Alex: A Virtual Peer that Identifies Student Dialect

Samantha Finkelstein, Amy Ogan, Caroline Vaughn, and Justine Cassell

Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
{slfink, aeo, cvaughn, cassell}@cs.cmu.edu

Abstract. Our work in educational technology supports the deeper understanding of cultural phenomena that affect students in an educational environment. In this demo, we show a Virtual Peer system that supports collaborative science dialogue in two dialects: Mainstream American English and African-American Vernacular English, an often-stigmatized dialect. The most recently developed module in this system allows for the automatic classification of the students' dialect in order to adjust the VP's own dialect in response, and provide metalinguistic support. We are hopeful that this system will provide data to help close pervasive achievement gaps between students of different cultural backgrounds, while also providing important insights about the relationship between culture and learning that will impact educational technology design.

1 Introduction and motivation

Learning technologies until very recently focused on the cognitive processes involved in learning, while measuring how quickly or robustly the student gains understanding. More recently, cognitive scientists have recognized that other factors such as *metacognitive skills* (e.g., students' understanding of their own abilities as a learner) heavily influence student success. One type of metacognitive process, though, has been even less extensively explored in the context of learning technologies: students' metalinguistic awareness. Benveniste writes that metalinguistic ability refers to "the possibility of raising ourselves above language, of abstracting ourselves from it, of contemplating it, whilst making use of it in our reasonings and observations" (as cited in [1]). This metalinguistic knowledge plays a substantial role in the classroom, despite how rarely it is addressed with students [2].

Indeed, how students speak in the classroom, make meaning of the events around them, and structure explanations of this meaning have been identified as key factors in school success. However, these skills are evaluated with particular school-ratified perspectives on how to speak, even though all children do not share the same cultural perspectives implied by this mainstream school culture, nor do they all come to school speaking the same dialect of their native language. Dialect refers to a regional variety of a language distinguished by pronunciation, grammar, or vocabulary. The number of speakers of dialects worldwide is uncountable, but taskforces such as the European Charter for Regional or Minority Languages have been formed in response to frequent discrimination against the use of such regional languages and dialects, particularly in schools [3].

adfa, p. 1, 2011.

Additionally, because most learning technologies are designed and marketed for speakers of mainstream dialects, many children are at a disadvantage in using them. Indeed, our most recent results suggest that in the United States, children who speak African-American Vernacular English (AAVE) can learn more when educational technologies communicate in their own dialect [4]. In this work, we hope to make contributions with interactive technologies in two primary ways: by supporting students' outcomes using contextually- and culturally-appropriate dialects, while supporting students' development of their own meta-linguistic awareness to help them more intentionally consider their own language use.

2 Our Virtual Peer system

To investigate students' responses to different cultural stimuli, we develop virtual peer (VP) technologies that allow us to examine students' behaviors across dimensions including learning, collaboration, and cultural identity. Alex is a gender- and race-ambiguous VP partner, who works with children on a variety of age-appropriate science tasks. This allows children to develop perceptions of the child's ethnicity based on the agent's verbal and non-verbal behaviors, rather than surface level features of appearance [5].

Alex and the child interact in real-time in a shared-reality environment [6]. The child stands at a table, while the virtual peer is projected onto a large screen facing the child, supporting the illusion that the two are working together at the same table. We are then able to examine how children participate in a collaborative dialogue with agents who either share, or don't share, some subset of distinct culturally-based verbal and non-verbal features. In the original version of our system, Alex was able to speak in either Mainstream American English (MAE) or AAVE. The results of this work showed that even with a very simple language model – one in which the virtual peer speaks in AAVE during collaborative play and then switches into MAE for formal practice of a science presentation – students responded very positively to Alex. In fact, they shifted their own dialect use to better reflect the use of MAE demonstrated by the VP in the presentation practice – even more so than when they were working with human peers [7].

However, language use is a complex social process, and this type of simple language model does not have great ecological validity. Indeed, people often switch fluidly between language dialects and styles for many reasons, including the content of their speech, the situation context, and the race and dialect of their interlocutor [8]. Due to the meaningful social implications of dialect use, it is thus also critical that our virtual peer technologies are able to respond in an appropriate dialect in a way that addresses both the cognitive and social needs of the student.

In this demo, we will show the most recent addition to our system which will allow us to dramatically improve the nuance and complexity of our VP's language model by *automatically classifying* the dialect of a students' speech. We can then begin addressing the problem of modeling real-time dialect selection for the virtual agent by using the child's own speech as a meaningful input that guides dialect choices.

3 Dialect identification module

To determine the effects of dialect congruence in educational technologies, we need an automatic method of determining the amount of African American Vernacular English present in a child's speech. As a first step towards such a metric, we have created a dialect identification system that uses only paralinguistic speech features to recognize individual utterances as either AAVE or MAE.

We first used the openSMILE system [9] on a set of .wav files containing speech samples recorded in our two target dialects to extract three types of audio features: Mel Frequency Cepstral Coefficients (MFCCs), prosodic features, and voice quality features. These features are a subset of those used by Metze et al. [10] for emotion recognition. MFCCs are a description of a sound sample's cepstral space, and are commonly used in the areas of Automatic Speech Recognition, Speaker Identification, and Emotion Identification. The prosodic features that we extracted included F0 (i.e., pitch), energy, and duration. Our voice quality features were jitter (the derivative of the F0), shimmer (the derivative of the loudness), and voicing probability.

Using the Weka machine learning tool [11], these features were then used to train a Support Vector Machine model with a linear kernel. The resulting dialect identification system takes individual utterances as input and, using this model, classifies them as AAVE or MAE. This current system is a preliminary attempt at classifying speech into separate dialects, and ignores the existence of a dialect continuum. Future work will use this model to create a more fine-grained metric to place speech along this continuum rather than making a binary assignment.

The speech used to train and test the model was produced by an AAVE-MAE bilingual speaker who is a professional actor. Our speaker recorded 158 sentences in AAVE and 159 in MAE. The system was trained on 212 of these 317 utterances (evenly distributed between dialects) and tested on the remaining 105. The model classifies these 105 test utterances as either AAVE or MAE with 71.4% accuracy. While this result clearly leaves room for improvement, it is significantly above the baseline of 50%, and suggests that this purely paralinguistic system might serve as a foundation for future work that will take more types of features into account, such as word choice or grammatical differences between dialects.

The dialect identification model is currently independent from the Alex system for testing purposes, but we plan to integrate it into the complete virtual peer system. Such an integrated system will then have real-time information about the dialect of the student, and will be able to tailor its own dialect in response or provide better feedback to the student to support metalinguistic awareness.

4 Discussion and conclusion

Our work in technology-enhanced learning supports the deeper understanding of cultural phenomena that affect students in an educational environment. In this demo, we show the newest module in our virtual peer system that can automatically classify a students' speech as either the mainstream or vernacular dialect of English, with the dual goals of allowing our TEL system to adjust its own dialect in response and pro-

vide metalinguistic support. This dialect modeling addresses some of the current substantial problems for addressing language use in TEL environments, for example, that students fluidly switch between dialects based on contextual and social cues. The addition of this dialect identification module allows us to take the next steps to develop a more complex model for an agent's dialect selection, by eventually enabling the child's own dialect to be one of a number of features in a real-time decision about how the virtual agent should respond. Subsequently, this module allows us to evaluate the impact of dialect congruence on students' task performance compared to alternative dialect selection strategies. We are hopeful that the use of these systems will also provide data that will help to close the pervasive achievement gap between African Americans and Caucasian students that exists in the United States, while also providing important insights about the relationship between culture and learning that will impact the future of educational technology design.

5 References

1. Gombert, J. E. (1992). Metalinguistic development. University of Chicago Press.
2. Wheeler, R. S. (2006). "What do we do about student grammar–all those missing-ed's and-s's?". Using comparison and contrast to teach Standard English in dialectally diverse classrooms. *English Teaching: Practice and Critique*, 5(1), 16-33.
3. International conference on the European charter for regional or minority languages (1998). Retrieved from http://book.coe.int/GB/CAT/LIV/HTM/l1314.htm
4. Finkelstein, S., Scherer, S., Ogan, A., Morency, L.P., Cassell, J. (2012) "Investigating the Influence of Virtual Peers as Dialect Models on Students' Prosodic Inventory". in *Proceedings of WOCCI at INTERSPEECH 2013*, September 14-15, 2012, Portland, OR
5. Iacobelli, F. & Cassell, J. (2007). "Ethnic Identity and Engagement in Embodied Conversational Agents" In *Proceedings of Intelligent Virtual Agents*, Sept. 17-19, Paris, France.
6. Ryokai, K., Vaucelle, C., Cassell, J. (2003) "Virtual Peers as Partners in Storytelling and Literacy Learning". *Journal of Computer Assisted Learning* 19(2): 195-208.
7. Rader, E., Echelbarger, M. Cassell, J. (2011). "Brick by Brick: Iterating Interventions to Bridge the Achievement Gap with Virtual Peers". In *Proceedings of the CHI'11 Conference*, May 9-12, Vancouver, BC. 2971-2974.
8. Rickford, J. R. (1991). Variation theory: Implicational scaling and critical age limits in models of linguistic variation, acquisition, and change. *Crosscurrents in Second Language Acquisition and Linguistic Theories*. Amsterdam, The Netherlands: Benjamins, 225-246.
9. Eyben, F., Wöllmer, M., Schuller, B. (2010). "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", In *Proceedings of ACM Multimedia (MM)*, ACM, Florence, Italy, pp. 1459-1462.
10. Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., & Steidl, S. (2010). Emotion recognition using imperfect speech recognition. In *INTERSPEECH*, pp. 478-481.
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11(1).