The Perceptual Contribution of Consonants and Vowels to Sentence Recognition: Effect of Dialect Variation in American English

Daniel Fogerty

Department of Communication Sciences and Disorders, University of South Carolina
fogerty@sc.edu

ABSTRACT

In background noise, the amplitude fluctuations of speech commonly provide for momentary glimpses of high intensity portions of speech, predominantly from vowels. Previous investigations have provided glimpses of consonants or vowels to determine the perceptual contribution of different speech acoustics to sentence intelligibility. The present study investigated the consistency of perceptual contributions across eight American English dialects for a group of listeners from the southern United States. Results demonstrated that sentences preserving predominant vowel acoustics were consistently more intelligible across dialects for all participants. The significant contribution of vowels does not appear dependent on familiarity with properties of dialectal variation but may represent the preservation of more general acoustic cues important for sentence recognition. Acoustic analyses of temporal amplitude modulation suggest important cues present during vowels and highlight gradient differences across dialects associated with intelligibility.

Keywords: sentence intelligibility, vowels, dialect, temporal envelope, amplitude modulation.

1. INTRODUCTION

Everyday communication settings often entail listening to speech in the presence of background noise that limits the available spectro-temporal detail of the stimulus. Under such conditions, listeners may have to base the perceptual analysis of speech on spectro-temporal fragments (e.g., glimpses) that provide momentarily favorable signal-to-noise ratios (SNR) [1]. For over half a century, it has been documented that the proportion of available speech information is an important predictor of performance [2]. However, speech is an acoustic signal characterized by high variability in the spectral and temporal domains. Furthermore, certain classes of speech sounds, such as consonants and vowels, are generally represented by different types of acoustic features [3,4]. Recent evidence has suggested that, even when the proportion of speech preserved is controlled, glimpses constrained to different portions of the speech signal can reflect highly variable contributions to intelligibility [e.g., 5,6,7]. For example, a number of studies have now documented that sentences which preserve vocalic glimpses result in word recognition scores twice as high (and in Mandarin, three times as high [8]) as sentences constrained to preserve consonantal acoustics [e.g., 6,9]. Due to high temporal overlap of acoustic-phonetic cues to consonant and vowel identity, these studies did not investigate the contribution of discrete category information. Rather, they reflect the relative distribution of information important for intelligibility over the highly dynamic speech signal. Empirical and acoustic analyses from these studies have suggested that the source of this effect is related to greater preservation of temporal amplitude modulation during the preserved vocalic glimpses [e.g., 10, 11], not phonetic identity [12]. Indeed, interruption conditions based on other analysis-based methods of segmentation, such as entropy, also seem to reflect properties of temporal amplitude modulation [7] and intensity differences [13]. The perceptual contribution of amplitude modulation appears to be greatest during high intensity glimpses [13, 14], such as provided by vowels [11].

In addition to linguistic information, the speech signal also codes dialectical information [15]. For example, the Southern American dialect is distinguished from other dialects based on fundamental frequency (F0) change [15], formant dynamics [16], and duration [16]. This dialectical information results in systematic variability in the acoustic-phonetic signal [16], contributing to vowel variants that are a principle identifier between dialects [17]. Therefore, vowel contributions to sentence intelligibility could be reduced for less familiar dialects. Alternatively, vowel production across all dialects still generates dominant amplitude modulation cues. Therefore, vowel contributions may be preserved regardless of the talker's dialect.

This study investigated these two alternative hypotheses. The intelligibility of sentences interrupted to preserve primarily vowel or consonant segments was studied across eight different dialects that introduce acoustic-phonetic variation. General acoustic measures investigated factors that might explain performance across dialects.

2. METHODS

2.1 Listeners

Fourteen young adults were originally recruited to participate in this experiment. All listeners were native speakers of American English and had normal audiograms with octave pure tone thresholds ≤ 20 dB HL. For the current analysis, four listeners were excluded given residential background histories outside of the Southern United States. The final group of ten listeners had resided in the South for at least 20 years, with localities primarily in North and South Carolina. Eight had always lived in the South, and the remaining two had lived no more than three years outside the Southern United States (Participants 3 and 10). The final group for analysis had a mean age of 23 years (22-25 years). Participants had between 2-4 years of formal language instruction, with two others having a total of 12 years of study (Participants 9 and 14). Years of musical instruction varied between 0-4 years, with two different participants completing 13-14 years of formal instruction (Participants 10 and 11). Testing was completed at the University of South Carolina.

2.2 Stimuli

Stimuli consisted of sentences selected from the TIMIT database [18]. Two lists of 14 sentences were created for each of eight dialect regions from the TIMIT database. Within a list, sentences were balanced evenly between male and female speakers (7 sentences each). The two lists for each dialect were selected to contain the same number of words per list. Total words per list varied from 109-112 words across the eight dialects. Across the entire corpus of selected sentences, there was an average of 17 consonants (SD = 4.3) and 11 vowels (SD = 2.6) per sentence. TIMIT sentences have been used in several studies using segmentally interrupted speech to selectively preserve vowel segments [e.g., 6,9,11] as the database provides time markings for segmental boundaries that were confirmed by expert phoneticians.

2.3 Signal Processing

Signal processing followed the methods used in previous studies of segmentally interrupted speech [6,8,9,11]. Sentences were processed to segmentally interrupt speech using the segmental boundaries identified in the TIMIT database and adjusted within 1-ms to the nearest local minima (i.e., zero-crossing). The interruption preserved segment durations, but replaced the neighboring segments with a low-level speech shaped noise (16 dB SNR) based on the long-term average spectrum of the sentence corpus. This processing resulted in sentences that contained predominantly consonant or vowel segments, as defined by the TIMIT boundary markings. Postvocalic /r/ was considered a rhotacized vowel and intervocalic glottal closure was also grouped as a vowel segment.

2.4 Procedures

All testing was completed in a sound attenuating booth. The participants listened to stimuli presented at a sampling rate of 16 kHz via Sennheiser HD 280 PRO headphones. The level of the speech (prior to replacement) was calibrated to be presented at 70 dB SPL. Stimulus presentation was blocked by dialect with consonant and vowel conditions randomized within a single block of 28 sentences. The order of presentation for the dialect blocks was randomized across participants. Prior to testing, participants first completed demonstration trials to familiarize them with the stimulus processing and the task prior to completing the experimental conditions.

All sentences were processed twice, once to preserve consonant segments and once to preserve vowel segments. Participants were split into two groups, such that half heard a given list with the consonants preserved and half heard the same list with vowels preserved. This procedure ensured that any difference in the consonant/vowel condition was not a result of the particular sentence lists constructed. Two lists were presented such that all listeners completed both vowel and consonant conditions.

Listeners completed the open-set test by repeating each sentence and were encouraged to guess. Sentence presentation was self-paced. Listener responses were audio recorded for offline analysis by trained raters who scored all words. Percent correct scores were transformed to rationalized arcsine units (RAU) to stabilize the error variance.

3. RESULTS AND DISCUSSION

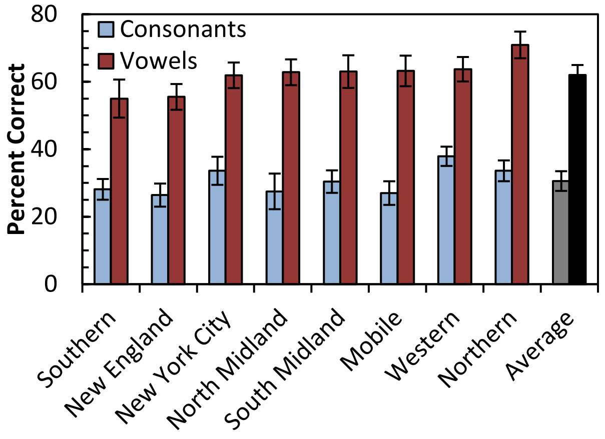
3.1 Mean Performance Across All Dialects

A 2 (segment) X 8 (dialect region) repeated-measures analysis of variance was conducted. Results demonstrated a significant main effect of segment [F(1,9) = 620.0, p<.001, $\eta_p^2 = .99$] and of dialect [F(7, 63) = 3.2, p=.006, $\eta_p^2 = .26$]. No significant interaction was found.

Fig. 1 displays the mean results across dialect. Overall, the results indicate that vowel-preserved sentences resulted in higher recognition rates than consonant-preserved sentences at a ratio of

2:1 (62% versus 31%, respectively). In addition, the main effect of dialect suggests that some dialects were more intelligible than others, and progressed from the Southern and New England dialects as the least intelligible to the Northern dialect as the most intelligible. This later finding is notable given that all participants in this analysis had primarily lived in the South (the Carolinas) and had self-identified as speaking the Southern dialect (albeit a heterogeneous classification).

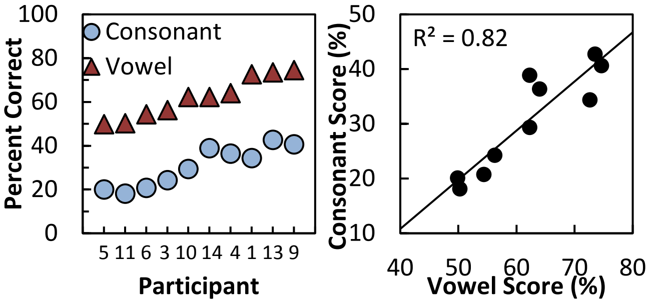
Figure 1: Average performance across the eight dialect regions and averaged across all dialects.



The lack of a significant interaction demonstrates that the advantage observed for vowel-preserved sentences is consistent across dialect regions, suggesting that greater familiarity with the properties of specific dialect variations does not influence the perceptual contribution of vowels in these sentences, relative to consonant contributions. Rather, performance in these two segmental conditions appear to reflect sensitivity to more general perceptual cues useful for speech intelligibility that remain relatively robust in the context of indexical variation in dialects.

As consistent effects were observed across dialects, results were also analysed across individual participants for average performance (Fig. 2, left panel). Notable is the large variability across participants. However, it is clear that all listeners were more accurate in reporting words in the vowel condition. The right panel in Fig. 2 also indicates a significant correlation between average vowel and consonant performance, perhaps indicating a more general ability for glimpsing speech fragments interrupted by noise. However, these consistencies are most notable for average scores, as correlations between dialects, e.g., Southern and Northern dialects, were not significant for consonant or vowel performance (ps<.05). This indicates variability in participant ranking across dialects.

Figure 2: Ranking (left) of average participant performance and correlation (right) of consonant and vowel scores for these participants.



3.2 Contribution of Temporal Amplitude Modulation

Previously, a measure of preserved amplitude modulation has provided a good fit for explaining intelligibility across consonant and vowel conditions [10]. A similar analysis of the temporal envelope was applied in the current case for vowel and consonant sentences. Stimuli were halfwave rectified and low-pass filtered using a Butterworth filter to 50 Hz. Stimulus envelopes were then correlated with envelopes extracted from the original version of the stimulus prior to interruption. This correlation provided an envelope correlation index (ECI) as to the degree to which the interruption condition preserved the amplitude modulation of the full sentence. Likewise, the envelope modulation spectrum was derived by analysing the temporal envelope by the FFT and summing the modulation area (ModA) up to a modulation rate of 32 Hz. Results of these analyses are plotted in Figure 4. Overall, these results demonstrate good classification of acoustic differences between consonant and vowel conditions, but lack specificity to account for performance variation across the eight dialects.

Dialect differences were noted in the modulation spectrum. This was quantified by calculating the difference in modulation area between 1-4 Hz and 8-32 Hz octave bands for the uninterrupted versions of the sentences (i.e., low-high modulation difference, LHMD). These scores (Fig. 5), account for the rank ordering of difficulty across dialects most notably for vowel-preserved sentences.

Fig. 6 displays performance for vowel sentences across dialects (from Fig. 1) in comparison to the LHMD score, which was transformed to facilitate comparison to the percent correct scores. Dialects that have relatively less energy at higher modulation rates (e.g., Southern) result in poorer overall intelligibility as compared to those with more similar modulation energy at low and high rates (e.g., Northern). Sentences were interrupted at a rate characterizing the syllabic alternation between consonants and vowels, which overlaps with modulation in the low-rate band. Those dialects with more relative energy in the modulation spectrum above the interruption rate may have greater preservation of some acoustic-phonetic cues for intelligibility. This is consistent with theories of modulation masking [19] which would suggest that the segmental interruption rate interferes with the transmission of speech information in low-rate modulation bands, leaving fast-rate modulation cues relatively more preserved.

Figure 4: Acoustic measures of preserved amplitude modulation for vowel and consonant conditions. ECI = Envelope Correlation Index; ModA= Modulation area

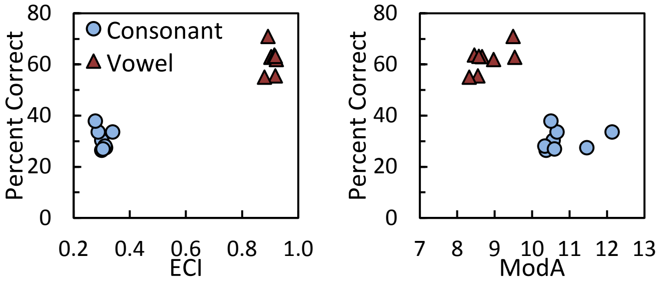


Figure 5: Correlation between LHMD and performance with consonant (left) or vowels (right).

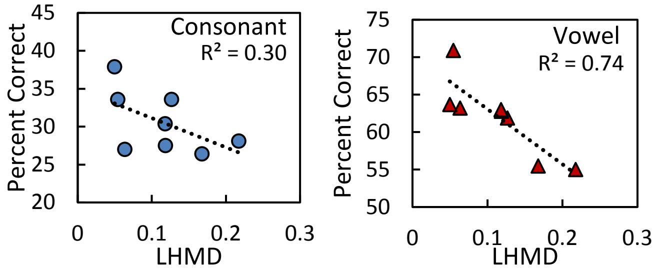
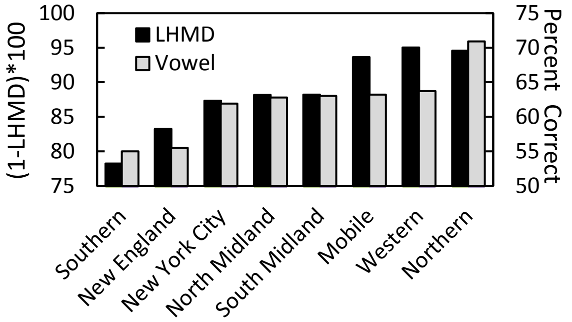


Figure 6: Comparison of transformed LHMD scores across dialects and vowel condition performance.



5. SUMMARY & CONCLUSIONS

This study investigated the recognition of sentences interrupted to provide predominant vowel or consonant acoustics for eight dialects of American English within a group of listeners from the Southern dialect region. Results confirmed better recognition of vowel sentences across dialects, despite dialect-specific vowel variability [16,17]. Results demonstrated significant listener variability. Consistent with Clopper and Bradlow [20], greater intelligibility for some dialects did not interact with the listener's dialect. These authors suggested that the processing demand imposed by highly degraded speech, as tested here, may cause listeners to bypass dialect-level representations. These results are different from Jacewicz and Fox [21] who found greater intelligibility of Southern versus General American (i.e., Midland) dialect when presented in multitalker babble to listeners from Central Ohio. However, as suggested in by Jacewicz and Fox, there may be an advantage when the target speech does not match the listener's dialect. Experimental differences may also be the source of this discrepancy. The significantly greater F0 change characteristic of the Southern dialect may be more advantageous for talker segregation in babble [21]. Indeed, dynamic F0 cues facilitate talker segregation [22]. However, this greater variability may be less useful in the current study when the speech is interrupted, which may introduce some discontinuities in the F0 contour.

Regardless, the current results suggest that listeners are able to extract general properties important for speech recognition in adverse listening conditions (i.e., interruption) across all dialects. Initial acoustic analyses suggest that sentences interrupted to preserve vowel segments result in greater preservation of the amplitude modulation contour of the sentence, consistent with other findings suggesting the importance of the temporal envelope for speech recognition [e.g., 10, 11, 12]. Furthermore, differences in performance across dialects, particularly for the vowel condition, appear to reflect dialect patterns in the relative distribution of modulation energy between low- and high-rate modulation bands. More intelligible dialects had relatively more energy in high-rate modulation bands compared to less intelligible dialects. This is interesting given that low-rate modulation bands contribute more to intelligibility [23]. However, high-rate modulation cues may be relatively more preserved when speech is interrupted at a segmental rate [see 11, 14], perhaps due to decreased modulation masking [19]. Overall, listeners require access to speech amplitude modulations for maximal sentence recognition. Differences in the modulation spectrum explain differences in the intelligibility of segmentally interrupted speech for different dialects.

6. ACKNOWLEDGEMENT

This work was supported, in part, by a grant from the NIH/NIDCD R01-DC015465.

7. REFERENCES

[1] Cooke, M. 2003. Glimpsing speech. *J. Phonetics*, 31, 579-584.

[2] Miller, G. A., Licklider, J. C. 1950. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.* 22, 167–173.

[3] Ladefoged, P., Disner, S. F. 2012. *Vowels and consonants*. John Wiley & Sons.

[4] Stevens, K. N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am*. 111, 1872-1891.

[5] Chen, F., Loizou, P. C. 2012. Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise. *J. Acoust. Soc. Am*. 131, 4104–4113.

[6] Fogerty, D., Kewley-Port, D. 2009. Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *J. Acoust. Soc. Am*. 126, 847-857.

[7] Stilp, C. E. 2014. Information-bearing acoustic change outperforms duration in predicting intelligibility of full-spectrum and noise-vocoded sentences. *J. Acoust. Soc. Am*. 135, 1518–1529.

[8] Chen, F., Wong, L.L., Wong, E.Y.W. 2013. Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility. *J. Acoust. Soc. Am*. 134, EL178-EL184.

[9] Kewley-Port, D., Burkle, T. Z., Lee, J. H. 2007. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am*. 122, 2365–2375.

[10] Fogerty, D. 2013. Acoustic predictors of intelligibility for segmentally interrupted speech: Temporal envelope, voicing, and duration. *J. Speech Lang. Hear. Res*. 56, 1402–1408.

[11] Fogerty, D. 2014. Importance of envelope modulations during consonants and vowels in segmentally interrupted sentences. *J. Acoust. Soc. Am*. 135, 1568–1576.

[12] Fogerty, D. 2015. Indexical properties influence time-varying amplitude and fundamental frequency contributions of vowels to sentence intelligibility. *J. Phon.* 52*,* 89-104.

[13] Oxenham, A. J., Boucher, J. E., Kreft, H. A. 2017. Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy. *J. Acoust. Soc. Am*. 142, EL264–EL269.

[14] Miller, R. E., Gibbs, B. E., Fogerty, D. 2018. Glimpsing speech interrupted by speech-modulated noise. *J. Acoust. Soc. Am.* 143, 3058-3067.

[15] Fox, R. A., Jacewicz, E., Hart, J. 2013. Pitch pattern variations in three regional varieties of American English. *Proc. Interspeech*, Lyon, France, 123-127.

[16] Fox, R. A., Jacewicz, E. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *J. Acoust. Soc. Am*. 125, 2603–2618.

[17] Clopper, C. G., Pisoni, D. B., De Jong, K. 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. *J. Acoust. Soc. Am*. 118, 1661-1676.

[18] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N. 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. *National Institute of Standards and Technology*. NTIS Order No. PB91-505065.

[19] Stone, M. A., Moore, B. C. 2014. On the near non-existence of "pure" energetic masking release for speech. *J. Acoust. Soc. Am.* 135, 1967-1977.

[20] Clopper, C. G., Bradlow, A. 2008. Perception of dialect variation in noise: Intelligibility and classification. *Lang. Speech* 51, 175–198.

[21] Jacewicz, E., Fox, R. A. 2015. The effects of dialect variation on speech intelligibility in a multitalker background. *Appl. Psycholinguist.* 36, 729-746.

[22] Darwin, C. J., Brungart, D. S., Simpson, B. D. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am.* 114, 2913-2922.

[23] Chait, M., Greenberg, S., Arai, T., Simon, J. Z., Poeppel, D. 2015. Multi-time resolution analysis of speech: evidence from psychophysics. *Front. Neurosci.* 9, 214.