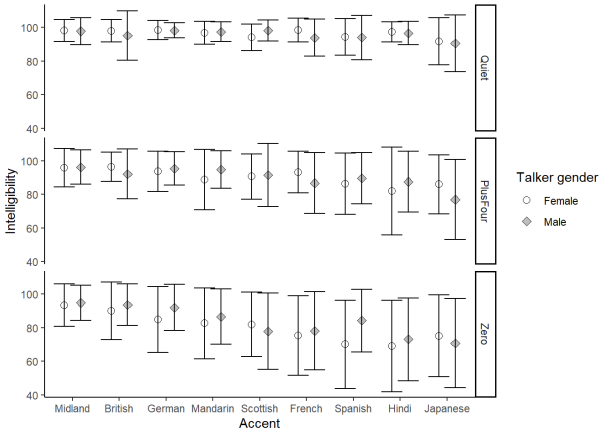

Figure 1: Intelligibility scores by accent and talker across the three listening conditions.

All three main effects were significant. The main effect of Levenshtein scores arose because items with higher Levenshtein scores were less intelligible than those with lower Levenshtein scores. That is, productions that diverged more from the local accent were more difficult for listeners to understand than those closer to the local dialect. The main effect of SNR resulted from highest accuracy in quiet and lowest accuracy in the 0 dB SNR with intermediate performance in the +4 dB SNR condition. The main effect of accent was due to the differences in intelligibility across accents with highest accuracy for the Midland American English and the Southern Standard British English accents and lowest accuracy for the Hindi and Japanese accents. The two-way interaction between Levenshtein distance and SNR was not significant suggesting that the effect of pronunciation distance from the local accent was similar across listening conditions. The other two-way and the three-way interaction were significant. The SNR by Accent interaction arose because some accents were highly intelligible even in the most difficult SNRs (e.g., Midland and British) whereas other accents showed much larger intelligibility declines, particularly in the most difficult SNR. The Levenshtein by Accent interaction and the three-way interaction among Levenshtein, Accent, and SNR suggest that effects of phonemic distance from the local accent were not consistent across accents and further, that these impacts were influenced by listening environment.

4. DISCUSSION

This study investigated intelligibility in quiet and two noise-added conditions across nine accents with two talkers representing each accent. The explanatory value of two variants of the Levenshtein distance metric, which measures phonemic differences from a reference accent, were evaluated. Results showed that higher Levenshtein distance scores were associated with lower intelligibility. Further, the unweighted Levenshtein metric was a better fit for the data than the weighted one. This result further supports the inclusion of these pronunciation distance metrics into studies of intelligibility to capture how differences in segmental productions across both unfamiliar native and non-native accents can impact intelligibility. They also suggest that relatively simple metrics (i.e., unweighted scores that are not adjusted for frequency of pronunciations within a large corpus) may be sufficient to capture the impact of phonemic differences on intelligibility. That said, the continued evaluation of other distance metrics (e.g., PMI, NDL) would still be valuable to determine the extent to which they can capture variance in intelligibility across talkers and accents, beyond their utility with explaining accent strength ratings [15-17]. This study extends the use of these pronunciation distance metrics to both native and non-native varieties whereas most previous studies (except [11]) used only non-native varieties. These results also cast doubt on any strong claims about non-native accents causing more disruption to word recognition than unfamiliar native ones.

These distance measures do not capture any pronunciation differences beyond the phoneme level, such as prosodic differences. Fully explaining why a specific talker or accent causes word recognition challenges, particularly for stimuli longer than a word, will very likely require measures of prosodic characteristics, such as rhythm, intonation, and speaking rate. Indeed, results in the full model suggest that intelligibility is impacted by interactions among phonemic distance, accent features, and listening conditions. Establishing standardized measures of prosodic distance from the home accent will be an essential next step in explaining the source of the accent effects. Similarly, a finer-grained analysis is necessary to interpret the impacts of listening condition (i.e., SNR). As has been seen for regional American English accents [8] and non-native accents [7], intelligibility differences across varieties tend to be much larger in adverse listening conditions. Additional analyses of the current data set will allow for pinpointing which pronunciation characteristics underlie the challenges that arise only in noisy environments. Finally, future studies should also include other social or sociolinguistic factors, such as listener familiarity with accents or language attitudes, to determine their impact on communication success.

5. ACKNOWLEDGEMENTS

We would like to thank our research assistants who supported this project: Lindsey Altum, Payton Bastie, Jessica Bell, Megan Hancock, Malachi Henry, Holly Lind-Combs, Yi Liu, Alondra Rodriguez, Ali Stallons, and Amy Warrington. We are grateful for the contributions of Lilian Golzarri Arroyo, who provided statistical consulting support, and for our funding from the National Science Foundation (Award Numbers: 1941691 and 1941662).

6. REFERENCES

[1] Pierrehumbert, J. B. 2016. Phonological representation: Beyond abstract versus episodic. *Annu. Rev. Linguist*. 2, 33–52.

[2] Kleinschmidt, D. F., Jaeger, T. F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev*. 122, 148–203.

[3] van Wijngaarden, S. J., Steeneken, H. J. M., Houtgast, T. 2002. Quantifying the intelligibility of speech in noise for non-native listeners. *J. Acoust. Soc. Am*. 111, 1906–1916.

[4] McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A. 2018. Van Engen, K. J. Coping with adversity: Individual differences in the perception of noisy and accented speech. *Atten. Percept. Psycho*. 80, 1559–1570.

[5] Rogers, C. L., Dalby, J., Nishi, K. 2004. Effects of noise and proficiency on intelligibility of Chinese-accented English. *Lang. Speech* 47, 139–154.

[6] Munro, M. J. 1998. The effects of noise on the intelligibility of foreign-accented speech. *Stud. Second Lang. Acq*. 20, 139–154.

[7] Wilson E. O., Spaulding, T. J. 2010. Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English. *J. Speech Lang. Hear. R*. 53, 1543–1554.

[8] Clopper, C. G., Bradlow, A. R. 2008. Perception of dialect variation in noise: Intelligibility and classification. *Lang. Speech* 51, 175–98.

[9] Bent, T., Baese-Berk, M., Borrie, S. A., McKee, M. 2016. Individual differences in the perception of regional, nonnative, and disordered speech varieties. *J. Acoust. Soc. Am*. 140, 3775–3786.

[10] Adank, P., Evans, B. G., Stuart-Smith, J., Scott, S. K. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol. Human* 35, 520–529.

[11] Bent, T., Holt, R. F., Van Engen, K. J., Jamsek, I. A., Arzbecker, L. J., Liang, L., Brown, E. 2021. How pronunciation distance impacts word recognition in children and adults. *J. Acoust. Soc. Am*. 150, 4103–4117.

[12] Sumner, M. 2011. The role of variation in the perception of accented speech. *Cognition* 119131–136.

[13] Tajima, K., Port, R., Dalby, J. 1997. Effects of temporal correction on intelligibility of foreign-accented English. *J. Phonetics* 25, 1–24.

[14] Sereno, J., Lammers, L., Jongman, A. 2016. The relative contribution of segments and intonation to the perception of foreign-accented speech. *Appl. Psycholinguist.* 37, 303–322.

[15] Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., Nerbonne, J. 2014. Measuring foreign accent strength in English. *Lang. Dynamics Change* 4253–269.

[16] Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., Baayen, R. H. 2014. A cognitively grounded measure of pronunciation distance. *PloS One* 9, e75734.

[17] Bartelds, M., Richter, C., Liberman, M., Wieling, M. 2020. A new acoustic-based pronunciation distance measure. *Fr. Art. Int.* 3, 39.

[18] Levy, H., Konieczny, L., Hanulíková, A. 2019. Processing of unfamiliar accents in monolingual and bilingual children: Effects of type and amount of accent experience. *J. Child Lang.* 46, 368–392.

[19] Gooskens, C., Heeringa, W. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Lang. Var. Change* 16, 189–207.

[20] Nilsson, M., Soli, S. D., Gelnett, D. J.1996. *Development of the Hearing in Noise Test for Children (HINT-C)*. House Ear Institute.

[21] Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. K. 2019. PsychoPy2: Experiments in behavior made easy. *Behav. Res.* 51, 195–203.

[22] Woods, K. J. P., Siegel, M. H., Traer, J., McDermott, J. H. 2017. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psycho.* 79, 2064–2072.

[23] Bosker, H. R. 2021. Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies *Behav. Res*. 53, 1945–1953.