Proceedings of Meetings on Acoustics

Volume 18, 2012 http://acousticalsociety.org/

164th Meeting of the Acoustical Society of America
Kansas City, Missouri
22 - 26 October 2012

Session 4pSC: Speech Communication

4pSC11. Modeling intrinsic intelligibility variation: vowel-space size and structure

Daniel McCloy*, Richard Wright and Pamela Souza

***Corresponding author's address: Institute for Learning and Brain Sciences, University of Washington, 1715 NE Columbia Rd, Seattle, WA 98195-7988, drmccloy@uw.edu**

This paper describes a model of talker intelligibility in read sentences, based on a variety of vowel-space and prosodic predictors. Data are analysed using mixed-effects regression, including explicit modeling of random variation due to talker, listener, and sentence. Intelligibility is found to correlate with the overall area of the vowel space, the mean area of individual vowel phonemes, and the degree of crowding or encroachment between adjacent phonemes in F2xF1 space. The relationship between talker intelligibility and the prosodic predictors tested remains unclear. The model was tested using a new set of talkers and listeners from a different dialect region, using the same set of sentences. The model did not fully generalize to the second group of talkers and listeners; however, it is argued that differences between the two groups reflect properties of the talker samples rather than genuine dialectal differences.

Published by the Acoustical Society of America through the American Institute of Physics

©2014 Acoustical Society of America [DOI: 10.1121/1.4870070]
Received 24 Jan 2014; published 20 Mar 2014

I INTRODUCTION

This paper is an expansion of "Modeling talker intelligibility variation in a dialect-controlled corpus" (McCloy *et al.*, 2012), presented at the 164th Meeting of the Acoustical Society of America. Additional statistical modeling has been performed since the conference presentation, and the title has been changed to better reflect the results of that modeling. The author list has changed to remove August McGrath and to include Pamela Souza.

II BACKGROUND

It is well known that the intelligibility of speech can vary both across individuals within styles or tasks, and within individuals across styles or tasks. Variation across individuals is sometimes referred to as *intrinsic intelligibility*; variation within an individual across speech tasks or situations might likewise be called *adaptive intelligibility*. In this paper, we investigate acoustic correlates of variation in intrinsic intelligibility across talkers in a fixed speech task (sentence reading). A number of past studies have tried to find reliable acoustic predictors of both intrinsic and adaptive intelligibility, but the reported findings from those studies varies considerably, as do the acoustic measurements and statistical methods underlying those findings. This section reviews those prior studies, and is organized topically (rather than by study) to highlight differences in methodologies and/or results.

A Duration, speech rate, and rhythm

One of the earliest studies of intelligibility and speech rate was by Tolhurst, who found significant differences in intelligibility among three rate-related styles (prolonged, normal, and stacatto speech), with the fastest speech ("staccato") being the least intelligible (Tolhurst, 1957). In more recent studies, however, evidence has been mixed as to whether speech rate is predictive of intelligibility. Bond & Moore (1994) found differences in isolated word durations between two talkers of varying intelligibility, but no difference in words in sentential context. Bradlow *et al.* (1996) found no correlation between mean sentence duration and intelligibility, and Krause & Braida (2002) found that the intelligibility difference between (high-intelligibility) clear speech style and (lower-intelligibility) conversational speech style is apparently independent of speaking rate. In contrast, Sommers *et al.* (1994) found a significant effect of *variation* in speech rate on intelligibility, though this result may reflect informational masking due to stimulus uncertainty (cf. Sommers and colleagues' results on blocked *vs.* randomized talkers in the same paper). Looking across several speech styles, Mayo *et al.* (2012) examined the perception of plain, infant-directed, computer-directed, foreigner-directed, and shouted speech, finding a correlation between sentence duration and word error rate across styles, as well as significant segment length differences across speech styles for stops, fricatives, nasals, vowels, and diphthongs (but not affricates, liquids or glides).

Given these results, it is not clear whether speech rate *per se* is a predictor of intelligibility. The disagreement in results is undoubtedly due in part to speaking-task differences and probably also due in part to listening-task and scoring differences between the studies. Because speaking rate, as well as word and segment duration, are rarely achieved independently of other pronunciation adjustments, it is a slightly problematic measure if not modeled explicitly together with other acoustic dimensions (such as those reviewed here).

There is also some evidence that patterns of duration (*i.e.*, rhythm) can contribute to speech intelligibility. First, with regard to duration in foreign-accented speech, Quené & van Delft (2010) report that the speech reception thresholds (SRT) of high-intelligibility non-native Dutch sentences were improved when the sentences were resynthesized to match the durational patterns of a native Dutch speaker, while the SRT of native Dutch speech was degraded when the native sentence was resynthesized to have the durational patterns of non-native speech. Second, in a 2×2 design investigating native duration patterns and syllable-isochronous duration patterns where all stimuli had been monotonized, it was found that target speech perception was above chance only when the target had native-like duration patterns and the competing speech had syllable-isochronous duration patterns (Cushing & Dellwo, 2010). In other words, subjects performed near chance when either the masker had native-like durations, or the target had syllable-isochronous durations. Both of these studies suggest an intelligibility benefit for stimuli with native-like durational patterns, but the relationship between rhythmic variations (*e.g.*, between native-like and non-native-like rhythms) and summary measures of duration like mean sentence length or mean syllable duration is not well understood.

A third study of longer-time-scale duration patterns examined differences in temporal structure between conversational and clear speech, and found differences in occurrences of vowel and consonant reduction and deletion, leading to differences in the number of prosodic phrases, but an overall consistency across speech styles in the ratio of consonantal to vocalic stretches and the variability of each (Smiljanić & Bradlow, 2008). In other words, consonants and vowels were both subject to durational reduction in conversational style (in roughly equal measure), and the variability in consonant and vowel duration remained constant across speaking styles (relative to the overall speech rate). Research like that of Smiljanić and Bradlow into the relationship between longer-time-scale aspects of speech prosody and intellibility is still relatively rare and therefore still poorly understood. However, as prosody is well known to correlate with the larger discourse context (*e.g.*, Pierrehumbert & Hirschberg, 1990; Hirschberg, 2002) it makes sense to hypothesize that differences in prosody may be related to differences in intelligibility, and the relationship is therefore worth investigating (cf. the discussion of individual differences in prosody in McCloy 2013, chap. 5).

B Vowel space size

From both word identification and sentence comprehension experiments, there seems to be consistent evidence for the overall size of the vowel space as a reliable predictor of intelligibility, though different studies have used different metrics of vowel space size. Size-related measures reported to positively correlate with intelligibility include F1 range (Bradlow *et al.*, 1996), F2 range (Hazan & Markham, 2004), mean distance of vowel tokens from the center of the vowel space (Bradlow *et al.*, 1996), and area of the polygon formed by vowel means (Neel, 2008).[1]

While there seems to be consistency in the finding that vowel space expansion/contraction is generally well correlated with differences in intelligibility, it is not yet clear which measure best captures the relationship, nor why the relationship exists. It is also not clear whether or not overall expansion, as opposed to selective expansion of specific words, is the best predictor. The relationship between intelligibility and other aspects of vowel space structure (*e.g.*, phoneme overlap, near-mergers, *etc*) is not yet well understood.

C Gender, f_0, and intonation

Some studies have found that speech produced by female talkers is more intelligible than speech produced by male talkers (Bradlow *et al.*, 1996; Hazan & Markham, 2004), while others show no correlation between intelligibility and talker gender for normal-hearing listeners (Kiliç & Ögüt, 2004; Neel, 2008). Since the fundamental frequency (f_0) of speech is typically much lower for adult males than females, a natural question to ask is whether gender differences in intelligibility are due (in part, at least) to differences in f_0. In fact, static measures of pitch such as mean f_0 do not correlate well with speech intelligibility (Picheny *et al.*, 1986; Bradlow *et al.*, 1996; Hazan & Markham, 2004; Lu & Cooke, 2009), suggesting that gender differences in intelligibility are merely indexed by differences in mean pitch, rather than caused by them (cf. discussion in Bradlow *et al.*, 1996).[2]

In contrast, measures that reflect a talker's dynamic use of pitch do seem to be relevant to intelligibility; Bradlow *et al.* (1996) reports a positive correlation between a talker's f_0 range and intelligibility, regardless of talker gender. Further evidence that dynamic properties of f_0 matter to intelligibility come from studies of speech stimuli with manipulated f_0 contours. Binns & Culling (2007) report a significant increase in speech reception threshold (SRT) in an English competing speech task when the f_0 contour of target speech is flattened or inverted, but no effect when the masker speech is similarly manipulated. Watson & Schlauch (2008) report similar results for speech with flattened f_0 in white noise. Predictably, flattening pitch contours also impacts intelligibility in lexical tone languages like Modern Standard Chinese (Patel *et al.*, 2010). Changes in f_0 may also have corollary effects on the signal that impact intelligibility. Laryngealization (or “creaky voicing”) is common as a prosodic marker (Lehiste, 1979; Kreiman, 1982; Dilley *et al.*, 1996), and is associated with a drop in both f_0 and intensity (Gordon & Ladefoged, 2001). The drop in intensity makes creaky-voiced words particularly susceptible to energetic masking.

In studies that have found gender based differences, female speakers are typically more intelligible than their male counterparts; however, since females typically have a more dynamic use of pitch, and since pitch dynamics are correlated in at least some studies with intelligibility,

[1] Note that Bradlow *et al.* (1996) (who did not find polygonal area to significantly correlate with intelligibility) used the vowels /i o a/ to construct the polygons, while Neel (2008) used /i æ ɑ u/.

[2] Hazan & Markham (2004) also report a gender difference in intelligibility using word list stimuli, though they admit that the gender difference seen in their data was much less clear than that reported by Bradlow, Torretta, and Pisoni.

it is likely that the apparent gender differences are likely due differences in pitch use. There is clearly much to be learned from new measures of pitch use as it relates to intelligibility.

D Clear speech style

Early work by Tolhurst on adaptive intelligibility and speech styles was discussed in Section II–A with regard to speaking rate (Tolhurst, 1957); related studies examined the effect of instructions to talkers to speak clearly (Tolhurst, 1954, 1955). Most recent research into the intelligibility of speech styles has likewise focused on "clear speech" — speech directed toward listeners who are hard of hearing and produced with intent to aid comprehension — a line of research pioneered by Durlach, Braida and colleagues (*e.g.*, Picheny *et al.*, 1985, 1986, 1989; Uchanski *et al.*, 1996; Krause & Braida, 2004). These and related studies show intelligibility improvements in the range of 10–20% over conversational speech.

The acoustic differences between clear and conversational speech span a range of acoustic dimensions, including changes in vowel formant frequencies, speech rate (due to both longer words and more pauses in clear speech), intonation patterns, and segmental reduction or deletion (see Picheny *et al.*, 1986; Li & Loizou, 2008; Smiljanić & Bradlow, 2008; Hazan & Baker, 2011, inter alia). However, precise characterization of each parameter's contribution to intelligibility is elusive, due at least in part to the difficulty of manipulating these parameters independently, and the processing artifacts that arise when attempting to do so via speech resynthesis (cf. discussions in Picheny *et al.*, 1989; Uchanski *et al.*, 1996; Liu & Zeng, 2006; Krause & Braida, 2009).

E Dialectal differences

It is well known that regional English dialect and accent differences are not uniform across the different regions in North America nor are they uniform across speech sounds. Some dialects are more similar to each other on one or more dimensions than they are to others, and some individual speech sounds vary more by region than others. For example, in several dialects spoken in New England and in the Northern Cities, urban areas on the southern shores of the Great Lakes such as Chicago, Detroit, Cleveland, and Buffalo, there is a phonemic contrast between an open-mid vowel /ɔ/ and a low-central vowel /a/ (as in the words "caught" and "cot" respectively), whereas in much of the West and the South this contrast is replaced with a single low-back vowel /ɑ/ (Clopper *et al.*, 2005; Labov *et al.*, 2006). Even when dialects share the same phoneme inventories and allophonic processes, they may differ in the phonetic realizations of the shared sounds. For example, while the symbol /u/ is used to represent a vowel that is present as a phoneme in all dialects of North American English, the acoustic realization of this vowel varies from a high back variant in Wisconsin to a high central variant in Southern California and parts of the Deep South (Clopper *et al.*, 2005; Labov *et al.*, 2006; Jacewicz *et al.*, 2007).

It is reasonable then to expect that dialectal differences may interact with intelligibility even in sentence level tasks. Adank & McQueen (2007) conducted a noun-animacy (animate *vs.* inanimate) decision task in which listeners were presented with auditory stimuli in a familiar and an unfamiliar accent. In the study, subjects' response times were slower for words presented in the unfamiliar accent. The effect persisted even when listeners were exposed to twenty minutes of speech in the unfamiliar accent prior to the animacy decision task, indicating a lasting effect for unfamiliar accents. Similarly, Floccia *et al.* (2009) found that unfamiliar accents in certain tasks impose a processing delay that is long lasting; it continues long after intelligibility scores reach ceiling.

Most relevant to the current study, in a study of cross-dialect intelligibility in noise, Clopper & Bradlow (2008) found that dialects that were more similar to General American English, which included New England, the West, and the Midland regions, were more intelligible than regional dialects that were more distant from General American including Mid-Atlantic, Northern, and Southern regions. While not all dialectal regions were equally represented so true regional effects were not fully probed, on the whole their results indicate that there is a negative effect on intelligibility of mismatch between the listener's dialect and the dialect in the stimuli in noisy listening environments.

F Background summary

Taken as a whole, the research summarized above paints a somewhat blurry picture of intelligibility research. Intelligibility seems to be affected by a talker's use of linguistic resources like vowel formant frequencies, speaking rate, or pitch. However, the precise contribution of each is difficult to resolve, and it remains unclear whether variation in intrinsic intelligibility results from the same acoustic modulations as style-based intelligibility differences. The current study aims at addressing one piece of this puzzle — the role of vowel space size and structure in intelligibility — while controlling for other variables known or suspected to be relevant. A secondary goal is to explore new ways of quantifying dynamic aspects of speech prosody.

III METHODS

A Participants

Listeners were drawn from the Pacific Northwest (PN) dialect region. 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, 137). By chance, all PN listeners were natives of Washington state (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 the PN. The mean age of the listener group was 20.5 years. 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 (re: ANSI, 2004). Overall, 15 listeners participated in the task.

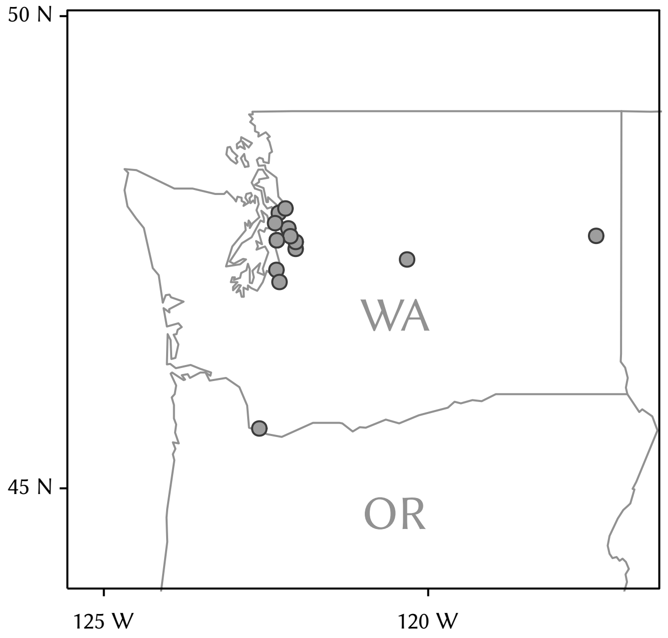

Figure 1: Hometown locations of Pacific Northwest listeners.

B Materials

The stimuli created for this study were a subset of the IEEE “Harvard” sentences (Rothauser *et al.*, 1969). From the full set of 720 sentences, 200 were selected based on absence of alliteration or rhyming, avoidance of focus/contrast readings, and lack of marked locutions (*e.g.*, “the juice of lemons” instead of “lemon juice”). Five male and five female talkers were recorded reading three repetitions of the block of 200 sentences. A shorter, fourth block of sentences was read in cases where specific sentences were problematic in all performances from the first three blocks. All talkers were natives of the PN region.

Sentences were recorded using a head-mounted close-talking microphone (Shure SM10–A) to ensure consistent and maximal signal-to-noise ratio (SNR) of the raw recordings. Talkers were coached to speak in a natural, relaxed manner, with no special effort or emphasis. Talkers who exhibited list intonation across sentences were alerted to their behavior and coached to produce normal (falling) declarative prosody on every sentence. Three trained phoneticians chose the best reading of each sentence for inclusion in the corpus, determined by lack of mic overloading or clipping, absence of hesitations and disfluencies, *etc.* 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 amplitude normalized. From the 200 sentences recorded, 20 were reserved for task familiarization, yielding a final corpus of 1800 stimuli (180 sentences × 10 talkers).

C Perception task

Stimuli 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 concatenated 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. 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. An audio recording was made of listener responses, and scoring uncertainties were resolved offline by a second researcher. Talker-sentence-SNR assignments were random and unique for each listener, with the following constraints: (a) each listener heard each talker an equal number of times; (b) within each talker, each listener heard each SNR an equal number of times; (c) each listener heard each sentence only once.

D Acoustic measures

Vowel space characteristics were calculated based on hand-measurements of 500 vowel tokens (10 talkers × 5 tokens per vowel for the 10 vowels /i ɪ e ɛ æ ɑ o ʊ u ʌ/). Measured vowels were drawn from keywords in positions throughout the sentence, with a preference for vowels with obstruent flanking consonants to avoid coloring by adjacent nasals, rhotics, or laterals. These hand-measured formant values were converted to a perceptual scale using the bark transform (Traunmüller, 1990) prior to data analysis.

Four acoustic predictors were calculated from the formant data: mean Euclidean distance from the center of the vowel space (cf. Bradlow *et al.*, 1996), area of the vowel polygon (cf. Bradlow *et al.*, 1996; Neel, 2008), total repulsive force of the vowel system (cf. Liljencrants & Lindblom, 1972; Wright, 2004), and mean vowel cluster size (see Figure 2).[3] The measure of polygonal area differs from previous studies in being based on a large number of vowel phonemes, in contrast to the /i o a/ triangle used in Bradlow *et al.* 1996, or the /i æ ɑ u/ quadrilateral used in Neel 2008. Additionally, we calculated polygonal area as the area of the convex hull encompassing all

[3] It is noteworthy that we chose not to include measures of F1 and F2 range in our models, especially given that F1 range was found to be a significant predictor of intelligibility by Bradlow *et al.* 1996, and F2 range was reported as significant by Hazan & Markham 2004 (though only for adult males). The reason we omitted these measures is twofold: first, 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 & Markham 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, ch. 4). The second reason for omitting these measures is that they are often highly correlated with other measures of vowel space size. Thus we opted for what we deemed to be more fine-grained measures of vowel space topology.

measured vowel tokens, rather than defining the polygon based on the mean value for each phoneme (as Bradlow *et al.* and Neel did). One reason for this method of calculation is based on the idea that a polygon based on vowel means contains information about a range of reduced and unreduced forms of each phoneme (and thus indexes a talker's prosodic habits to some degree), whereas a convex hull is more representative of a talker's unreduced pronunciations and thus might abstract away from individual differences in prosody (see McCloy, 2013, chap. 5, for discussion).

Repulsive force (sometimes called "total energy") 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 caclulation seen here differs from both Liljencrants & Lindblom 1972 and Wright 2004 in calculating force based on individual vowel tokens rather than mean values for each vowel.

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

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.

In addition to the vowel space predictors, speech rate (syllables/second), one measure of intensity, and two measures of f_0 were also included. Because stimuli were RMS normalized, mean intensity across stimuli is 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.

For measures of pitch, pitch tracks were automatically extracted using Praat (Boersma & Weenink, 2012) and a random subset of 15 of the 180 sentences were selected for hand-correction. This yielded a total of 300 pitch tracks for data analysis (15 per talker × 20 talkers). From those 15 sentences the average pitch range magnitude was calculated for each talker,[4] as well as the mean absolute value of rate of change in pitch ("pitch dynamicity," cf. measures of "pitch velocity" used in Gauthier *et al.* 2007a,b, 2009). Pitch dynamicity was included as a measure of how dramatic the rises and falls in f_0 were across each sentence, irrespective of overall f_0 downtrend.

[4] The choice to use mean size of pitch range rather than 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 from the difficulty of automatic pulse detection and pitch tracking algorithms in dealing with creaky voicing).

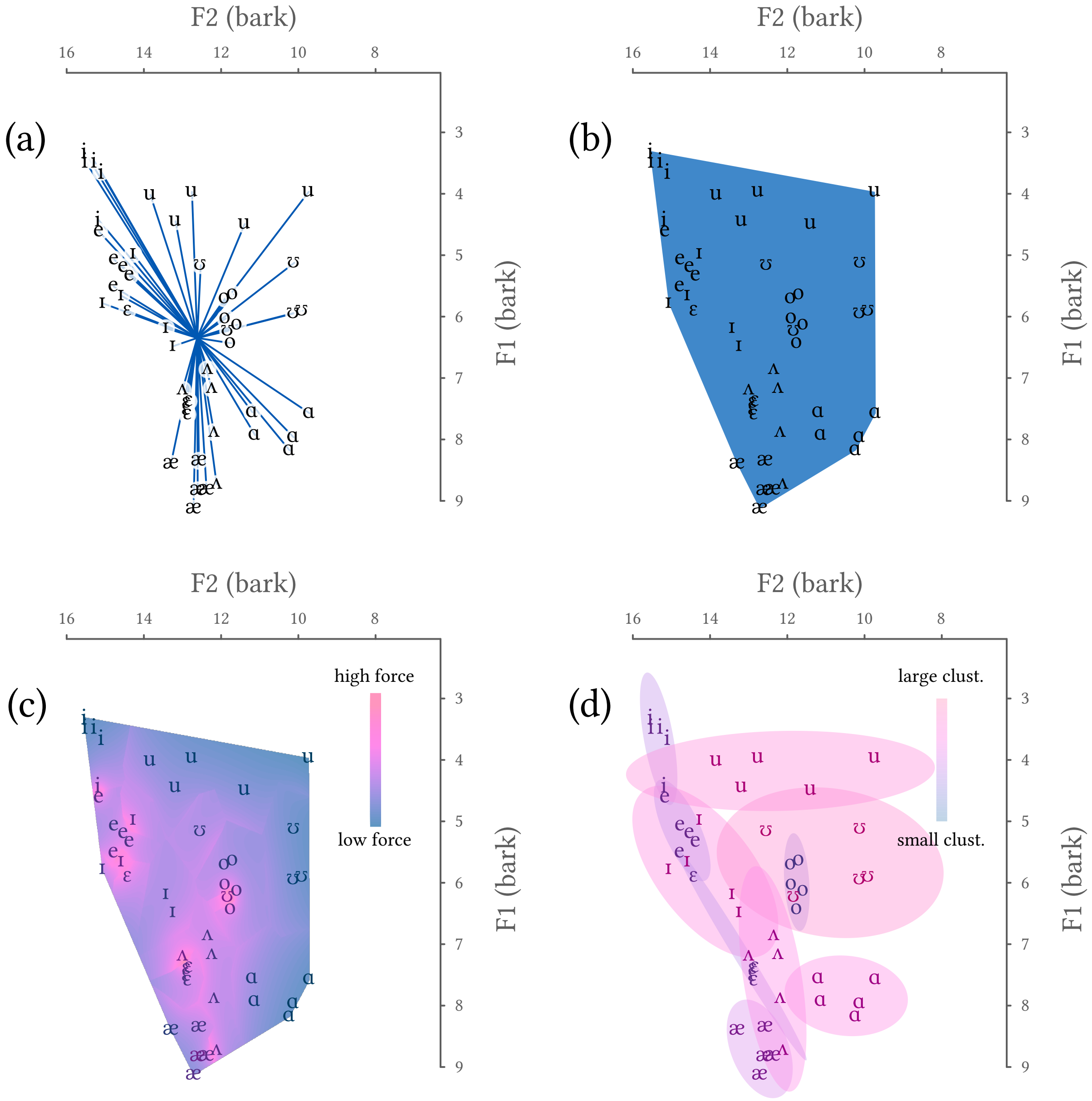

Figure 2: Illustration of the acoustic measures of the vowel space used in the statistical models, for an arbitrarily chosen talker. (a) Mean Euclidean distance from center. (b) Area of the convex hull encompassing all measured vowel tokens. (c) Repulsive force. (d) Mean vowel cluster size.

E Statistical modeling

The acoustic predictors (along with a binary predictor for talker gender) were entered into a mixed-effects logistic regression model in R (R Development Core Team, 2012) using the lme4 package (Bates *et al.*, 2012). The outcome (number of correct keywords) was reduced to a binary score (1 = all keywords correct),[5] and all predictors except gender were normalized prior to modeling. Data from the clear and +6 dB SNR noise conditions were not included in the models due to ceiling effects. Full model specification prior to likelihood ratio tests is given here:

```
glmer(score ~ speechRate + meanDistFromCenter + polygonalArea + repulsiveForce +
meanClusterSize + meanPitchRange + meanPitchDynamicity + meanIntensityVelocity +
talkerGender + (1|talker) + (1|listener) + (1|sentence), family=binomial)
```

From the full model, poor predictors were eliminated via likelihood ratio tests using a p-criterion of 0.3.[6] The random effect for talker was also eliminated, based both on the likelihood ratio tests and on the observation that the estimated variance for those effects was effectively zero.[7]

IV RESULTS

The final model (after model simplification based on likelihood ratio tests) is shown in Table I. Among the vowel space predictors, polygonal area was positively correlated with intelligibility, while repulsive force, cluster size, and mean distance from center were negatively correlated with intelligibility. The negative correlation of mean distance from center is somewhat surprising, though one possible explanation for this finding is that the inclusion of more vowels than previous studies (especially mid-vowels /e ɛ o ʌ/ and non-peripheral vowels /ɪ ɛ ʊ ʌ/) may have reduced the extent to which this measure correlates with the overall expansion of a talker's vowel space, effectively adding noise to the measurement (cf. Figure 2a). The negative correlations between intellibility and both vowel cluster size and repulsive force are expected: talkers with higher within-category variability and/or more cross-category encroachment were predicted to be less intelligible.

Regarding the prosodic predictors, mean pitch range was positively correlated with intelligibility, while pitch dynamicity was negatively correlated with intelligibility. There was no correlation between intelligibility and intensity velocity. The negative correlation with pitch dynamicity is unexpected: we predicted that talkers who speak with a more dynamic intonation

[5] This was done to avoid the complexities of multinomial modeling (a discretized 0–5 score is not appropriately modeled as continuous). Another alternative, averaging across sentences within talker-listener pairs to approximate a continuous outcome, was avoided because it did not allow modeling by-sentence random variation.

[6] Though it may seem paradoxical, using a relatively high p-criterion for likelihood ratio tests is in a sense a *conservative* strategy, in that it discourages the elimination of predictors that might turn out to be statistically significant components of the final fitted model, even if they appear to be marginally non-significant contributors in the context of the full model. Of course, once the final model is fit, the statistical significance of each of the remaining predictors is individually assessed with a more stringent p-criterion.

[7] The zero-estimates of talker variance can be attributed to the inclusion of the talker-level acoustic predictors, based on the fact that a null model of **score ~ (1|talker) + (1|listener) + (1|sentence)** shows substantial by-talker variance. The details of this comparison to the null model are not reported for reasons of space.

Table I: Summary of fixed effects predictors for the PN talkers and listeners.

	Coef.	*SE*	*z*	*p*
Intercept	1.155	0.132	8.735	$< 10^{-15}$
Mean dist. from center	−0.712	0.318	−2.241	< 0.05
Polygonal area	1.379	0.466	2.961	$< 10^{-2}$
Repulsive force	−0.203	0.082	−2.471	< 0.05
Mean vowel cluster size	−0.837	0.177	−4.726	$< 10^{-5}$
Mean pitch range	2.835	0.477	5.942	$< 10^{-8}$
Mean pitch dynamicity	−3.060	0.567	−5.399	$< 10^{-7}$
Mean intensity velocity	−0.202	0.124	−1.624	> 0.1

would be more intelligible, reasoning that high pitch dynamicity would indicate greater use of pitch excursions to mark prominent or low-context words in the sentence. However, another possibility is that measures of pitch dynamicity are being artificially inflated in some talkers by large drops in pitch associated with creaky voicing, and that talkers with a high occurence of creaky voicing may be less audible in noise due to the reduced energy in creaky-voiced portions of the signal. Thus we are cautious of over-interpreting the negative correlation between pitch dynamicity and intelligibility.

A Testing the model on novel talkers and listeners

To assess the generalizability of the model, the same initial set of acoustic predictors and model simplification procedure were used to model data from a separate set of talkers and listeners, from the Northern Cities (NC) dialect region. The NC region was defined following Labov *et al.* (2006, 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 likewise a sub-region of the "North" region described in studies by Clopper *et al.* (2005, 2006), which largely follow Labov *et al.* (2006). Ten talkers and 13 listeners were recruited; inclusion criteria (other than region) were identical to the PN study. The mean age of the listeners was 24.5 years, and their hometowns are shown in Figure 3.

This region was chosen partly for convenience (the third author is located within the region) and partly because known differences between NC and PN dialects are relatively small. Analysis of the NC data was identical to the PN data, with the exception that in place of the PN /ɑ/ phoneme, two phonemes (/a/ and /ɔ/) were measured. This changed the overall number of vowel tokens from 500 to 600, and affected calculations of repulsive force by increase in vowel categories (from 10 to 11). This makes direct comparison of repulsive force values across the two dialects problematic, but such direct comparison is unnecessary since the comparison of interest is between the statistical models of the two regions, rather than the raw acoustic values for talkers from each region.

The picture that emerges from the NC data is less clear than in the model for the PN region; a summary of the NC model is given in Table II. Among the vowel space predictors, polygonal

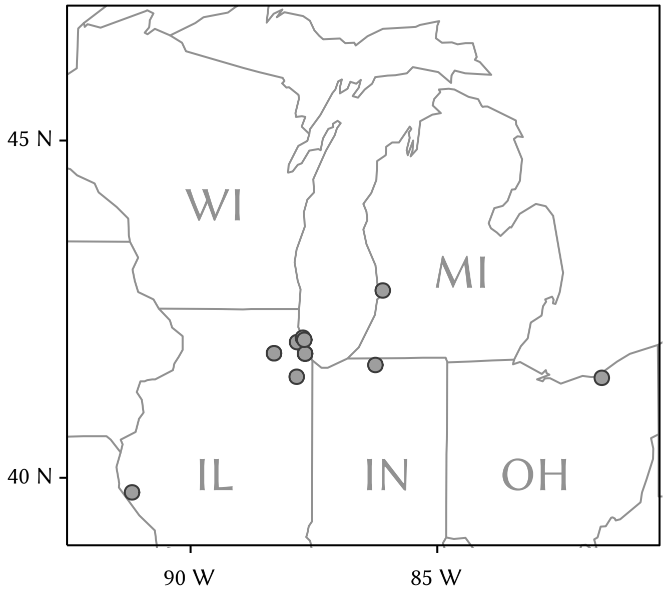

Figure 3: Hometown locations of Northern cities listeners.

area, mean distance from center, and repulsive force are significantly correlated with intelligibility, but the direction of correlation for polygonal area is the opposite of the PN model, and the opposite of what we expect based on previous studies (*i.e.*, according to this model a larger vowel space predicts *lower* intelligibility). The correlation between intelligibility and mean distance from center is also in the opposite direction as in the PN model, though as mentioned in Section III–E, the reliability of this predictor is subject to some doubt (due to the inclusion of several mid-vowels). Repulsive force patterns as expected (higher repulsive force predicts lower intelligibility in this model), and was the only predictor that was consistent in both significance and direction between the models for the two dialects. None of the remaining predictors show significant relationships to intelligibility in the NC model.

Table II: Summary of fixed effects predictors for the NC talkers and listeners.

	Coef.	*SE*	*z*	*p*
Intercept	1.749	0.243	7.210	$< 10^{-12}$
Mean dist. from center	0.772	0.320	2.411	< 0.05
Polygonal area	−0.901	0.310	−2.904	$< 10^{-2}$
Repulsive force	−0.563	0.253	−2.223	< 0.05
Mean pitch range	0.325	0.172	1.897	> 0.05
Mean pitch dynamicity	−0.476	0.279	−1.710	> 0.05
Mean intensity velocity	−0.228	0.209	−1.091	> 0.2
Talker gender (f=0, m=1)	−0.602	0.545	−1.105	> 0.2

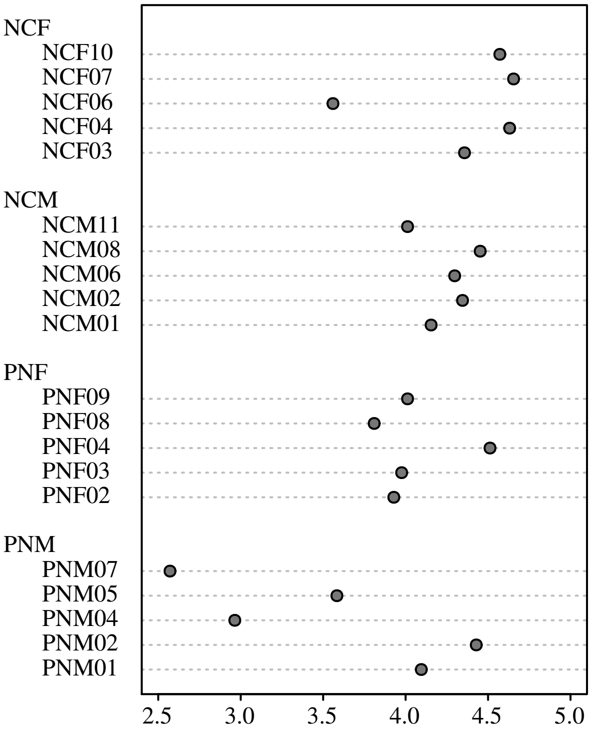

Figure 4: By-talker mean keywords correct (across dialect-matched listeners) for the +2 dB SNR condition.

B Comparing the regional models

The differences between the PN and NC models are quite surprising, given the parallel corpora and similarity of the dialects and the task. To better understand these results, consider the distribution of intelligibility scores by talker across the two regions, as seen in Figure 4. The most striking characteristic is the tight clustering of the NC males (especially in comparison to the PN males). This suggests that even with five talkers per gender per region, there is still a problem with the representativeness of the sample: variation in intelligibility is not being adequately captured in the NC males. Note that this does not seem to be purely a ceiling effect in the perception task, given the distribution of intelligibility scores among the PN males.

Another striking characteristic of Figure 4 is the unusual pattern in the NC females (with NCF06 seeming to be an outlier). Unfortunately there is not enough residual variability in the NC intelligibility scores to meaningfully interpret the model with NCF06 excluded.

V DISCUSSION

This study set out to clarify the relationship between intrinsic intelligibility and various acoustic dimensions of speech, particularly the properties of the talker's vowel space. The PN model suggested that vowel space expansion (indexed here by polygonal area) is positively correlated with intelligibility, consistent with past findings. The measure “mean distance from center” as a metric of vowel space expansion appears less reliable than expected given previous findings in the literature, possibly due to our inclusion of non-peripheral and/or mid vowels. Polygonal area (as defined here, *viz.* the area of the convex hull encompassing all vowel tokens) looks to be a more robust metric of vowel space expansion. A further finding from the PN model was that both within-category vowel variability (represented by mean vowel cluster size) and vowel overlap or encroachment (represented by repulsive force) were useful predictors of intelligibility. This makes sense in terms of vowel recognition, where smaller within-category variability and smaller across-category overlap both increase recognition accuracy.

Mean pitch range was also positively correlated with intelligibility, though the utility of the prosodic predictors (pitch dynamicity and intensity velocity) was unclear. More fine-grained prosodic measures may be needed to understand the precise role of prosody in intelligibility (*e.g.*, stressed/unstressed vowel ratios, a measure of pitch movement that disentangles use of pitch accents from creaky voicing, *etc*). Another interesting result was the lack of a significant effect for gender in the PN model (its absence in the NC model is easily explained by the low intelligibility of NCF06, seen in Figure 4). This suggests that the acoustic predictors used in the model may have adequately captured the dimensions of gender-based speech variation that contribute to higher intelligibility scores for female talkers in many studies.

Speech rate was not significantly correlated with intelligibility either of our models. As mentioned in Section I, past findings regarding speech rate have been mixed; we believe that it is not speech rate *per se* that is important for intelligibility, but rather the concomitant reduction and deletion that often occurs in fast speech (cf. Krause & Braida, 2002). Overall, the generalizability of the findings from the PN model to other dialects is uncertain, due to questions about the representativeness of the NC talker sample mentioned in Section IV–A. Nonetheless, the general consistency of the PN model findings with previous literature suggest that our findings from that model are likely to generalize to other populations.

Finally, an important (though unintended) result of this study is the way it illustrates the importance of a representative sampling of talkers. Many studies of speech perception are carried out with one or just a few talkers, and are often restricted to one gender. In some cases this is a necessary step to control variability that might confound the measurement of interest; in others it an expedient choice for studies requiring a lot of stimulus preparation or manual measurement. But in some cases (including, we believe, almost all cross-dialect studies), it would seem that a larger sample of talkers is required to ensure reliable results. In this study we were fortunate that our PN talker sample exhibited a wide range of intelligibilities, and as such we are confident in the results reported, but the size and representativeness of talker samples is an issue that demands greater attention in future studies.

ACKNOWLEDGMENTS

This research was supported by NIH grant #R01DC006014 to Pamela Souza. We are grateful to Namita Gehani, Jennifer Haywood and August McGrath for help in stimulus preparation, and to members of the UW Phonetics lab and two anonymous reviewers for helpful suggestions on earlier drafts of this paper.

REFERENCES

Adank P., & McQueen J. M. (2007). The effect of an unfamiliar regional accent on spoken word comprehension. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 1925–1928.

American National Standards Institute (2004). ANSI/ASA S3.21-2004 (R2009): methods for manual pure-tone threshold audiometry. Acoust. Soc. Am., Melville, NY.

Bates D., Maechler M., & Bolker B. (2012). lme4: Linear mixed-effects models using S4 classes. http://CRAN.R-project.org/package=lme4.

Binns C., & Culling J. F. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *J. Acous. Soc. Am.*, *122*(3), 1765–1776. DOI: 10.1121/1.2751394.

Boersma P., & Weenink D. (2012). Praat: Doing phonetics by computer. http://www.praat.org/.

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

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

Clopper C. G., & Bradlow A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Lang. 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. *J. Acous. Soc. Am.*, *119*(1), 566–574. DOI: 10.1121/1.2141171.

Clopper C. G., Pisoni D. B., & de Jong K. J. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *J. Acous. Soc. Am.*, *118*(3), 1661–1676. DOI: 10.1121/1.2000774.

Cushing I. R., & Dellwo V. (2010). The role of speech rhythm in attending to one of two simultaneous speakers. Paper presented at the 5th International Conference on Speech Prosody. In *SP-2010*, paper 039.

Dilley L. C., Shattuck-Hufnagel S., & Ostendorf M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *J. Phonetics*, *24*(4), 423–444. DOI: 10.1006/jpho.1996.0023.

Floccia C., Butler J., Goslin J., & Ellis L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguist. Res.*, *38*(4), 379–412. DOI: 10.1007/s10936-008-9097-8.

Gauthier B., Shi R., & Xu Y. (2007a). Learning phonetic categories by tracking movements. *Cognition*, *103*(1), 80–106. DOI: 10.1016/j.cognition.2006.03.002.

Gauthier B., Shi R., & Xu Y. (2007b). Simulating the acquisition of lexical tones from continuous dynamic input. *J. Acous. Soc. Am.*, *121*(5), EL190–EL195. DOI: 10.1121/1.2716160.

Gauthier B., Shi R., & Xu Y. (2009). Learning prosodic focus from continuous speech input: A neural network exploration. *Lang. Learn. Devel.*, *5*(2), 94–114. DOI: 10.1080/15475440802698524.

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

Hazan V., & Baker R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acous. Soc. Am.*, *130*(4), 2139–2152. DOI: 10.1121/1.3623753.

Hazan V., & Markham D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *J. Acous. Soc. Am.*, *116*(5), 3108–3118. DOI: 10.1121/1.1806826.

Hirschberg J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Commun.*, *36*(1–2), 31–43. DOI: 10.1016/S0167-6393(01)00024-3.

Jacewicz E., Fox R. A., & Salmons J. C. (2007). Vowel space areas across dialects and gender. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 1465–1468.

Kiliç M. A., & Ögüt F. (2004). The effect of the speaker gender on speech intelligibility in normal-hearing subjects with simulated high frequency hearing loss. *Revue de laryngologie - otologie - rhinologie*, *125*(1), 35–38.

Krause J. C., & Braida L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *J. Acous. Soc. Am.*, *112*(5), 2165–2172. DOI: 10.1121/1.1509432.

Krause J. C., & Braida L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acous. Soc. Am.*, *115*(1), 362–378. DOI: 10.1121/1.1635842.

Krause J. C., & Braida L. D. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *J. Acous. Soc. Am.*, *125*(5), 3346–3357. DOI: 10.1121/1.3097491.

Kreiman J. (1982). Perception of sentence and paragraph boundaries in natural conversation. *J. Phonetics*, *10*(2), 163–175.

Labov W., Ash S., & Boberg C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. New York: Mouton de Gruyter. DOI: 10.1515/9783110167467.

Lehiste I. (1979). Perception of sentence and paragraph boundaries. In B. Lindblom, S. E. G. Öhman, & G. Fant (Eds.) *Frontiers of speech communication research*, 191–201. London: Academic Press.

Li N., & Loizou P. C. (2008). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *J. Acous. Soc. Am.*, *124*(6), 3947–3958. DOI: 10.1121/1.2997435.

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

Liu S., & Zeng F.-G. (2006). Temporal properties in clear speech perception. *J. Acous. Soc. Am.*, *120*(1), 424–432. DOI: 10.1121/1.2208427.

Lu Y., & Cooke M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Commun.*, *51*(12), 1253–1262. DOI: 10.1016/j.specom.2009.07.002.

Mayo C., Aubanel V., & Cooke M. (2012). Effect of prosodic changes on speech intelligibility. In *INTERSPEECH-2012*.

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

McCloy D. R., Wright R. A., & McGrath A. T. D. (2012). Modeling talker intelligibility variation in a dialect-controlled corpus. *The Journal of the Acoustical Society of America*, *132*(3), 2078. DOI: 10.1121/1.4755663.

Neel A. T. (2008). Vowel space characteristics and vowel identification accuracy. *J. Speech Lang. Hear. Res.*, *51*(3), 574–585. DOI: 10.1044/1092-4388(2008/041).

Patel A. D., Xu Y., & Wang B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. Paper presented at the 5th International Conference on Speech Prosody. In *SP-2010*, paper 890.

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. *J. Speech Hear. Res.*, *28*(1), 96–103.

Picheny M. A., Durlach N. I., & Braida L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *J. Speech Hear. Res.*, *29*(4), 434–446.

Picheny M. A., Durlach N. I., & Braida L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *J. Speech Hear. Res.*, *32*(3), 600–603.

Pierrehumbert J., & Hirschberg J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds.) *Intentions in communication*, 271–311. Cambridge, MA: MIT Press.

Quené H., & van Delft L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Commun.*, *52*(11-12), 911–918. DOI: 10.1016/j.specom.2010.03.005.

R Development Core Team (2012). R: A language and environment for statistical computing. http://www.R-project.org/, Vienna, Austria.

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

Rothauser E. H., Chapman W. D., Guttman N., Hecker M. H. L., Nordby K. S., Silbiger H. R., Urbanek G. E., & Weinstock M. (1969). IEEE recommended practice for speech quality measurements. *IEEE T. Acoust. Speech*, *17*, 225–246. DOI: 10.1109/TAU.1969.1162058.

Smiljanić R., & Bradlow A. R. (2008). Temporal organization of English clear and conversational speech. *J. Acous. Soc. Am.*, *124*(5), 3171–3182. DOI: 10.1121/1.2990712.

Sommers M. S., Nygaard L. C., & Pisoni D. B. (1994). Stimulus variability and spoken word recognition I: Effects of variability in speaking rate and overall amplitude. *J. Acous. Soc. Am.*, *96*(3), 1314–1324. DOI: 10.1121/1.411453.

Tolhurst G. C. (1954). The effect on intelligibility scores of specific instructions regarding talking. Tech. Rep. AD0051810, Ohio State University Research Foundation, Columbus, OH.

Tolhurst G. C. (1955). The effects of an instruction to be intelligible upon a speakers intelligibility, sound pressure level, and message duration. Tech. Rep. AD0094574, Ohio State University Research Foundation, Columbus, OH.

Tolhurst G. C. (1957). Effects of duration and articulation changes on intelligibility, word reception and listener preference. *J. Speech Hear. Disord.*, *22*(3), 328–334.

Traunmüller H. (1990). Analytical expressions for the tonotopic sensory scale. *J. Acous. Soc. Am.*, *88*(1), 97–100. DOI: 10.1121/1.399849.

Uchanski R. M., Choi S. S., Braida L. D., Reed C. M., & Durlach N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *J. Speech Hear. Res.*, *39*(3), 494–509.

Ward M. (2003). *Portland dialect study: The fronting of /ow, u, uw/ in Portland, Oregon.* Masters thesis, Portland State University, Portland, OR.

Watson P. J., & Schlauch R. S. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *Am. J. Speech Lang. Path.*, *17*(4), 348–355. DOI: 10.1044/1058-0360(2008/07-0048).

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