The Role of Rhythm and Vowel Space in Speech Recognition

Li-Fang Lai, Janet G. van Hell & John Lipski

The Pennsylvania State University, USA
{lkl5302@psu.edu, jgv3@psu.edu, jlipski@psu.edu}

Abstract

This paper explores the role of rhythm and vowel space in automatic speech recognition (ASR), with a particular focus on Midland and Southern American English in the Appalachian region. Three sets of analysis were conducted. First, we computed the word error rates between the ground truth and the transcripts generated by DARLA. Consistent with previous studies, the results show higher error rates for Southern English (59.5%) than for Midland English (47.2%), suggesting a dialect gap in speech recognition. Next, we examined whether the error rates are influenced by rhythm. The results show that neither %V nor ΔV reliably predicted ASR performance. We also sought to draw a link between vowel space, speech intelligibility, and ASR performance. Three vowel space metrics were considered: convex hull, formant dispersion, and the polygon area. We noticed that as convex hull and formant dispersion increase, the error rates decrease, particularly for Midland speakers. This aligns with our hypothesis that more expanded vowel space enhances speech intelligibility, thus reducing the error rate for the Midland cohort. No clear connection between the polygon area, speech intelligibility, and error rates was found. These results, albeit suggestive, point out some promising directions for improving acoustic modeling in speech recognition.

Index Terms: rhythm, vowel space, speech recognition, dialect gap, speech intelligibility

1. Introduction

As voice-activated technology continues to reshape our life, advances in deep learning for speech recognition and natural language processing have significantly improved the accuracy of automatic speech recognition (ASR) systems and related applications.

One common metric for evaluating ASR performance is to compute the word error rate (WER), defined as the number of word substitutions, deletions, and insertions in the machine-generated transcript, divided by the total number of words in the ground truth [1]. The value is then multiplied by 100 to represent it as a percentage. The lower the value, the better the performance of the ASR system because more words are accurately identified during recognition.

However, there has been increasing concern about racial disparities and dialect gap in speech recognition [2, 3, 4, 5, 6], which urges researchers and ASR builders to incorporate diversity, equity, and inclusion into technology to make speech recognition tools accessible to all.

In what follows, we begin the discussion of dialect gap in speech recognition in American English. We acknowledge that dialect classification in the U.S. differs across researchers and that our sample of participants technically came from the northern (Pennsylvania) and southern (Tennessee) Appalachian areas. Since northern and southern Appalachians overlap with Midland and the Southern regions respectively based on [7]’s classification, we adopted the general terms – Midland and Southern American English – throughout this paper to avoid confusion.

1.1. Dialect bias in ASR performance

Evaluations of the accuracy of YouTube’s automatic captions of isolated words as produced by speakers from California (Western dialect), Georgia (Southern dialect), and New England (Northeastern dialect) in the U.S. yielded the lowest WER (i.e., highest accuracy) for California English and the highest error rates for Georgia English, with New England English falling in between [5]. Another cross-dialect comparison [6] compared the accuracy of Microsoft Bing Speech API and YouTube’s automatic captions of the reading passage “Comma gets a cure” in four acoustically distinct dialects of American English: General American, Northern Cities, Southern, and California English. The results show that both Bing Speech and YouTube’s captions performed best for General American English and had high error rates for California and Southern English.

Two points are worthy of mention here. First, ASR systems overall are worse in decoding Southern speech. Second, the discrepancies in ASR performance in California English between [5] and [6] draws our attention. We posit that the seemingly conflicting results between [5] and [6] may be related to the type of spoken input. Specifically, [5] used isolated words as input and found low error rates for California English, whereas [6] used reading passage and reported the highest error rates for California English. It is therefore hard to determine whether California English is more or less recognizable than other dialects based on these results.

Now that we have seen differences in ASR performance across dialects, we are interested in probing what drives these differences. We posit that rhythm and vowel space may be two possible mechanisms accounting for the dialect gap.

1.2. Regional rhythmic variation in American English

Over the past two decades, several metrics have been proposed to capture rhythm, the durational variability of speech. The most commonly analyzed metrics include: %V, the percentage of the total duration of vocalic intervals within an utterance; ΔC, the standard deviation of the duration of consonantal intervals within each utterance; ΔV, the standard deviation of the duration of vocalic intervals within each utterance [8]; VarCoC and VarCoV which normalize ΔC and ΔV respectively for speaking rate [9]; and nPVI-V [10] and rPVI-C [11] which exploit the derivational contrast between pairs of consecutive vocalic or consonantal intervals.

[12] examined the rhythmic structure of the Goldilocks and the Rainbow passages produced by New England, Mid-Atlantic,
Northern, Midland, Southern, and Western American English speakers and found that Southerners exhibited smaller VarcoC and rPVI-C values than New England speakers, but had greater ΔV, VarcoV, and rPVI-V values and relatively longer %V than Northern and/or Western speakers.

Following up on [12], [13] investigated the seven rhythmic metrics of the Goldilocks passage produced by Midland and Southern English speakers, and also found that Southerners produced greater ΔV and longer %V than Midland speakers.

1.3. Articulatory vowel space and speech intelligibility

A classic study by [14] proposed several metrics to measure vowel space. First, the vowel space area, defined by the F1 and F2 means for each vowel under investigation. Second, the vowel formant dispersion, defined as the mean Euclidean distance between individual vowel’s F1 and F2 values and the center point of each speaker’s vowel space. More recently, the convex hull encompassing all measured vowel tokens for each speaker has emerged as another important vowel space metric [15]. For all metrics, a larger value indicates a more expanded vowel space, and theoretically, higher degree of speech intelligibility.

Results from perception tasks, however, have exhibited a mixed profile regarding the correlations between vowel space and speech intelligibility in adult speech. While [14] found that more intelligible speakers had larger vowel triangle areas and more extended formant dispersion, [15] noted that more intelligible speakers in fact had smaller vowel polygon and that formant dispersion was not a reliable predictor of speech intelligibility. For convex hull, more intelligible speakers tended to have larger convex hulls [15].

1.4. Aims of the study

The overarching goal of this paper is to relate dialect bias (if any) in ASR to speech rhythm and vowel space, with a focus on Midland and Southern American English. We had three general hypotheses. First, based on [5, 6], we expected to see higher error rates in Southern speech than in Midland English. Second, since studies have shown that Southern English is rhythmically distinct from other dialects [12, 13], we anticipated that regional rhythmic features in Southern speech may cause difficulty to the machine, leading to higher error rates. Last, we seek to relate vowel space to speech intelligibility and then to ASR performance, such that more expanded vowel space, which is theoretically more intelligible, would lead to lower error rates.

We acknowledge that speech recognition is mainly based on spectral (formant) structure. Nevertheless, it is worth exploring the role of durational and spatial cues in speech recognition, which may yield insights into the acoustic modeling of ASR systems.

2. Analyses and results

Nine Midland participants (4 females + 5 males, ages 19-22) and 10 Southern participants (5 females + 5 males, ages 19-61) participated in an online speech production study. All were monolingual native speakers of American English who had lived in their own dialect region until the age of 18. Both parents of each participant were also raised in the same region.

The experiment was conducted by two native speakers of American English. All participants completed an interview session, followed by multiple reading tasks. The first paragraph of the Goldilocks fairytale was analyzed in this paper.

We tested our hypotheses in three sets of analysis: ASR performance, the relationship between rhythm and error rates, and the association with vowel space and error rates.

2.1. ASR performance

The unprocessed audio files were uploaded to an open-access website, DARLA [16, 17], to generate transcripts. Then, each audio file was segmented into individual intonational phrases (IPs) signaled by a perceivable following pause and final pitch lowering [18]. Next, we aligned the ground truth with machine-generated transcript for each IP. ASR performance was evaluated by computing the WER per IP.

In total, 292 IPs were analyzed. A linear mixed effects model was built for WER, with dialect as the fixed factor and speaker as the random effect. The results show a somewhat higher mean error rate for Southern dialect (59.5%) than for Midland dialect (47.2%), and this difference was marginally significant (p = .0987).

![Figure 1: WER by dialect.](image1)

2.2. Rhythm and WER

To measure rhythm, each IP was segmented into individual consonant or vowel intervals based on the first author’s auditory impression and visual inspection of the waveforms and spectrograms in Praat 6.0.43 [19]. A reliability check on the segmentation was performed by a trained labeler who is a native speaker of American English. Next, a Praat script [20] was used to extract the duration of all labelled intervals for analysis. Here we considered %V and ΔV because [13] reported longer %V and larger ΔV in Southern speech than in Midland English (Figure 2).

![Figure 2: Durational differences in vowel intervals by dialect.](image2)

A linear mixed effects model for WER was built to test the effects of %V and ΔV on ASR performance. Speaker was included as the random effect. Although error rates increased when the IPs are produced with longer %V, the effect was not statistically significant (p = .116). No clear trend between error rates and ΔV was observed (p = .798).
2.3. Vowel space and WER

This section focused solely on the vowel intervals. Measured vowels were drawn from lexically stressed syllables in content words, pronouns, and prepositions in each IP. As with [15], vowels adjacent to nasals, rhotics, or laterals were eliminated. In total, 11 phonemes /ɪ, i, e, æ, ə, ɔ, o, u, ʌ/ comprising 773 vowel tokens were analyzed. The same Praat script [20] was used to extract F1 and F2 values from the midpoint of each labelled vowels. These formant values were converted to a perceptual scale using the bark transform [21] and then z-score normalized within each dialect region (see [15]) prior to statistical analysis and graphing. Note that in this analysis, vowel space and error rates were calculated at the speaker level, rather than at the IP level.

Three vowel space measures were considered: (1) the area of the convex hull encompassing all measured vowels; (2) vowel formant dispersion; and (3) the area of the \( \Delta V \) polygon. The results show that overall, Southern speakers had larger convex hulls and vowel polygons than Midland speakers, whereas Midland speakers had more expanded formant dispersion than Southerners (Table 1).

Table 1: Mean normalized vowel space (Bark) by dialect.

<table>
<thead>
<tr>
<th>Dialect</th>
<th>Measure</th>
<th>Convex hull</th>
<th>Vowel formant dispersion</th>
<th>Vowel polygon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midland</td>
<td>(n = 9)</td>
<td>6.36</td>
<td>1.63</td>
<td>1.68</td>
</tr>
<tr>
<td>Southern</td>
<td>(n = 10)</td>
<td>6.68</td>
<td>1.53</td>
<td>1.91</td>
</tr>
</tbody>
</table>

When relating the three vowel space measures to WER, some interesting trends emerged. First, as convex hulls increase, the error rates decrease for Midland speakers (Figure 3b). Second, as formant dispersion becomes more expanded, the error rates slightly decrease for Midland speakers (Figure 3b). We also found that, as the vowel polygons increase, the error rates decrease for Southern speakers (Figure 3c).

Figure 3: Tendencies between normalized vowel space measures (Bark) and WER.

3. General discussion

In general, the results adhere with previous studies [2, 5, 6], suggesting that ASR systems are less accurate in decoding Southern speech than in recognizing other dialects of American English. This raises the question of why Southern speech is more error-prone to machines than other dialects? We posit that rhythm and vowel space may help account for the dialect gap.

Two rhythmic metrics (%V and \( \Delta V \)) were considered. Given that the Southern dialect exhibits longer %V and greater \( \Delta V \) (i.e., more variable duration in vocalic intervals), we expected to see positive correlations between the two durational metrics and error rates. For %V, the results suggest a trend that when IPs are produced with longer %V, the error rates increase. It is important to emphasize that while Southern speakers tend to have a slower speaking rate [12, 22, 23], longer %V in the Southern dialect is independent of speaking rate because there is no significant difference in speaking rate in our sample of Southern and Midland speakers [14]. It could be that the longer %V in Southern speech may cause some “misalignment” in timing during recognition, thus leading to higher error rates. For \( \Delta V \), the results did not support our hypothesis because greater vowel duration variability did not directly associate with higher error rates. Taken together, the relationship between rhythm and error rate is not very straightforward. More research is needed to theorize the role of durational cues in speech recognition.

We also measured vowel space in terms of the convex hull, formant dispersion, and vowel polygon. Since more intelligible speech has been found to be associated with more expanded vowel space [14, 15], we anticipated that ASR systems would also benefit from vowel expansion, leading to better performance. Along this line of reasoning, we expected to see
smaller vowel space in all three vowel measures in Southern speech. Smaller vowel space makes Southern speech less intelligible, therefore leading to higher error rates. In contrast, Midland speakers may produce larger vowel space. This makes their speech more intelligible, thereby lowering the error rates. The results, however, offer very limited support for our hypotheses – while Midland speakers indeed had more expanded formant dispersion, Southern speakers in fact had larger convex hulls and vowel polygons than Midland speakers (Table 1).

Larger average vowel space among Southerners may be accounted for by a higher level of linguistic consciousness. To illustrate, Southern U.S. English exhibits distinct phonological and morphosyntactic features from General American English, and folk perception often conveys stereotypes about the South [24, 25, 26]. We noticed that overall, our southern participants spoke naturally and were highly engaged in the interview session. But when proceeding to the reading task, a few speakers behaved very differently from the interview session by slowing down their speech, an indicator of hypercorrection in speech. This may explain why Southern speakers had larger vowel space. Midland speakers, on the other hand, did not show style shift in the experiment. Still, it remains unclear why ASR performance is worse for the Southern group. A tentative explanation is that reporting group mean may have masked some important information, which makes the interpretation difficult.

To tackle this issue, we turn to mean vowel space by individual speakers to gain a broader picture of how vowel space may be related to ASR performance. Now, we noticed that for Midlanders (lower error rates), as convex hulls and formant dispersion increase (i.e., more expanded space), the error rates decrease (Figures 3a, 3b). The findings align with our hypothesis that as vowel space increases, the speech is more intelligible. Increased intelligibility then enhances ASR performance. Critically, when relating ASR performance from the present study to literature on human perception, our results are in line with [15] who noted that more intelligible speakers had larger convex hulls and with [14] who found that more intelligible speakers had more extended formant dispersion.

Vowel polygons should be treated with caution because we did not see a negative trend, in which as vowel polygons increase, error rates decrease in Midland speech. Instead, the negative trend was observed in Southern speech (Figure 3c), which in fact showed poorer ASR performance. When relating our results to human perception studies, we noticed that while [14] found a positive correlation between vowel triangle area and intelligibility, [15] in fact reported a negative correlation between vowel polygon and intelligibility. Taken together, it seems that vowel triangle/polygon is a less reliable spatial cue for both human and machine speech recognition.

4. Summary

This paper attempts to explore whether ASR performance is influenced by rhythm and vowel space. The results did not return reliable relationships between durational cues (%V and ΔV) and ASR performance. For vowel space, the results hinted that convex hull and formant dispersion are more reliable spatial cues than vowel polygon for speech recognition.

These results, albeit suggestive, point out some promising directions for future work on improving acoustic modeling in speech recognition.

In the future, more data and larger sample of speakers will be needed to further test and revisit our proposal. In addition, analysis of vowel identification accuracy [27] may help spot algorithmic biases in the systems.

5. Acknowledgements

We thank our participants for providing the research data and Carly Danielson and Aleah Combs, former lab managers at Penn State’s Center for Language Science, who helped conduct the online experiments. We also thank Penn State’s Center for Language Science for their financial support for this study.

6. References


