Talker versus dialect effects on speech intelligibility: a symmetrical study

Daniel R. McCloy* and Richard A. Wright

University of Washington, Seattle WA

Pamela E. Souza

Northwestern University and Knowles Hearing Center, Evanston IL

Running head: Talker vs. dialect intelligibility

*Author to whom correspondence should be addressed. Electronic mail: drmccloy@uw.edu. Postal mail: Daniel R. McCloy, University of Washington, Institute for Learning & Brain Sciences, Box 357988, Seattle WA, 98195-7988.

Acknowledgments: This research was supported by the National Institutes of Health, National Institute on Deafness and Other Communication Disorders [grant R01-DC006014]. The authors are grateful to Namita Gehani, Jennifer Haywood and August McGrath for help in stimulus preparation, and to members of the UW Phonetics laboratory and anonymous referees for helpful suggestions on earlier drafts of this paper. Portions of the research described here were previously presented at the 164th Meeting of the Acoustical Society of America in Kansas City, MO, and published in *Proceedings on Meetings in Acoustics*.

ABSTRACT

This study investigates the relative effects of talker-specific variation and dialect-based variation on speech intelligibility. Listeners from two dialects of American English performed speech-in-noise tasks with sentences spoken by talkers of each dialect. An initial statistical model showed no significant effects for either talker or listener dialect group, and no interaction. However, a mixed-effects regression model including several acoustic measures of the talker's speech revealed a subtle effect of talker dialect once the various acoustic dimensions were accounted for. Results are discussed in relation to other recent studies of cross-dialect intelligibility.

Keywords: speech intelligibility, dialect variation, speech perception in noise, acoustics, mixed models

INTRODUCTION

This study investigates the relative effects of talker-specific variation and dialect-based variation on speech intelligibility. Unlike some previous studies of cross-dialect speech perception, it involves a fully-crossed design (all dialects studied are represented among both the talkers and the listeners) and explicit control for talker-level contributions to intelligibility that could otherwise contribute to sample bias (making one dialect appear more or less intelligible due to dialect-irrelevant aspects of the talkers' pronunciations). A variety of acoustic measures are used to quantify differences in talker-intrinsic intelligibility that emerge in the study.

BACKGROUND

It is well established that the pronunciation of an utterance can vary significantly from talker to talker even when the content of the utterance and the communicative conditions are fixed (Black, 1957; Bond & Moore, 1994; Hood & Poole, 1980). Idiosyncratic pronunciation characteristics that underlie this variation are sometimes referred to as *indexical traits*, for which two broad classes of explanation have been proposed. Scholars following Abercrombie (1967) attribute indexical traits to a talker's linguistic, social, and life experiences such as dialect, sociolect, and context exposure (in addition to the idiosyncratic physiological and kinematic characteristics of each talker's vocal tract). These are typically treated as relatively constant influences on a talker's speech, although their manifestation in the speech signal is not necessarily "constant" in the sense of being observable at all points or time scales. In contrast, Silverstein (2003) and others view indexical traits as manifestations of the sociolinguistic, discourse, and pragmatic signaling behaviors that a particular talker engages in for a particular communicative task. These behaviors can be seen as varying from utterance to utterance as the communicative context changes.

Within these two broad sources of between-talker variation, dialect in particular has been much scrutinized, especially as it relates to speech intelligibility. For example, when listeners hear speech in their own dialect it is usually more intelligible than when it is in an unfamiliar dialect (Labov & Ash, 1997; Mason, 1946). Listeners perform increasingly poorly in vowel identification tasks when synthetic-vowel stimuli are modeled on dialects that are progressively divergent from their own (Wright & Souza, 2012). Similarly, Oder and colleagues (2013) demonstrated that vowel identification error patterns are consistent with vowel acoustic differences between dialects using naturally produced vowels. Listeners also perform worse on a variety of language-processing tasks when stimuli are drawn from an unfamiliar dialect than when stimuli are in a familiar dialect, and worse yet when the stimuli are in a non-native accent (Adank, Evans, Stuart-Smith, & Scott, 2009). However, listeners do show adaptation to unfamiliar dialects or accents through exposure or training (Baese-Berk, Bradlow, & Wright, 2013; Bent & Holt, 2013; Bradlow & Bent, 2008; Kraljic, Brennan, & Samuel, 2008); this suggests that intelligibility differences between dialects are at least in part, if not predominantly, a reflection of listener experience.

Nonetheless, a few studies have suggested that some dialects are intrinsically more intelligible than others. This is analogous to the many studies showing that talker-intrinsic pronunciation traits can lead to talker-intrinsic intelligibility differences (e.g., Bradlow, Torretta, & Pisoni, 1996; Payton, Uchanski, & Braida, 1994; Picheny, Durlach, & Braida, 1985). Two recent studies of dialect-level differences in intrinsic intelligibility will be reviewed here: Clopper and Bradlow's (2008) study of dialect intelligibility and classification in noise, and Jacewicz and Fox's (2014) study of dialect intelligibility in multitalker backgrounds.

Based on two earlier studies of dialect similarity and classification (Clopper, Levi, & Pisoni, 2006; Clopper & Pisoni, 2007), Clopper and Bradlow (2008) examined the intelligibility of four American English dialects: Mid-Atlantic, Northern, Southern, and General American (the last of which is a meta-dialect comprising speakers from the Midlands, Western U.S., and New England). Clopper and Bradlow found differences in intelligibility between the dialects increased in more challenging noise levels, with General American talkers being most intelligible and Mid-Atlantic talkers being least intelligible across a range of noise levels regardless of listener dialect. They interpret their results as indicating that General American is the most intrinsically intelligible dialect to their listeners, who were drawn from three groups: Northern, General American, and "mobile" (listeners who had lived in two or more dialect areas before the age of 18). Somewhat surprisingly, the General American talkers were more intelligible than the Northern talkers *even to Northern listeners*. This could be taken (somewhat implausibly) to indicate that Northern listeners were somehow more familiar with General American speech than they were with their own dialect, or (more plausibly) that the difference in intelligibility of the dialects (which favored General American) outweighed the difference in familiarity (which presumably favored Northern speech).

One limitation of the Clopper and Bradlow study is that the listener-talker groups weren't fully symmetrical. That is, while the Northern and General American dialects were represented in both the listener and talker groups, there were no Southern or Mid-Atlantic listeners and no "mobile" talkers. Therefore, it is difficult to definitively attribute intelligibility differences to intrinsic dialect traits rather than listener familiarity with those dialects. Indeed, part of Clopper and Bradlow's explanation for the intelligibility advantage of General American is that listeners are likely to have encountered it and be at least somewhat familiar with it. A second limitation is

that all listeners heard a practice set of 12 sentences from a different set of General American talkers before hearing the test stimuli. This could have resulted in a dialect-familiarization advantage for the stimulus-set of General American talkers. Of most relevance to the current study, there was very little effort to control for or measure intrinsic intelligibility of the talkers. While they did find that vowel space area and vowel space dispersion were not significantly different across dialects, there were intelligibility advantages for gender groups (differing by dialect) that were not predictable from the vowel space measures, indicating that there is at least some other dimension contributing to gender-based intelligibility effects. The study had a moderate number of talkers from each dialect group (three male and three female) but their statistical analysis does not include talker as random effect to see if there might be individual pronunciation differences within the set of regional talkers that may be unintentionally contributing to the apparent intelligibility advantage of the General American talkers.

In a more recent study, Jacewicz and Fox (2014) found that Southern talkers were more intelligible than General American talkers when the listener group was General American, and that the Southern-talker intelligibility advantage increased as listening conditions deteriorated. They also found that, in more difficult listening conditions, Southern talkers were more effectively masked by Southern-talker babble than they were by General American-talker babble, but that General American talkers were masked equally well by both dialects. The first result contradicts Clopper and Bradlow's (2008) findings (discussed above) showing that General American was most intelligible to all listener groups that they studied (which included General American listeners). Like the Clopper and Bradlow study, the Jacewicz and Fox study was also asymmetrical, with listeners only from the General American region.

As Jacewicz and Fox point out, the tasks in the two studies are not identical: the sentence material and the masker noise both differed between their study and Clopper and Bradlow's study. They also note that their General American speakers are drawn from a narrower geographic area than Clopper and Bradlow's General American area, so dialect homogeneity may play a role in explaining the difference in findings. Importantly, Jacewicz and Fox note that their Southern talkers produced speech with a pitch pattern that differed from the General American talkers despite producing speech at a comparable tempo. They attribute the differences in pitch to dialect-specific differences, but, like the Clopper and Bradlow study, Jacewicz and Fox do not conduct a detailed analysis of the talkers' signals to look for acoustic traits that might contribute to the intelligibility differences, nor do they use talker as a random variable in their statistical design. Therefore, it remains to be seen if their apparent dialect-intrinsic intelligibility effect favoring Southern talkers is a genuine dialect effect rather than an effect of the individuals who made up their samples of talkers.

In light of the questions raised by the two studies just discussed, the current study probes the relative contribution of dialect and individual pronunciation variability in a symmetrical two-dialect sample. It includes a detailed acoustic analysis of talker traits that are thought to contribute to intelligibility. The study involves talkers and listeners from two dialect regions with fairly narrow definitions: Pacific Northwestern (PN) English (a subregion of the General American variety used in both studies) and Northern Cities (NC) English (a subregion of the Northern region in Clopper and Bradlow's study).

METHODS

Stimulus Set

Stimuli were sentences drawn from the PN/NC corpus (McCloy et al., 2013), a subset of the IEEE "Harvard" sentences (Rothauser et al., 1969) read by five males and five females from each of two dialect regions: the Pacific Northwest (PN) and the Northern Cities (NC). The PN region was defined as Washington, Oregon, and Idaho, and is a sub-region of "the West" as defined in both Clopper et al (2005) and Labov et al (2006, p. 137). The NC region was defined following Labov et al (2006, pp. 121–124) as the sub-region of the "Inland North" that preserves the low-back distinction between /a/ and /ɔ/ in both production and perception. It is a sub-region of the "North" region described in studies by Clopper and colleagues (Clopper et al., 2006, 2005), who largely follow Labov et al (2006).

From the full set of 720 Harvard sentences, 200 were selected based on avoidance of contrast of focus readings, absence of alliteration or rhyming, and lack of marked locutions (e.g., "the juice of lemons" instead of "lemon juice"). Talkers were fitted with a head-mounted close-talking microphone (Shure SM10–A) to ensure consistent and maximal signal-to-noise ratio (SNR) of the raw recordings, and the 200-sentence block was recorded three times per talker. Talkers were coached to speak in a relaxed manner with no special effort or emphasis. Talkers exhibiting list intonation across sentences were notified of this behavior and coached to produce falling declarative intonation on every sentence. Three trained phoneticians chose the best instance of each sentence from each talker for inclusion in the corpus, determined by two criteria: lack of mic overloading or clipping, and absence of hesitations and disfluencies. All stimuli were hand-trimmed (with careful attention to low-amplitude edge phones such as [h], [f] and [θ]), padded with 50 ms of silence at the beginning and end, and RMS normalized. From the 200 sentences

recorded, 20 were reserved for task familiarization, yielding a final corpus of 3600 stimuli (5 talkers per group × 2 genders × 2 dialect regions × 180 sentences).

Perception Task

Each listener first heard 20 unique training sentences (one from each talker) presented at three SNRs (6 sentences in clear and 7 at each of +2 dB and +6 dB SNR levels). The training sentences were all distinct from the test sentences. Each listener then heard all 180 test sentences in the corpus (each containing five keywords), drawn in equal numbers from the 20 talkers. Sentences were presented in quiet and in two levels of background noise (+6 dB and +2 dB SNR). The masker in the noise conditions was gaussian noise filtered to match the long-term spectral average of the corpus. To ensure target audibility, the level of the speech was held constant at 68 dB SPL (dB RMS in a 6 cc coupler) and different levels of masker noise were digitally added to the speech to achieve the desired SNRs. The combined signal was presented in a sound-insulated booth over closed-back supra-aural headphones (Sennheiser HD 25–1 II). Listeners were instructed to repeat each sentence they heard, to give partial answers when they only heard some words, and to guess when they were unsure. Trials were scored 0–5 on keywords correct during the task. An audio recording was made of listener responses, and scoring uncertainties were resolved offline by a second researcher. The 900 keywords were all content words, with the following exceptions: 7 instances of pronouns (it, you, your, she, her, he, him) and 25 instances of prepositions (across, against, beside, into, from, off, under, when, with, without). 81% of the keywords were monosyllabic, the remaining 171 were disyllabic; no sentence had more than one disyllabic keyword. Talker-sentence-SNR assignments were random and unique for each listener, with the following constraints: (a) each listener heard 9 sentences

from each talker; (b) each listener heard each talker at all three SNRs (3 sentences at each SNR); (c) each listener heard each sentence only once.

Participants

Listeners were drawn from the same two dialect regions used to create the corpus: the Pacific Northwest (PN) and Northern Cities (NC). By chance, all PN listeners were natives of Washington state; NC listeners were natives of northern Ohio, northern Indiana, northern Illinois, and Michigan (see Figure 1). All listeners were required to have lived in-region for ages 5–18, and to have not lived more than 5 years total outside their region. The mean age of the listener group was 20.5 years for the PN listeners and 24.5 years for the NC listeners. All listeners had bilaterally normal hearing, defined as pure-tone thresholds of 20 dB HL or better at octave intervals from 250 Hz to 8 kHz (American National Standards Institute, 2004). Fifteen PN listeners and thirteen NC listeners participated in the perception task.

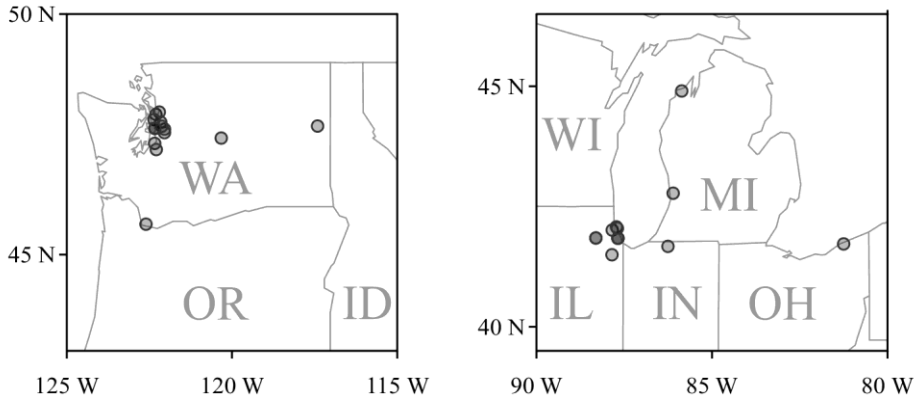

Figure 1. Hometown locations of all 28 listeners.

Acoustic measurements

In order to characterize difference between talkers, a variety of acoustic measures of the corpus were performed. Measures related to intensity, fundamental frequency (f_0) and several characteristics of the F2 × F1 vowel space were obtained. Speech rate (syllables per second) was

also calculated for each talker as the sum of the number of syllables in each sentence divided by the sum of the durations of each sentence.

Vowel space measures Vowel space characteristics for each of the talkers were calculated from the corpus materials based on hand-measurements of the first and second formants of 1100 vowel tokens (2 dialects × 10 talkers × 5 tokens per vowel for the 11 vowels /i ɪ e ɛ æ a ɔ o ʊ u ʌ/; for PN talkers the vowel /ɑ/ replaced both /a/ and /ɔ/, and had 10 measured tokens instead of 5 due to its correspondence to both /a/ and /ɔ/ of the NC talkers' vowel inventory). Measured vowels were drawn from lexically stressed syllables in keywords in positions throughout the sentence, with a preference for vowels with obstruent flanking consonants to avoid coloring by adjacent nasals, rhotics, or laterals. Selection of vowels to measure was balanced across vowel category, such that there were equivalent numbers of tokens from keywords early, middle, and late in the sentence for each vowel type. Vowel boundaries were marked following the methods of Peterson and Lehiste (1960): vowel onset was marked at the release burst of a preceding plosive or at the start of periodicity when preceded by a fricative, and vowel offset was marked at the cessation of periodicity. The hand-measured formant values were taken from the 50% point of each vowel and converted to a perceptual scale using the bark transform (Traunmüller, 1990) prior to statistical analysis.

Five measures of the vowel space were calculated from the formant data: mean Euclidean distance from the center of the vowel space (Bradlow et al., 1996), area of the polygon defined by F1 and F2 means for each vowel (Bradlow et al., 1996; Neel, 2008), area of the convex hull encompassing all measured vowel tokens (McCloy, 2013), total repulsive force of the vowel system (Liljencrants & Lindblom, 1972; Wright, 2004), and mean vowel cluster size (see Figure

2).[1] Calculation of the area of the vowel polygon differs from previous studies in being based on a large number of vowel phonemes (all vowels measured excluding /ʌ/), in contrast to the /i o a/ triangle used by Bradlow and colleagues (1996), or the /i æ ɑ u/ quadrilateral used by Neel (2008). The measurement of the area of the convex hull encompassing all vowel tokens is based on the idea that a polygon based on F1 and F2 means for each vowel contains information about a range of reduced and unreduced forms of each phoneme (and thus likely indexes a talker's prosodic habits to some degree), whereas a convex hull is a representation of the vowel space based on a talker's most extreme unreduced pronunciations, and thus might abstract away from individual differences in prosody (see McCloy, 2013, Chapter 5 for discussion).

Repulsive force (sometimes called "total energy" of the vowel system) was calculated as the sum of inverse squared distances between all pairs of vowel tokens not belonging to the same phoneme, as in Equation 1 (where /i/ and /j/ represent the phonemic categories of the vowel tokens being compared, and *r* is the Euclidean distance formula). This measures the degree to which neighboring vowel phonemes in a system encroach on one another, with higher values of repulsive force corresponding to greater degrees of phoneme overlap or encroachment. The calculation seen here differs from both Liljencrants and Lindblom (1972) and Wright (2004) in calculating force based on individual vowel tokens rather than mean values for each vowel.

[1] It is noteworthy that we chose not to include measures of F1 and F2 range in our models, since both have previously been reported to be significant predictors of intelligibility (Bradlow, Torretta, & Pisoni, 1996; Hazan & Markham, 2004). When we included these measures in our statistical model, the model failed to converge (possibly because F1 and F2 range are highly correlated with other measures of vowel space size, or possibly due to insufficient statistical power for the large number of predictors included). We chose to omit these rather than other measures because they are rather coarse measures of vowel space size that can easily be influenced by dialectal variation, depending on the vowels measured. For example, Hazan and Markham (2004) measured F2 range based only on tokens of /i/ and /u/, which could be strongly influenced by dialectal or gender differences in /u/-fronting (a known feature of PN speech, cf. Reed, 1952; Ward, 2003, Chapter 4).

$$\sum_{i=1}^{n-1} \sum_{j=n+1}^{n} \frac{1}{r_{ij}^2}, \qquad /i/\neq/j/$$

Equation 1. Formula used to calculate repulsive force of the vowel space. /i/ and /j/ represent vowel phoneme categories, and r_{ij} is the Euclidean distance formula $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

Finally, mean vowel cluster size was calculated for each talker as the mean of the areas of the 95% confidence ellipses for each vowel category (based on bivariate normal density contours). Low values of cluster size are associated with low degrees of within-category variation and therefore (we predict) a more predictable perceptual target and higher intelligibility. We are not aware of any previous studies of intelligibility that make use of this measure.

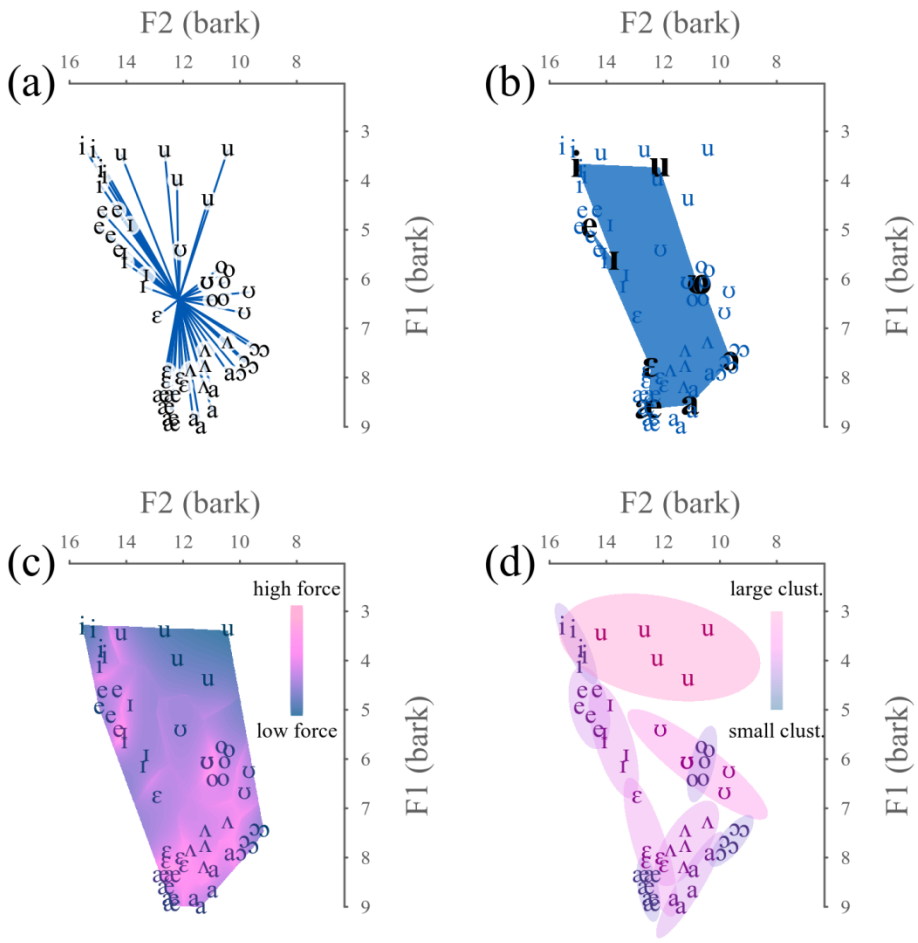

Figure 2. Illustration of acoustic measures of the vowel space used in the statistical models. (a) Mean Euclidean distance from center. (b) Area of the polygon formed by vowel means. (c) Repulsive force (color) and area of convex hull (shape). (d) Mean cluster size.

Intensity- and pitch-related measures Because stimuli were RMS normalized, mean intensity across stimuli is virtually identical, but the mean rate of change of intensity "intensity velocity") was calculated for each stimulus in hopes of capturing a talker's tendency to "trail off" at the ends of utterances, or conversely to maintain a more consistent level across all the keywords in the sentence. First, mean-subtracted intensity values were calculated using a gaussian moving window with 50 ms bandwidth and 13.3 ms step size (i.e., a window corresponding to a 60 Hz pitch floor and step size corresponding to 0.25 × effective window length). From these intensity values, the rate of change of intensity was calculated at each point by subtracting the intensity

value at the previous time point and dividing by the step size. Intensity velocity was defined as the mean of these rate-of-change values across the sentence, while "intensity dynamicity" was defined as the mean of the absolute value of the rate of change of intensity, as a measure of how dramatic the rises and falls in intensity were across each sentence, irrespective of overall intensity downtrend.

For measures of vocal pitch, f_0 tracks were automatically extracted using Praat (Boersma & Weenink, 2013) and a random subset of 15 of the 180 sentences were selected for hand-correction. This yielded a total of 300 f_0 tracks for data analysis (15 per talker × 20 talkers). From those 15 sentences the absolute and average f_0 range magnitudes were calculated for each talker,[2] as well as the mean rate of change in f_0 ("pitch velocity") and mean absolute value of rate of change in f_0 ("pitch dynamicity"). Like the intensity measures, pitch velocity indexes overall sentence-level changes in pitch, while pitch dynamicity indexes how dramatic the rises and falls in f_0 were across each sentence, irrespective of overall f_0 downtrend.

Dialect-related measurement issues Acoustic measures of the vowel space across talkers from the PN and NC regions involves some dialect-related complexities due to the presence of the low-back split preserved in the speech of NC talkers, and the corresponding merger of /a/ and /ɔ/ to a single phoneme /ɑ/ in the PN talkers. This affects the calculation of vowel-space-related measures that reflect the *relationships* among vowels in F2 × F1 space, such as polygonal vowel space area, mean cluster size, and repulsive force. To address this issue, measures of polygonal vowel space area, mean cluster size, and repulsive force were calculated based on the dialect

[2] The choice to use mean size of pitch range in addition to absolute pitch range was motivated by the fact that a given sentence may be uttered in a fairly monotone fashion even by a talker that has a large overall pitch range. Thus we reason that a talker's *typical* range across utterances is more indicative of their linguistic use of pitch than their *maximal* range. Ideally, pitch range would be a stimulus-level predictor rather than a talker-level aggregate, but reliable measures of pitch range for all 3600 stimuli was not possible given the need for hand correction (stemming primarily from the difficulty of automatic pulse detection and pitch tracking algorithms in dealing with creaky voicing).

region of the talker, and z-score normalized within dialect (rather than across all talkers) prior to statistical modeling. In this way, the measures reflect each talker's use of the vowel space relative to their within-dialect peers, thereby decoupling those predictors from other dimensions of speech that might be dialect-linked and therefore might co-vary with a vowel space measure that co-varied with talker dialect.

Data Analysis

Listener scores for keywords correct were modeled using mixed-effects regression with the lme4 package (Bates, Maechler, Bolker, & Walker, 2013) in R (R Development Core Team, 2013). As can be seen in Figure 3, the clear and +6 dB SNR conditions showed ceiling effects, and consequently were excluded from further analysis.

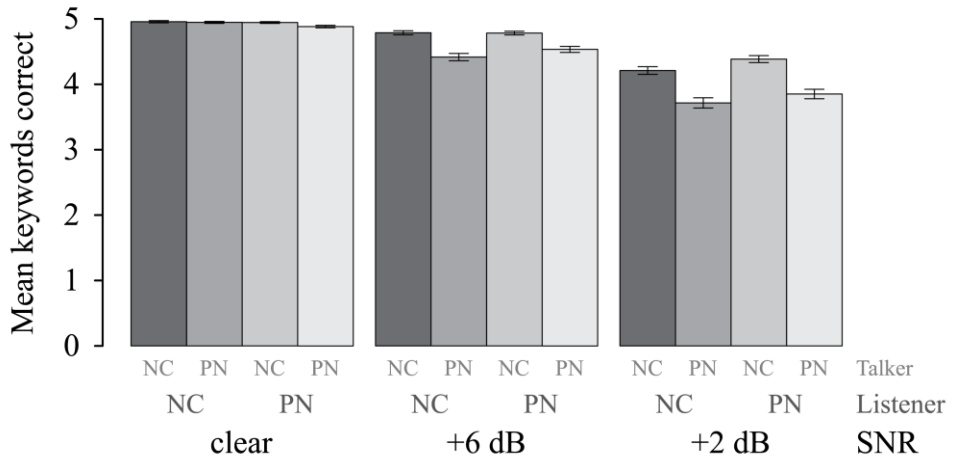

Figure 3. Mean keywords correct for the three SNR conditions, two talker groups, and two listener groups. Error bars show one standard error. Ceiling effects are evident in the clear and +6 dB conditions.

Although during the experimental session the sentences were scored 0-5 on keywords correct, scores were reduced to binary (1 = all keywords correct) prior to statistical analysis and were modeled using a logistic link function. This was done for two reasons. First, it is inappropriate to model a discretized 0–5 keyword score as continuous (because the data are restricted to six discrete values and bounded between zero and five, violating the model assumption that the

dependent variable is continuous and unbounded). Second, in this case modeling perception at the level of individual keywords is less desirable given that our acoustic predictors are all sentence- or talker-level measures. Another alternative, averaging across sentences within talker-listener pairs to better approximate a continuous outcome, was rejected because it did not allow modeling by-sentence random variation.[3]

As mentioned above, acoustic measures related to the internal structure of the vowel space (polygonal area defined by formant frequency means, mean cluster size, and repulsive force) were z-score normalized within each talker dialect group prior to statistical modeling. The other acoustic measures (mean distance from center of the vowel space, and measures related to intensity, f_0, and speech rate) are not known to systematically differ between these two dialects in ways that would affect their calculation, so those measures were z-score normalized across all talkers. A binary factor of talker gender was also included. The five acoustic measures of the vowel space were analyzed for collinearity by computing the condition index of the predictor matrix (also known as "kappa"). The kappa value was 8.9, indicating mild-to-moderate collinearity (kappa < 10 is a commonly used criterion for predictor inclusion).

RESULTS

The initial statistical model set out to test whether there was a (possibly asymmetrical) effect of cross-dialect listening on talker intelligibility. Model results are shown in Table 1; there is no significant interaction between talker dialect region and listener dialect region, nor are there significant main effects for talker or listener dialect region separately. Not surprisingly, the random effects show relatively high estimates of variance for sentence and talker (suggesting

[3] Although there is arguably an important difference in listener performance between 4 keywords correct and no keywords correct that is lost when converting scores to binary, this distinction may be less important with low-context stimuli such as the IEEE "Harvard" sentences used here (in which missed words are more difficult to recover from surrounding context, so that reporting 4 words correctly does not necessarily correspond to a listener getting the gist of the sentence's meaning).

that sentences varied in their difficulty and talkers varied in their intelligibility), and a relatively low estimated variance for listener (suggesting that by and large, listeners were equally good at the task). In fact, variation in intrinsic intelligibility of the talkers spanned a wide range, with the least intelligible talker averaging only 2.6 keywords correct per sentence in the hardest noise condition, and the most intelligible averaging 4.7 (see Figure 4).

Table 1. Summary of mixed-effects regression model testing the interaction between talker and listener dialects. S.E. = standard error; s^2 = estimated variance.

Fixed effects (N = 1680; log-likelihood = −1031.637)					*Random effects*	
Predictor	**Coefficient**	**S.E.**	**Wald *Z***	***p***	**Group**	**s^2**
Intercept	0.4759	0.3067	1.551	0.1208	Sentence	0.8575
Talker from PN	−0.6502	0.3972	−1.637	0.1016	Talker	0.6500
Listener from PN	0.4043	0.2152	1.879	0.0603	Listener	0.1408
TalkerPN:ListenerPN	−0.2626	0.2294	−1.145	0.2523		

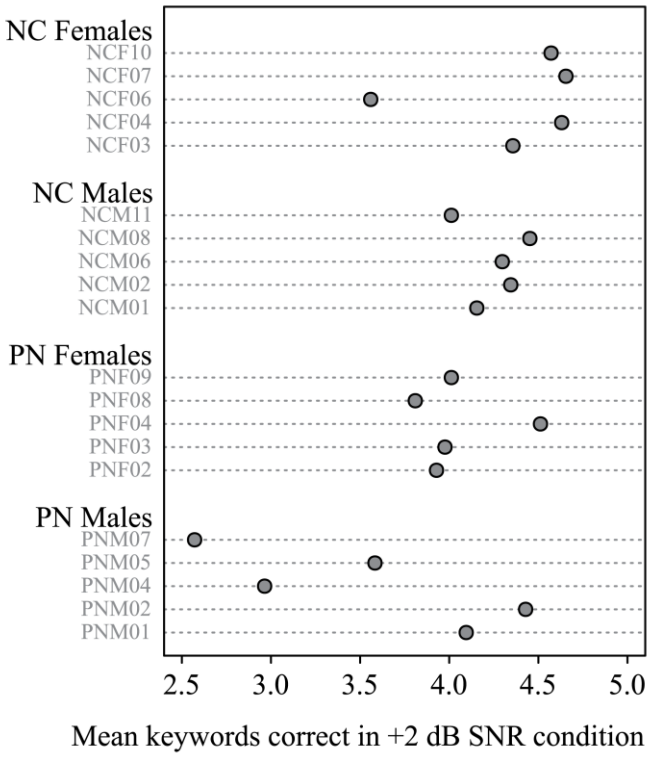

Figure 4. Dotchart of mean intelligibility by talker in the +2 dB SNR condition. Dramatic differences across the groups are evident in the horizontal spread within each group.

Accounting for talker variation using acoustic predictors

Given that there were no significant differences between the dialect groups overall, we ran a second statistical model that included a variety of acoustic predictors, to see if a more subtle dialect effect would emerge in a model that controlled for various acoustic dimensions of speech known to impact intelligibility. The results of that statistical model are summarized in Table 2.

Table 2. Summary of mixed-effects regression model testing the interaction of talker and listener dialects while controlling for several acoustic predictors. S.E. = standard error; s^2 = estimated variance.

Fixed effects (N = 1680; log-likelihood = −1004.420)					*Random effects*	
Predictor	**Coefficient**	**S.E.**	**Wald *Z***	***p***	**Group**	**s^2**
Intercept	−1.45545	0.40325	−3.609	0.000307	Sentence	0. 8868
Talker from PN	2.09153	0.47977	4.359	0.0000130	Talker	0.00001
Listener from PN	0.40793	0.21681	1.881	0.059906	Listener	0.1394
Talk.PN:List.PN	−0.26921	0.23273	−1.157	0.247380		
Talk. gender: male	1.13436	0.42235	2.686	0.007235		
Speech rate	−0.09063	0.19922	−0.455	0.649167		
Mean dist. center	−0.36363	0.28063	−1.296	0.195058		
Repulsive force	−0.53792	0.09024	−5.961	2.51×10^{-9}		
Mean v. clust. size	−0.80192	0.30576	−2.623	0.008723		
Convex hull area	1.86465	0.60866	3.064	0.002187		
V. means polygon	−0.53241	0.22967	−2.318	0.020442		
Overall pitch rng.	1.00874	0.26968	3.740	0.000184		
Mean pitch range	2.99947	0.66325	4.522	6.12×10^{-6}		
Pitch velocity	2.37561	0.52072	4.562	5.06×10^{-6}		
Pitch dynamicity	−1.55075	0.36468	−4.252	0.0000211		
Intensity velocity	−0.43184	0.10204	−4.232	0.0000231		
Intensity dynam.	0.49245	0.11678	4.217	0.0000248		

Examining the random effects, we see that the estimated variance across sentences and across listeners has not changed substantially from the previous model, but the estimated variance across talkers is dramatically smaller in the second model. This result is expected, since many of the aspects of the talkers' speech that differentiate them in intelligibility are now being accounted for by the acoustic predictors, rendering the random effect for talker virtually unnecessary.

More interestingly, the fixed effect for talker dialect has reversed direction: in the first model, a talker being from PN predicted *lower* intelligibility (i.e., the model coefficient was *negative*, though not significantly so), whereas in the model that controls for several acoustic predictors, a talker being from PN predicts *higher* intelligibility (i.e., the model coefficient is *positive*).

Vowel space measures Among the acoustic predictors related to the vowel space, we see mostly expected results: talkers are predicted to be more intelligible if they have larger convex hulls encompassing their vowel space, smaller mean vowel cluster sizes, and lower repulsive force (i.e., less phonemic crowding). Unexpectedly, we see a negative correlation between intelligibility and the size of the vowel polygon defined by F1 and F2 means. One possible explanation for this finding is that a polygon based on vowel means contains information about a range of reduced and unreduced forms of each phoneme. Given that in our data vowel tokens were all drawn from lexically stressed syllables in positions throughout the sentence, to the extent that there is any reduction present, it is likely to be prosodically driven rather than phonologically driven. As such, a large polygon might indicate that a talker is less effective at making use of the vowel space to differentiate prominent from non-prominent words, which might explain the negative correlation between intelligibility and the area of the polygon in this data.

Pitch-related measures Among the acoustic predictors related to pitch, we see mostly expected results: talkers are predicted to be more intelligible if they have a larger overall pitch range, larger average pitch range magnitude across sentences, and higher mean pitch velocity (i.e., less downdrift). An unexpected finding is that higher intelligibility is correlated with *lower* pitch dynamicity, implying that larger pitch excursions to mark important words is actually detrimental to intelligibility. However, another possibility is that the measure of pitch dynamicity

conflates mid-sentence pitch excursions with sentence-final drops into creaky voicing, which is likely to reduce intelligibility due to the concomitant drop in intensity during creaky phonation (Gordon & Ladefoged, 2001; cf. McCloy, 2013, pp. 83–86). In other words, the negative correlation between pitch dynamicity and intelligibility may be dominated by a single, large, utterance-final pitch excursion, masking the effect on intelligibility of earlier, smaller, prominence-marking pitch excursions. A thorough investigation of this hypothesis would require a more detailed measurement and transcription of pitch excursions than can be performed automatically, so for this study we must remain agnostic regarding the role of pitch dynamicity in intelligibility. Recall also that pitch measures were based on a random selection of 15 of the 180 stimulus sentences, and although we believe this to be a sufficient number to estimate the variation in these talkers' use of pitch, some readers may still view the pitch-related findings with some skepticism.

Intensity-related measures In this model, higher intelligibility is predicted by higher values of mean intensity dynamicity, and by lower (i.e., more negative) values of mean intensity velocity. The correlation with mean intensity dynamicity was expected, in that higher intensity dynamicity was hypothesized to correspond with more consistent use of intensity to mark prominence (and correspondingly to produce non-prominent words with reduced intensity). However, it is possible that this measure would be less effective in predicting intelligibility in a study where intelligibility were scored across all words in a sentence, rather than based on linguistically prominent keywords.

The negative correlation with intensity velocity is more difficult to explain. One possible explanation is that intensity that is more flat across the span of the sentence indicates less dramatic modulation of the vocal source by the articulators of the vocal tract, i.e., more

coarticulation, more frequent consonant lenition, etc. More research is necessary to understand the significance of this measure.

Talker gender The effect of talker gender on intelligibility is also statistically significant in this model, suggesting that, once the aforementioned speech parameters related to intensity, pitch, and the vowel space are accounted for, some aspect remains that makes the male talkers more intelligible than the females in this corpus. One possible locus of this difference is spectral properties of speech other than formant center frequencies, such as the bandwidths of the formants or the density of the harmonics of f_0 — both properties that are obscured when focusing on the F2×F1 vowel space defined only by the center frequencies of each formant. Another possibility is that consonant articulation varies systematically across gender groups in a way that impacts intelligibility in our sample of talkers. More research is needed to determine what factor(s) underlie this result.

DISCUSSION

This paper presents the results of a within- and across-dialect speech perception study involving talkers and listeners from the Pacific Northwest (PN) and Northern Cities (NC) dialects of American English. These dialects are known to differ phonolexically in several respects, but in general the differences are regarded as fairly subtle and the dialects are judged to be mutually intelligible (indeed, many of the listeners in our study did not even notice the presence of out-of-dialect speakers when debriefed after the perception task). Consistent with this view, our findings showed no systematic difference in intelligibility attributable to talker-listener dialect difference, but dramatic individual differences in talker intelligibility within talker dialect groups in the +2 dB SNR condition. There was a small difference in intelligibility between the talker populations (with PN speech being slightly more intelligible than NC speech) even after

accounting for a variety of acoustic predictors. That result is consistent with findings from Clopper and Bradlow (2008), who reported higher intelligibility for General American talkers than for Northern talkers, but cannot be easily related to the findings of Jacewicz and Fox (2014), who did not include Northern / Northern Cities speech in their study.

It is worth noting in this context that the similarity of the two dialects made the task easier than it might have been if the dialects had been more divergent. This is evidenced by the need for a fairly low SNR (+2 dB) to elicit a sufficient number of errors in the perception task. This is a level that can be readily found in daily life (cf. Hodgson, Steininger, & Razavi, 2007); with stronger divergence, dialect effects on intelligibility may have been more observable at this level. It is also worth noting that the results presented here may not generalize to speech perception tasks at more challenging SNRs, as the different acoustic dimensions discussed may not degrade at the same rate as noise increases.

Summary of acoustic findings

In this study, more intelligible talkers tended to have larger convex hulls encompassing their vowel spaces, smaller mean vowel cluster sizes, lower repulsive force of their vowel systems, and smaller vowel polygons defined by F1 and F2 means. More intelligible talkers also tended to have larger overall pitch ranges, larger average pitch range magnitudes across sentences, higher mean pitch velocities, and lower mean pitch dynamicities. Higher intelligibility was also correlated with higher values of mean intensity dynamicity, and by lower (i.e., more negative) values of mean intensity velocity. Finally, there was a slight tendency for males to be more intelligible than females in this sample of talkers.

Variation in intrinsic intelligibility

The fact that we found variation in intrinsic intelligibility across talkers within the PN/NC corpus is hardly surprising. It is well documented that talker-specific variation can influence the intelligibility of a particular utterance, in that some talkers' utterances are more intrinsically intelligible than others (e.g., Bradlow et al., 1996; Payton et al., 1994; Picheny et al., 1985). However, the relative benefit of talker-specific variation derives from a complex talker-listener relationship. For example, a listeners' word and speech sound identification accuracy is higher when the talker is held constant than when the talker varies across trials (Goldinger, Pisoni, & Logan, 1991; Nygaard, Sommers, & Pisoni, 1995; Palmeri, Goldinger, & Pisoni, 1993). This is usually interpreted as a benefit of exposure or training. This benefit appears to derive from dynamic-pronunciation aspects of the signal production rather than from production-independent variation, such as post-recording manipulations (Bradlow et al., 1996; Church & Schacter, 1994; Sommers & Barcroft, 2007). This indicates that listeners can use knowledge of a talker's idiosyncratic pronunciation traits to aid in performing the speech perception task. Talker-familiarity benefits are particularly robust when the listener is intimately familiar with a talker, as occurs in long-term friendships or in marriages; moreover, intimate-familiarity becomes more important under adverse listening conditions (Souza, Gehani, Wright, & McCloy, 2013).

The strong impact of listener experience on talker intelligibility raises the question of whether listener experience affects intelligibility in cross-dialect listening situations, and whether experience, exposure or training with a dialect might generalize to novel speakers of that dialect, as has been shown for foreign accented speech (Bradlow & Bent, 2008). More relevant to this study, however, is the question of how to quantify variability at the dialect level (and any listener

adaptation to dialect) in the face of talker-level variability and the known ability of listeners to adapt to individual talkers.

Separating dialect from idiolect

In creating stimulus corpora for speech perception research, there is an inherent trade-off in the amount and type of variability to include. On one end of the spectrum, large numbers of words or sentences are recorded from a single talker (at the expense of talker variability); at the other end of the spectrum, many different talkers are recorded saying just a few words or sentences each (at the expense of stimulus variability). If what is of interest is the difference between groups (e.g., age groups, genders, dialects, familiar/unfamiliar talkers), different talkers must necessarily be recruited to represent each of those groups. This introduces a potential source of error: the groups become confounded with the identity of the talkers representing those groups, and any difference (or lack thereof) may be attributable to the idiosyncrasies of the talkers' voices rather than the aspects of their speech that define them as a member of the group they represent. Put another way, it becomes difficult to separate idiolect, on one hand, from dialect or sociolect on the other.

The ideal solution to this problem is a sufficiently large number of talkers in each group such that the idiolectal variation averages out within groups, allowing cross-group differences to emerge. However, practical limitations prevent sample size from expanding indefinitely, and in any case it is not always obvious how much idiolectal variation is expected on the dimensions of interest, nor is it always clear how large a sample of talkers is necessary to adequately represent this variation. In this study we used ten talkers per dialect group; Clopper and Bradlow (2008) used six, Jacewicz and Fox (2014) used four. Whatever other advantages our methods may have over those previous studies, it is clear that even ten talkers per dialect was not enough to ensure

equivalence of the samples on aspects of intelligibility that are unrelated to dialect. Our statistical methods and acoustic measurements allowed us to deal with this aspect of our sampling, make some observations about which aspects of talker variability were most predictive of intelligibility differences, and detect a small effect of talker dialect on intelligibility once those acoustic differences had been accounted for. However, we are still unable to say to what extent that finding reflects genuine intelligibility differences between the dialects, or merely residual aspects of the talker idiolects in our sample that were not accounted for by the other acoustic predictors. Any inference to the intelligibility of dialects generally is therefore infeasible without a larger-scale symmetrical study encompassing many more dialects and many more talkers than were included here.

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. doi:10.1037/a0013552

American National Standards Institute. (2004). *American national standard methods for manual pure-tone threshold audiometry* (No. ANSI/ASA S3.21-2004 (R2009)). Melville, NY: Acoustical Society of America.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180. doi:10.1121/1.4789864

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.0-4). Retrieved from http://CRAN.R-project.org/package=lme4

Bent, T., & Holt, R. F. (2013). The influence of talker and foreign-accent variability on spoken word identification. *The Journal of the Acoustical Society of America*, *133*(3), 1677–1686. doi:10.1121/1.4776212

Black, J. W. (1957). Multiple-choice intelligibility tests. *Journal of Speech and Hearing Disorders*, *22*(2), 213–235. doi:10.1044/jshd.2202.213

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (Version 5.3.41). Retrieved from http://www.praat.org/

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, *14*(4), 325–337. doi:10.1016/0167-6393(94)90026-4

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. doi:10.1016/j.cognition.2007.04.005

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3-4), 255–272. doi:10.1016/S0167-6393(96)00063-5

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521–533. doi:10.1037/0278-7393.20.3.521

Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech*, *51*(3), 175–198. doi:10.1177/0023830908098539

Clopper, C. G., Levi, S. V., & Pisoni, D. B. (2006). Perceptual similarity of regional dialects of American English. *The Journal of the Acoustical Society of America*, *119*(1), 566–574. doi:10.1121/1.2141171

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics*, *35*(3), 421–438. doi:10.1016/j.wocn.2006.06.001

Clopper, C. G., Pisoni, D. B., & de Jong, K. J. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, *118*(3), 1661–1676. doi:10.1121/1.2000774

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 152–162. doi:10.1037/0278-7393.17.1.152

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383–406. doi:10.1006/jpho.2001.0147

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108–3118. doi:10.1121/1.1806826

Hodgson, M., Steininger, G., & Razavi, Z. (2007). Measurement and prediction of speech and noise levels and the Lombard effect in eating establishments. *The Journal of the Acoustical Society of America*, *121*(4), 2023. doi:10.1121/1.2535571

Hood, J. D., & Poole, J. P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *International Journal of Audiology*, *19*(5), 434–455. doi:10.3109/00206098009070077

Jacewicz, E., & Fox, R. A. (2014). The effects of dialect variation on speech intelligibility in a multitalker background. *Applied Psycholinguistics*, *prepub*(prepub), prepub. doi:10.1017/S0142716413000489

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81. doi:10.1016/j.cognition.2007.07.013

Labov, W., & Ash, S. (1997). Understanding Birmingham. In C. Bernstein, T. Nunnally, & R. Sabino (Eds.), *Language variety in the South revisited* (pp. 508–573). Tuscaloosa: University of Alabama Press.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. New York: Mouton de Gruyter. Retrieved from http://www.degruyter.com/viewbooktoc/product/178229

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–862.

Mason, H. M. (1946). Understandability of speech in noise as affected by region of origin of speaker and listener. *Speech Monographs*, *13*(2), 54–58. doi:10.1080/03637754609374918

McCloy, D. R. (2013, June). *Prosody, intelligibility and familiarity in speech perception* (Doctoral dissertation). University of Washington, Seattle. Retrieved from http://hdl.handle.net/1773/23472

McCloy, D. R., Souza, P. E., Wright, R. A., Haywood, J., Gehani, N., & Rudolph, S. (2013). The PN/NC corpus (Version 1.0). Seattle: University of Washington. Retrieved from http://depts.washington.edu/phonlab/resources/pnnc/

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*, *51*(3), 574–585. doi:10.1044/1092-4388(2008/041)

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics*, *57*(7), 989–1001.

Oder, A. L., Clopper, C. G., & Ferguson, S. H. (2013). Effects of dialect on vowel acoustics and intelligibility. *Journal of the International Phonetic Association*, *43*(01), 23–35. doi:10.1017/S0025100312000333

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328. doi:10.1037/0278-7393.19.2.309

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, *95*(3), 1581. doi:10.1121/1.408545

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703. doi:10.1121/1.1908183

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, *28*(1), 96–103.

R Development Core Team. (2013). R: A language and environment for statistical computing (Version 3.0.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reed, C. E. (1952). The pronunciation of English in the state of Washington. *American Speech*, *27*(3), 186–189.

Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., … Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246. doi:10.1109/TAU.1969.1162058

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, *23*(3-4), 193–229. doi:10.1016/S0271-5309(03)00013-2

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(02), 231–249. doi:10.1017/S0142716407070129

Souza, P. E., Gehani, N., Wright, R. A., & McCloy, D. R. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(8), 689–700. doi:10.3766/jaaa.24.8.6

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100. doi:10.1121/1.399849

Ward, M. (2003). *Portland dialect study: The fronting of /ow, u, uw/ in Portland, Oregon* (Masters thesis). Portland State University, Portland, OR. Retrieved from http://www.pds.pdx.edu/Publications/Ward.pdf

Wright, R. A. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation* (pp. 75–87). Cambridge, UK: Cambridge University Press.

Wright, R. A., & Souza, P. E. (2012). Comparing identification of standardized and regionally valid vowels. *Journal of Speech, Language, and Hearing Research*, *55*(1), 182–193. doi:10.1044/1092-4388(2011/10-0278)