Investigating the Influence of Virtual Peers as Dialect Models on Students' Prosodic Inventory

Samantha Finkelstein[1], *Stefan Scherer*[2], *Amy Ogan*[1], *Louis-Philippe Morency*[2], *Justine Cassell*[1]

[1]Human-Computer Interaction Institute, Carnegie Mellon University,
Pittsburgh, Pennsylvania, United States
[2]Institute for Creative Technologies, University of Southern California,
Playa Vista, California, United States

slfink@cs.cmu.edu

Abstract

Children who speak non-standard dialects of English show reduced performance not just in language-oriented topics in school but also in math and science. Technological solutions have been rare exactly because of the non-mainstream nature of their talk, and hence the difficulty in automatically recognizing their speech and responding to it with, for example, computer tutors. In order to work towards overcoming this achievement gap, in this work we investigate African American students' prosodic inventories in different contexts as a first-step towards building a system that will be able to automatically recognize, and respond to, the dialect in which a child is speaking. We presented children with recordings of a peer (confederate) speaking in either African American English (AAE) or Mainstream American English (MAE) during both a social task and a science task. We found that children showed decreased prosodic variation and peak slopes during speech segments which did not contain AAE features, resulting in more monotone and breathy utterances than when they are speaking in AAE. We also found that children who were speaking with a "peer" who uses AAE have increased articulation rates, energy, and pitch variation. We discuss potential interpretations of these results that are important to the design of a system to support linguistic diversity and decrease the achievement gap.

Index Terms: Virtual peers, dialect model, prosodic inventory

1. Introduction

While many schools are promoting cultural diversity amongst their students, the linguistic diversity that accompanies this movement may have a harder time being supported in schools [1]. Despite the powerful ties language has to cultural identity, there is a foundation of research on children's oral and written language development indicating an association between speaking non-standard dialects of English and reduced literacy skills and test scores [2, 3]. This issue becomes particularly salient in inner-city schools, where non-standard dialects, particularly African American English (AAE) may be one of, or the only, dialect that students speak, and where teachers may have quite strong beliefs about the "inferiority" of AAE. The Black-White achievement gap is well-known and persistent in the American educational system [4], and providing support in schools for students who speak non-mainstream dialects may help to improve students' test scores and promote achievement.

One method of addressing the issue may be through supporting students in acquiring awareness of dialect, and the different social contexts in which different dialects may be most successfully employed. [5] and shifting [6]. In its favor, these strategies suggest that the ability to selectively switch between MAE and AAE in appropriate social contexts is indicative of greater meta-linguistic awareness, which may help students succeed in school more generally. Children in diverse linguistic communities may be presented with more opportunities not only to notice differences between spoken dialects, but also to practice shifting between them for specific contexts. In particular, children's acquisition of and proficiency with dialect shifting ability may be difficult to acquire if they do not have classmates who happen to speak in MAE.

In response to this, we have previously built an embodied conversational agent, Alex, who was designed to strategically employ either an AAE or MAE dialect during interactions with students [7]. Our initial studies with Alex, who began speaking in AAE during a collaborative playing task and then switched into MAE for a science presentation task, found that 3rd grade students who spoke with Alex during a science presentation task used less AAE than did students who completed the same task with a human peer partner. Currently, we are working with both elementary school teachers and students to expand on our design to create an agent that can be deployed in classrooms as a stable support for students learning to dialect shift. As presented in our recent work on collaboration with virtual peers ([8], these agents may fill a lacuna in classrooms by having the ability to automatically

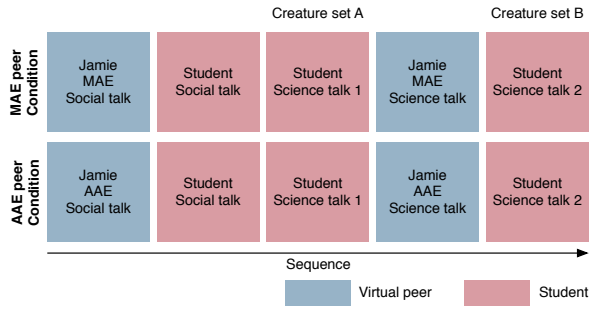

Figure 1: Visualization of the two recording conditions MAE speaking peer and AAE speaking peer. Blue segments indicate virtual peer's speech and red segments indicate students' segments.

recognize certain student skills and weaknesses, and addressing them through dialogue. Though it may be impractical (or impossible!) for teachers to focus on every students' dialectal level individually while also teaching the primary curriculum, agents who are already working with students may be able to automatically identify the dialect students are using, and respond accordingly. In this work, we focus on two questions in this area: (1) are we able to automatically detect when African American students are speaking in AAE or MAE at the utterance level, and (2) will students accommodate to virtual peers who are speaking in either AAE or MAE?

2. Study

2.1. Study design

In this work, we perform an acoustic analysis of 3rd and 4th grade African American students who were recording speech for a virtual partner during both a social task and a science task. Students were randomly-assigned virtual partners who all provided identical content either delivered in AAE or MAE. The virtual partner, Jamie, was represented by an African American avatar, and, it was explained, was a student in their grade at another school just like theirs. Jamie's talk was recorded by a native speaker of African American and Mainstream English whose voice was pitch-shifted to sound younger. Students heard Jamie give a 3.5 minute social introduction, as well as complete a 3.5 minute science comparison activity. In the study procedure, Jamie provided first provided a social introduction, and then student free-recorded their own 3-4 minute social introduction for Jamie. Afterwards, the student completed a science task that required them to free-record observations, hypotheses, comparisons, and scientific questions about a set of creatures we provided. Jamie then provided a 3.5 minute recording of the science with a different set of creatures. After listening to Jamie's recording, the students were asked to complete the same science task again with another set of creatures. This procedure allows us to investigate students' change in dialect features based on the dialect pattern of their virtual partner between a social context and a science context, as well as between two science tasks separated by a virtual peer modeling science talk.

Feature	MAE peer	AAE peer
$\mathbf{f_0}$	212.20 ($\pm$ 63.55)	210.28 ($\pm$ 58.62)
Energy	-43.60 ($\pm$ 17.42)	-44.08 ($\pm$ 17.82)

Table 1: Feature comparison of virtual peer's speech across conditions. With f_0 being the virtual peer's fundamental frequency. Mean and standard deviation values over utterances are reported.

Jamie's voice did not vary strongly over the two recording conditions MAE and AAE, in order to control for entrainment effects, see Table 1 for details.

2.2. Hypotheses

There are currently no ASR systems capable of dealing with childrens spontaneously produced AAE. In this work, then we focused our acoustic analysis primarily on suprasegmental features, and investigated two primary hypotheses: (1) There will be significant differences in the acoustic signal between utterances that had AAE features and utterances that did not have AAE features, and (2) free-speech recorded in a social setting to an AAE-speaking virtual peer would be significantly different acoustically from speech recorded to an MAE-speaking virtual peer.

2.2.1. Hypothesis 1: MAE vs. AAE

We expected to see a series of significant suprasegmental acoustic differences between child free-speech segments that were annotated as either having at least one phonetic or morphosyntactic AAE feature in the clause(AAE), or no AAE features in the clause(MAE). These annotated speech samples came from the science task component of our study. Although overall African Americans use a wider pitch range than European English speakers, particularly during free (rather than read) speech [9], it is unknown if bidialectal African American speakers will have different pitch ranges when they speak different dialects. We hypothesized that indeed increased pitch shifts would occur primarily during speech with AAE features. Additionally, we expected to find other differences when the children were speaking MAE, such as voice quality variation between the two dialects, of the kind that has been correlated with decreased comfort level or increased cognitive load.

2.2.2. Hypothesis 2: Speech to an MAE vs. AAE-speaking peer

We additionally expected that there would be differences in the suprasegmental acoustic speech signals between students who were recording a social introduction to a virtual partner who spoke AAE vs. MAE. Because there are no strong differences between the acoustic signals in the social recordings given by the virtual partners, differences in children's speech would likely be for social or cognitive reasons, rather than unconsciously entraining to the acoustic features of the model.

3. Feature extraction

In this section we briefly discuss the prosodic features used in the statistical analysis. We chose these features as they have proven to be robust representatives of various prosodic phenomena in previous analysis. The features include the following measures:

- **Energy (in dB)** is a measure of the intensity of the speech signal. Higher values indicate louder speech.
- **Articulation rate** is calculated by identifying the number of syllables per second. The syllables are detected by identifying vowels in the speech.
- **Fundamental frequency (f_0)** is the base frequency of the speech signal. It is the frequency the vocal folds are vibrating at during voiced speech segments.
- **Peak slope** is a measure suitable for the identification of breathy to tense voice qualities. Values closer to zero are considered as more breathy (see Figure 2).
- **Spectral stationarity** is a value that captures the fluctuations and changes in the voice signal. High values indicate a stable vocal tract and little change in the speech (e.g. during a hesitation or sustained elongated vowels).

The following sections detail each acoustic feature.

3.1. Energy in dB

The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms (i.e. 100Hz sample rate). This speech window $w(t)$ is filtered with a hamming window and the energy

$$e(t) = \sum_{i=1}^{|w(t)|} w_i(t)^2 \tag{1}$$

is calculated and converted to the dB-scale

$$e_{dB}(t) = 10 \cdot \log_{10}(e(t)). \tag{2}$$

3.2. Articulation rate

Detection of syllable nuclei to calculate the articulation rate were made using the method introduced in [10], which is based on intensity peak detection of voiced segments of speech in Praat. On an abstract level, the script introduced in [10] follows a simple peak detection using the signal intensity. Peaks that are preceded and followed by considerable dips in intensity (based on a predefined threshold, e.g. -2 dB) are considered syllable nuclei candidates. All detected peaks, that are not voiced are removed from the selection in order to obtain the syllable nuclei without the need of a preceding transcription.

3.3. Fundamental frequency f_0

In [11], a method for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions, is introduced. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process removes strong influences of noise and vocal tract resonances. For each $r(t)$ the amplitude spectrum $E(f)$ is computed, showing peaks for the harmonics of f_0, the fundamental frequency. Then, the summation of residual harmonics (SRH) is computed as follows [11]:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)], \tag{3}$$

for $f \in [f_{0,\min}, f_{0,\max}]$, with $f_{0,min} = 50$ and $f_{0,max} = 300$. The frequency f for which $SRH(f)$ is maximal is considered the fundamental frequency of this frame. By using a simple threshold θ, the unvoiced frames are discarded as in [11].

3.4. Peak slope

This voice quality parameter is based on features derived following a wavelet based decomposition of the speech signal [12]. The parameter, named *peak slope*, is designed to identify glottal closure instances from glottal pulses with different closure characteristics. It was used to differentiate between breathy, modal, and tense voice qualities in [13]. The following equation is used for decomposing the speech signal:

$$g(t) = -\cos(2\pi f_n t) \cdot \exp(-\frac{t^2}{2\tau^2}), \tag{4}$$

where the speech signal $s(t)$ is convolved with $g(\frac{t}{s_i})$, and $s_i = 2^i$ and $i = 0, 1, 2, \ldots, 5$. This essentially is the application of an octave-band filter bank with the center frequencies being: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz and 250 Hz. Then the local maximum is measured at each of the signals obtained from the decomposition and a regression line is fit to these peaks. In Figure 2, it can be seen that for an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities the slope of the

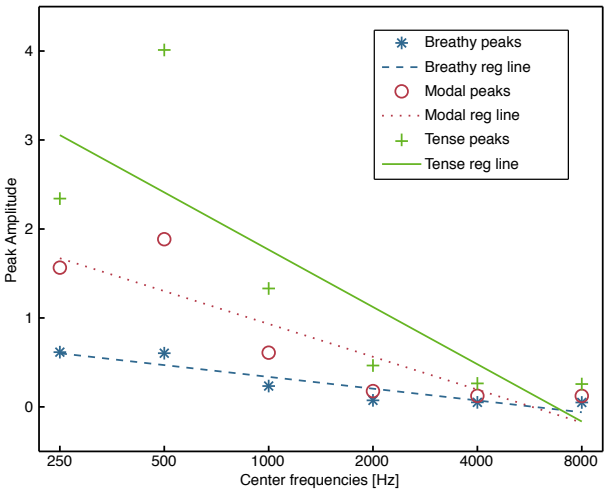

Figure 2: Peak amplitudes, from signals with different center frequencies, with regression lines for an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities

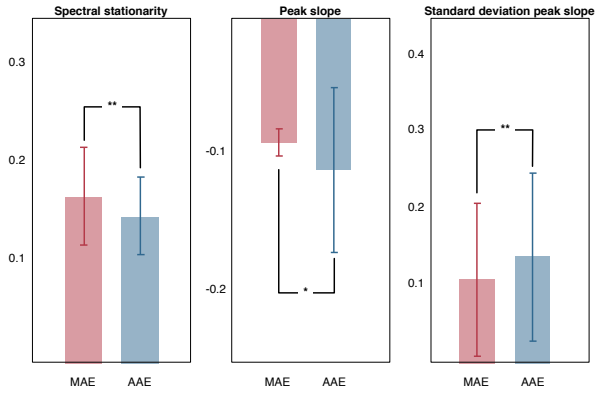

Figure 3: Plots with error bars comparing observations in mainstream American English utterances (MAE) vs. African American English utterances (AAE). Three features are displayed including spectral stationarity, peak slope, and standard deviation of the peak slope. Significant differences are highlighted with labeled brackets, * for $p < .05$ and ** for $p < .01$ in unpaired t-tests.

regression line is clearly different. Hence, the peak slope parameter is simply the slope coefficient of the regression line. In the original publication [12] this was carried out on individual phone segments. In the current study it is carried out on the frame level of the entire speech signal. A frame length of 32 ms and shift of 10 ms (i.e. 100Hz sampling rate) is used.

3.5. Spectral stationarity

To characterize the range of the prosodic inventory used over utterances, we make use of the so called *spectral stationarity* measure ss. This measurement was previously used in [14] as a way of modulating the transition cost used in the dynamic programming method used for f_0 tracking. Spectral stationarity, ss is measured with:

$$ss = \frac{0.2}{\text{itakura}(f_i, f_{i-k}) - 0.8} \quad \in [0, 1], \qquad (5)$$

where itakura(.) is the Itakura distortion measure [15] of the current speech frame f_i and f_{i-k} is the previous frame with $k = 1$. We use a relatively long frame length of 60 ms (with as shift of 10 ms; sampling rate 100Hz) and frames are windowed with a Hamming window function before measuring ss. The long frame length was used in the attempt to characterize relatively long periods of maintained vocal tract articulation. ss is close to 1 when the spectral characteristics of adjacent frames are very similar and goes closer to 0 if the frames show a high degree of difference.

4. Statistical evaluation

For the statistical analysis we employed standard unpaired t-tests comparing if the mean values of the prosodic features for one population are the same as those of the other population. In particular, we test the two hypothesis of Section 2:

- In Section 4.1 a more fine-grained analysis is conducted, by comparing the utterances of the science tasks of all kids that are annotated as containing AAE fragments (i.e. morphosyntactic or phonological) and the utterances that are annotated as being free of any AAE fragments, in order to test Hypothesis 1.

- Section 4.2 compares the prosodic parameters introduced in Section 3 on a session level, in order to test Hypothesis 2. In particular the social talk of the kids recorded within the MAE speaking peer condition is compared to the social talk of kids in AAE speaking peer conditions.

4.1. Hypothesis 1: MAE vs. AAE

Thirty of the science talk sessions have been annotated at an utterance level. Each utterance additionally received a label whether or not it contains one or more AAE elements (i.e. phonological, morphosyntactic or both AAE segments). In total we found 420 utterances containing AAE fragments and 414 without. The average length of AAE utterances is 6.19 s (standard deviation: 4.07) and the average length of MAE utterances is 5.59 s (standard deviation: 3.15). For this analysis, we compare the utterances with AAE elements with those that are purely MAE using the aforementioned prosodic parameters. The results are summarized in Table 2, and three features are visualized in Figure 3.

Feature	Mean		Standard deviation	
	MAE	AAE	MAE	AAE
Fundamental frequency	230.96 (±24.42)**	226.57 (±22.30)**	45.73 (±17.06)	45.386 (±12.64)
Spectral stationarity	0.17 (±0.05)**	0.15 (±0.04)**	0.13 (±0.02)	0.13(±0.02)
Energy	-32.56 (±4.42)**	-31.57 (±4.62)**	3.59 (±2.98)	3.68 (±2.43)
Peak slope	-0.09 (±0.01)*	-0.11 (±0.06)*	0.11 (±0.11)**	0.14 (±0.11)**
Articulation rate	0.23 (±0.20)	0.22 (±0.18)	0.06 (±0.10)**	0.09 (±0.12)**

Table 2: Statistics and comparison of prosodic values for AAE and MAE annotated utterances for all conditions. Mean and standard deviation (Std.) values over utterances are reported. For each the corresponding mean and standard deviation values are reported and separated by /. Significant differences are marked with * for $p < .05$ and ** for $p < .01$ in unpaired t-tests.

We could find significant results for the mean values of the fundamental frequency (f_0), spectral stationarity, speech energy and peak slope. Further, significant results were found for the standard deviations of values for peak slope and articulation rate.

In particular, the average f_0 varies significantly (MAE: 230.96 vs. AAE: 226.57; p = 0.006). The average spectral stationarity is significantly greater for utterances containing no AAE (MAE: 0.17 vs. AAE: 0.15; p = 0.005). Also, there is a significant difference in average energy measurements (MAE: -32.56 vs. AAE: -31.57; p = 0.002). Further, average (MAE: -0.09 vs. AAE: -0.11; p = 0.02) and standard deviations (MAE: 0.11 vs. AAE: 0.14; p = 0.001) of peak slope values vary significantly, and the standard deviation of articulation rate is significantly different for the two sets of data (MAE: 0.06 vs. AAE: 0.09; p = 0.002). All other prosodic values did not show significant differences.

4.2. Hypothesis 2: Speech to an MAE vs. AAE-speaking peer

In this section we report the statistical analysis of the prosodic parameters for the social talk segments, i.e. Hypothesis 2. Social talk segments in the MAE speaking peer condition (8 segments in total) are compared to those of the other conditions (20 segments in total). As mentioned in Section 2, the social talk segments in those conditions are distinct from one another solely with respect to the speech the virtual peer uses (MAE vs. AAE). The statistical results are reported in Table 3, and three features are additionally visualized in Figure 4.

Significant differences between the conditions could be found for the mean values of spectral stationarity, speech energy, and articulation rate. Additionally, the standard deviations for fundamental frequency (f_0), spectral stationarity and energy were statistically significantly different from each other.

Table 3, shows that the average f_0 of the children in the two different conditions is not significantly different (Condition 1: 227.75 vs. Condition 2+3: 235.18; p = 0.46), however, the standard deviations are (Condition 1: 43.63 vs. Condition 2+3: 64.80; p = 0.008). For the

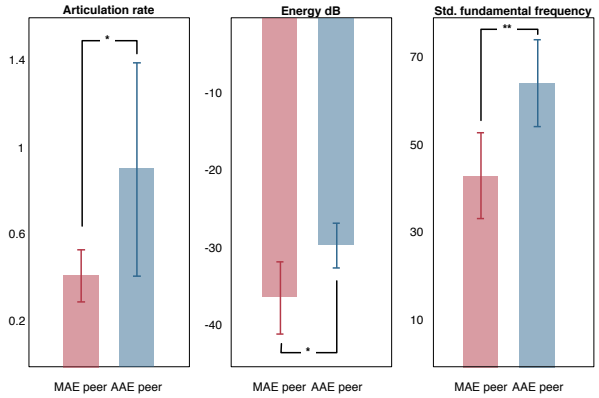

Figure 4: Plots with error bars comparing observations in MAE speaking peer condition vs. AAE speaking peer conditions. Three features are displayed including articulation rate, energy, and standard deviation of the fundamental frequency. Significant differences are highlighted with labeled brackets, * for $p < .05$ and ** for $p < .01$ in unpaired t-tests.

spectral stationarity the difference is significant for both average (Condition 1: 0.15 vs. Condition 2+3: 0.23; p = 0.04) and standard deviation (Condition 1: 0.13 vs. Condition 2+3: 0.15; p = 0.04). Similarly, the average energy (Condition 1: -36.15 vs. Condition 2+3: -29.45; p = 0.02) and standard deviations (Condition 1: 7.24 vs. Condition 2+3: 2.44; p = 0.001) vary significantly. Additionally, the average articulation rate (Condition 1: 0.42 vs. Condition 2+3: 0.91; p = 0.03) is significantly different for the two conditions. All other prosodic values did not show significant differences.

5. Discussion

As reported in Section 4, we found several statistically significant differences in prosodic parameters in students' speech dependent on a sequence level based on the different conditions, as well as on an utterance level across the conditions. These findings support the research hypotheses presented in Section 2, yielding two main find-

Feature	Mean		Standard deviation	
	MAE peer	AAE peer	MAE peer	AAE peer
Fundamental frequency	227.75 (±18.38)	235.18 (±24.74)	43.630 (±9.75)**	64.80 (±10.01)**
Spectral stationarity	0.15 (±0.03)*	0.23 (±0.02)*	0.13 (±0.01)*	0.15 (±0.03)*
Energy	-36.15 (±4.65)*	-29.45 (±2.90)*	7.24 (±2.18)**	2.44 (±1.57)**
Peak slope	-0.16 (±0.07)	-0.12 (±0.10)	0.27 (±0.17)	0.15 (±0.09)
Articulation rate	0.42 (±0.12)*	0.91 (±0.49)*	0.29 (±0.17)	0.58 (±0.56)

Table 3: Statistics and comparison of prosodic values for social talk segments in recording MAE speaking peer condition vs. AAE speaking peer conditions. Mean and standard deviation (Std.) values over segments are reported. For each the corresponding mean and standard deviation values are reported and separated by /. Significant differences are marked with * for $p < .05$ and ** for $p < .01$ in unpaired t-tests.

ings: (1) there are significant differences in the acoustic patterns of clauses produced by an African American child depending on whether that clause contained or not at least one phonetic or morphosyntactic AAE feature. Interpretations of these findings are presented below, and (2) young African American children are sensitive to the dialect of a virtual partner and it affects their prosody.

5.1. Hypothesis 1: MAE vs. AAE

Based on the manual annotation of 30 science talk sequences across all conditions, we analyzed prosodic parameters of utterances that either contained or not at least one AAE phonological and morphosyntactic feature. These sequences were all taken from a science task, where students were making observations and hypotheses about imaginary creatures. Indeed, as predicted, we found several parameters, as reported in Section 4.1, that vary significantly between the two groups of utterances. The exact differences are summarized in Table 2 and three distinctive measures are visualized in Figure 3. In particular, we found that the spectral stationarity is significantly higher for pure MAE utterances, which indicates that there is less variation in prosody present in those utterances than in the ones containing AAE fragments. We additionally find that MAE utterances have a lower peak slope, and are thus *breathier* than their AAE counterparts.

If these acoustic differences are reliable, it may allow us to train a model to automatically characterize the differences between AAE and MAE utterances, and then automatically recognize when a child is switching between AAE and MAE during free speech. In turn, this will allow us to build a system that responds to a child's dialect use with appropriate feedback, adapting the interaction with the child based on the child's dialect use. While it is known that African American free-speech typically contains more pitch variation than European American free-speech [9], our results additionally show that particular utterances from one single African American child that contain an AAE feature have more variation than utterances without any AAE features spoken by that same child. These results are promising for the development of a system to perform dialect identification; however, before the systems adaptive responses can be implemented, further analyses are needed to determine what is causing the differences we are seeing.

5.2. Hypothesis 2: Speech to an MAE vs. AAE-speaking peer

We found that the preceding interaction with an MAE speaking peer influenced the children's prosody in the subsequent social talk session (see Section 4.2). Our results demonstrate that children speaking with an MAE-speaking partner employed more limited pitch ranges (i.e. smaller variations in fundamental frequencies), limited loudness (i.e. reduced speech energy), and reduced speaking rates (i.e. lowered articulation rate), as shown in Table 3 and Figure 4. Though there was slightly more pitch variation in the MAE social recording we gave the students, children who were speaking to an AAE-speaking peer responded with more pitch variation than children who were speaking to an MAE-speaking peer. This indicates that students were not simply entraining to the pitch variation presented in the model.

5.3. Possible interpretations

We propose three possible interpretations of the data to explain these acoustic differences: *(1) the linguistic hypothesis, (2) the social hypothesis, or (3) the cognitive hypothesis.*

Our linguistic interpretation proposes that there are intrinsic features within the AAE and MAE dialects that produce these prosodic differences. In this case, it would be the expression of AAE features that produces prosodic differences, regardless of other social or cognitive phenomena. This would be the easiest interpretation for implementing automatic dialect detection at the utterance level with young African American students. However, dialect use is difficult to separate from identity and culture, which leads us to assess these results from a social lens.

It could be that African American children who are

talking to an MAE-speaking peer feel intimidated, different, or out of place. They may also feel like they need to try and produce more MAE to match that peer, and as such, they force themselves to hyper-articulate. This could explain the low-energy, slower, more monotone recordings of the children in the MAE social condition. If this interpretation holds true, it may be important to train models on African American children recording for an MAE-speaking virtual peer with whom they have already built a rapport and feel comfortable. If the social interpretation is true, our results may accurately identify dialect features in the beginning of an interaction, but may become less reliable over time as the child builds a comfortable relationship with the agent.

Finally, the cognitive interpretation would expect that the decreased speed, decreased energy, and increased monotony of the recordings to MAE-speaking peers is due to the increased difficulty of speaking a more foreign dialect. Children partnered with an MAE-speaking peer may have been exerting increased effort to speak MAE, causing these acoustic effects. Additionally, it could be that children spoke without using any AAE features when they were talking about more complicated science material, and the increased breathiness and reduced pitch variation during MAE utterances can be just due to the increased difficulty of the science, and not the difficulty of speaking the dialect. If the cognitive hypothesis is true, our results may hold in the beginning of an interaction but, as mentioned for the social hypothesis, may become less applicable over time as children master speaking in an MAE dialect.

Further analyses need to be done to identify which, or how many, of these interpretations are contributing to the results presented in this paper. Indeed, the type of system we can build varies widely depending on the underlying reasons behind children's dialect shifting.

6. Conclusion

Technological interventions to support linguistically-diverse students have great potential to increase student success, but may hinge on being able to automatically recognize certain features of children's speech even when ASR is still beyond reach. In this work, we analyze a corpus of students recording monologues for an imagined peer as a first step to determining if we can build an automatic recognizer of the dialect African American children are employing on the utterance level. We found that children who are recording speech for an MAE-speaking virtual partner talk more slowly, talk more quietly, and have reduced pitch variation than students who are recording for a virtual partner who speaks AAE. We also found that, across all conditions, students had more pitch-variation and less "breathy" tones during utterances that had at least one AAE feature than utterances that did not have any AAE features.

These results are promising, as they indicate it may be possible to train a model to recognize the different dialects African American children may be using in any given utterance, which could have important pedagogical implications for the design of a technological intervention to support students in dialect-shifting. If a pedagogical system can reliably recognize when a student is speaking AAE, it may be able to provide feedback in the moment, or change its approach to increase students' awareness of their dialect. However, because dialects are complicated expressions of identity, even when we provide all children the same model of partner speech within every condition, it is not possible to identify the underlying reasons behind the acoustic differences we see in our results. The results we found may additionally be due to social artifacts (such as African American children feeling self-conscious and trying to hyper-articulate when speaking with an MAE-speaking peer, affecting their acoustic patterns) or cognitive artifacts (such as children choosing to speak in MAE for more complicated science content, which is affecting their acoustic patterns.) To hone in on what is causing these acoustic differences, we look towards annotating the content of the recorded speech to see if there are clues in the content of what the children are saying to help us interpret the prosodic differences shown in this corpus. Once we can accurately predict the dialect of the students, we look to create a virtual peer dialogue agent that will be able to identify the child's dialect, provide feedback to the child, and adjust the pedagogical approach to best support the student.

7. Acknowledgements

We are grateful for generous funding under NSF grants ALT-0735664 and IIS-1018055, and Heinz Grants C2902 and C2996. Additionally, many thanks to Anna Donahoe and Peter Schwartz who transcribed and annotated the data.

8. References

[1] R. R. Gann, "Language, conflict, and community: linguistic accommodation in the urban us," *Changing English*, vol. 11, no. 1, pp. 105–114, 2004.

[2] F. J. Morrison, H. J. Bachman, and C. M. Connor, *Improving literacy in America: Guidelines from research*. Yale University Press, 2005.

[3] R. R. Gann, "Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice," *Handbook of early literacy research*, pp. 97–110, 2001.

[4] C. Jencks and M. Phillips, *The black-white test score gap*. Brookings Institutional Press, 1998.

[5] A. Charity, H. Scarborough, and D. Griffin, "Familiarity with school english in african american children and its relation to early reading achievement," *Child Development*, vol. 75, pp. 1340–1356, 2004.

[6] C. Connor and H. Craig, "African american preschoolers' language, emergent literacy skills, and use of african american en-

glish: A complex relation," *Journal of Speech, Language, and Hearing Research*, vol. 49, pp. 771–792, 2006.

[7] J. Cassell, K. Geraghty, B. Gonzalez, and J. Borland, "Modeling culturally authentic style shifting with virtual peers," in *Proceedings of the 2009 international conference on Multimodal interfaces*, ser. ICMI-MLMI '09. ACM, 2009, pp. 135–142.

[8] S. Finkelstein, A. Ogan, and J. Cassell, "Real collaboration with a virtual peer: Results from co-design in an early elementary context," in *Proceedings of Workshop on Intelligent Support in Exploratory Environments*, 2012.

[9] A. Hudson and A. Holbrook, "A study of the reading fundamental vocal frequency of young black adults," *Journal of Speech and Hearing Research*, vol. 24, no. 2, pp. 197–201, 1981.

[10] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[11] T. Drugman and A. Abeer, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech 2011*. ISCA, 2011, pp. 1973–1976.

[12] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Proceedings of Interspeech 2011*. ISCA, 2011, pp. 177–180.

[13] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech and Language*, 2012.

[14] D. Talkin, "A Robust Algorithm for Pitch Tracking," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, pp. 495–517.

[15] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-23, pp. 67–72, 1975.