Towards Assessing Communicative Competence Using Multimodal Learning Analytics

Lei Chen LCHEN@ETS.ORG
Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08540 USA

Gary Feng GFENG@ETS.ORG
Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08540 USA

Isaac Bejar IBEJAR@ETS.ORG
Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08540 USA

Abstract

Multimodal Learning Analytics (MLA) technology has received increasing attention in the learning science field. Recently, novel types of assessments using MLA to analyze the process of student activities have been advocated. In this paper, we suggest an extension of MLA to educational assessments measuring communicative competence. MLA is useful for the measurement of all aspects of the human communication process, e.g., verbal and nonverbal behaviors. Also, MLA's capability to track rich observations over a long time provides a means to conduct detailed analyses of the human communication process. To support this view, we conducted a brief but informative survey of related research from several areas. Then, we proposed a conceptual framework related to developing new assessments using MLA. One examplary case study was presented to illustrate this framework.

1. Introduction

Communicative competence (CC), in Hymes' (Hymes, 1972) original conception, concerns "knowing what to say to whom in what circumstances and how to say it." It includes not only traditional language competencies such as listening, speaking, reading, and writing, but also sociolinguistic competence and strategic competence (Canale & Swain, 1980) which are closely tied to the circumstance and the dynamics of the communicative interaction. It follows, then, that communicative competence is ideally measured in situations, using authentic communicative tasks and taking into account all aspects of communicative behaviors, verbal or non-verbal.

There remains quite some distance between the ideal and the current practices in assessing communicative competence. In this paper we argue that multimodal learning analytics (MLA) may enhance both the design and scoring of communicative competence assessment.

We take a broad view of communicative competence, one that focuses on the user of language as a social agent with a task to accomplish in a given set of circumstances and within a particular field of action, linguistic or nonlinguistic (Verhelst et al., 2009). Hence, the success of a communicative act is the overall effect of the speaker's actions, i.e., "what to say to whom in what circumstances and how to say it" (Hymes, 1972). As an assessment framework, the notion of communicative competence applies to assessment for native and non-native language learners as well as to communicative abilities in learning and the workplace. To illustrate with examples of a college student's academic life, she is required to present in front of a class. She needs to choreograph her performance across multiple modalities - e.g., the speech content, voice and intonation, facial expressions, head poses, hand gestures, and body postures. When participating in group discussions, she needs to monitor the verbal and non-verbal cues of other participants and signal her wishes for turn-taking. When participating in a job interview, she needs show strong social skills, establishing rapport with the interviewer and presenting herself confidently, through her speech register, voicing pattern, and composure. Needless to say, all of these facts will be doubly challenging when the language user is a non-native speaker and/or working with people from different cultures.

Valid and fair tests of communicative competence in these

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

situations require authentic performance-based assessment tasks and the ability to measure subtle behaviors during multimodal communicative acts. While there is no shortage of performance-based language tests, most, if not all, rely on human holistic judgment in scoring, which are susceptible to systematic biases and random fluctuations. We argue here that multimodal technology can potentially quantify communicative behaviors and improve assessment of communicative competence.

Part of the inspiration comes from the seminal work by Blikstein (Blikstein, 2013), who argued that a newly emerging technology, multimodal learning analytics (MLA), could offer new insights into students' learning trajectories. He provided a detailed description of MLA technology in (Blikstein, 2013), for example, "...in the well-established field of *multimodal interaction*, new data collection and sensing technologies are making it possible to capture massive amounts of data in all fields of human activity. These techniques include logs of computer activities, wearable cameras, wearable sensors, bio-sensors (e.g., skin conductivity, heartbeat, and EEG), gesture sensing, infrared imaging, and eye tracking". Compared to the traditional standard tests widely used in the educational testing field, new types of performance tests are closer to the tasks students encounter in their daily lives. MLA technology makes it possible to not only analyze the test outcomes (e.g., the artifacts created by students), but also to understand the process through which students use their own various skills to achieve the task goal.

MLA technology is useful not only for assessing students' hand-on activities, e.g., computer programming and the engineering construction project illustrated in (Blikstein, 2013), but also for measuring communicative competence. For example, the accurate machine perception enabled by MLA technology make it possible to measure key aspects of CC, e.g., sociolinguistic and strategic competence.

The remainder of the paper is organized as follows: Section 2 reviews the previous multi-disciplinary research on (a) communicative competence and nonverbal behaviors, (b) New assessments based on MLA technology, and (c) recent technical research in the multimodal interaction field related to scoring CC behaviors. Section 3 proposes a conceptual framework for using MLA technology to analyze and score a series of communication tasks. Section 4 reports on a case study we conducted in 2014 using MLA technology to assess oral presentation quality. Finally, Section 5 summarizes the findings of the paper and discusses future research directions.

2. Previous research

Research on the structure of CC (Canale & Swain, 1980) suggests that both sociolinguistic and strategic competence are closely related to nonverbal behaviors, e.g., facial expressions, eye gazes, hand gestures, speech prosody, and body postures. (Pennycook, 1985) described the important roles that nonverbal communication plays in human speaking skills. That paper strongly recommended that the language learning field considers adding more support to the instruction and assessment of nonverbal communication skills. There is clear evidence that examinees' nonverbal and paralinguistic behaviors can significantly affect assessment outcomes. For example, (Jenkins & Parra, 2003) investigated the impact of test takers' nonverbal behavior on oral proficiency scores. It was found that the effective use of nonverbal behavior by non-native speaking interviewees had positive effects on interviewers' perceptions of the English language ability of the test takers. Such effects also depend on cultural norms.

Recently, the researchers in the multimodal interaction field have utilized multimodal sensing to support the evaluation of the human communication process. The emergence of 3D motion tracking devices such as the Microsoft Kinect has greatly facilitated multimodal research. (Nguyen et al., 2012) proposed building a public speaking evaluation system. Students participating in a scientific presentation course were recorded using a Kinect device. Body posture and hand gestures were analyzed for identification of improper body language actions for feedback. (Batrinca et al., 2013) created a public speaking skill training system with a combination of advanced multimodal sensing and virtual human technologies. In particular, MultiSense (Scherer et al., 2013) was used to record and recognize a set of the presenters' multimodal behaviors. Meanwhile, a virtual audience would respond to the quality of the presentation in real time to provide feedback and training opportunities. Similar technologies have also been applied to coach job applicants to better prepare for interviews (Hoque & Picard, 2014; Baur et al., 2013). For example, (Hoque & Picard, 2014) developed My Automated Conversation coacH (MACH), a novel system that provides extensive access to social skills training. Using a virtual agent, MACH reads facial expressions, speech, and prosody, and responds with verbal and nonverbal behaviors in real time.

Using MLA technology to analyze and support the learning process has become popular in the multimodal interaction research field. For example, (Oviatt, 2013) analyzed the problem solving, domain expertise, and learning appearing in the Math Data Corpus (Oviatt et al., 2013), which involves multimodal data, i.e., audio and video recordings of groups of three students solving Math problems using digital pens. Based on the behavior patterns discovered by multimodal sensing, prediction techniques have been developed to identify expertise, which is critical to groups' learning outcomes. Using the same data set, (Scherer et al., 2012) investigated using low-level predictions from audio and writing modalities for the separation and identification of socially dominant learning leaders and experts within each study group.

At the ACM international conference of multimodal interaction (ICMI), a workshop/grand-challenge dedicated to MLA technology has been organized since 2012. Since 2013, the Math Data Corpus (Oviatt et al., 2013), for the first time, has been open for public use. This greatly helps the MLA research field to do more productive research on analyzing interesting learning patterns using multimodal cues. (Morency et al., 2013) summarized the research conducted in this grand challenge organized in 2013 from the viewpoints of several disciplines. In the latest MLA workshop (2014), besides running a grand-challenge on the Math Data Corpus, a new data set was added. The Oral Presentation Quality Corpus contains college students' oral presentations in their courses. These presentations were recorded using a sophisticated multimodal data collection system, including audio, video, and Microsoft Kinect depth camera.

To our knowledge, (Worsley & Blikstein, 2013) is one of the few studies using MLA to develop new performance tests. (Worsley & Blikstein, 2013) described an assessment prototype for understanding and identifying expertise as students engage in a hands-on building activity. Each student was instructed to build a tower to hold a mass by using everyday material. The process of manipulating these objects was recorded by using (a) an overhead camera for recording audio/video and (b) a Kinect device for tracking upper body motions. From the recorded multimodal data streams, a set of fine-grained action categories were coded, and the temporal sequence of these action categories were analyzed for predicting students' expertise levels. Also, directly from the Kinect tracking results, hand movements were analyzed and the motion traces were automatically segmented into several clusters. Both the action category sequence annotated from video streams and gesture analysis results inferred from Kinect motion traces were found to be useful for identifying students' expertise levels.

Three observations can be drawn from the above brief literature review. First, to better measure communicative competence, nonverbal behaviors need be tracked and included in the assessment. Second, in recent years, there has been a vibrant research community interested in improving MLA technology to better analyze the human communication process. Some research results have been applied to coaching people to improve different aspects of their communicative competence. Third, MLA has been utilized to develop a new generation of performance-based assessments that automatically track and evaluate test-takers' non-verbal behaviors. We see such technologies playing a critical role in future assessments of communicative competence.

3. Using MLA to Assess Communicative Competence

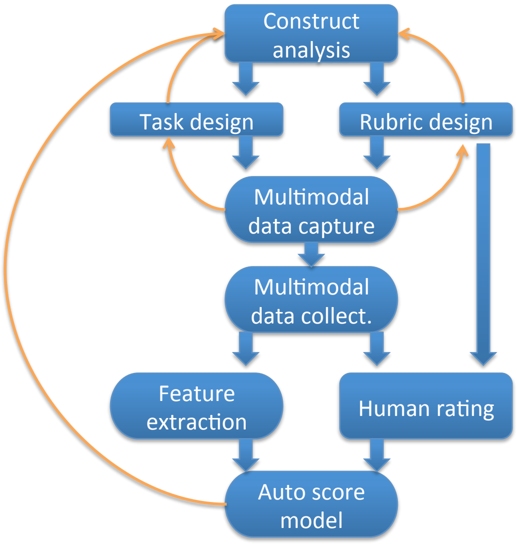

Figure 1. An diagram of the design of a multimodal assessment for communicative competence.

Taking advantage of the affordances of MLA technologies, we have been working since 2013 to develop multimodal assessments for communicative competence in various contexts. Figure 1 illustrates the typical work flow of the multimodal assessment design process. As in all assessment development, our research begins with a construct analysis of the domain, which drives both the task design and technology applications. The developed tasks will then be used in the data collection step to accumulate sufficient data samples for evaluating the assessment's psychometric properties. During this step, a proper data capture system will be determined based on considerations related to different factors, e.g., cost, convenience, and so on. Later, the collected multimodal performance samples will be processed by a set of computational algorithms to obtain features, and they will be rated by human raters for obtaining scores. Finally, using the obtained features and human-labeled scores, automatic scoring models will be constructed.

A few things are noteworthy. First, the figure depicts a generic model of development, with an ideal end goal of a fully automated scoring model that rivals human raters. Depending on the assessment goals and resources, this may vary. For high-stakes tests, the value of MLA may be to assist or validate human scoring, instead of replacing human scoring. Regardless, one of the goals of MLA-based assessment is to replicate human scores.

Second, we emphasize the mutual constraints between constructs and assessment technologies. Not all theoretically important constructs can be readily measured, and hence any particular assessment is a compromise between what we want to measure and what we can measure. Multimodal assessments are not exceptions. The feedback arrows in Figure 1 signifies the iterative nature of assessment development, in which the availability of new MLA technologies often enables us to expand the space of measurable constructs. At no time, however, should technology drive the interpretation of constructs in our design process. For example, in Figure 1, a feedback arrow exists between "task design" and "construct analysis" modules. For the public speaking assessment we will discuss in the next section, we used monologue-style presentations as the task, and we therefore would not be able to measure presenters' skills on addressing questions. Therefore, question-answering will be omitted from our construct.

Lastly, rubric design and human scoring are as critical in multimodal assessments as they are in traditional, human-scored assessments. Our research team includes strong expertise in construct analysis, human rater training, and psychometric analysis. Reliable and valid human scores are the basis for a reliable and valid automated scoring model, as illustrated in Figure 1.

We have begun to apply the general model in Figure 1 to a number of domains, such as public speaking competency, job interview skills, and teaching quality assessment. In what follows, we will illustrate our approach with one case study using MLA.

4. A Case Study

Focusing on the use of MLA technology to automatically score speakers' public-speaking skills, we have conducted research (Chen et al., 2014). To our knowledge, this is the first published work with an aim of using multimodal cues (collected using a Kinect depth camera device and a camcorder) to automatically score presenters' performance. In this section, we will briefly summarize our work published in (Chen et al., 2014) to provide a concrete example of the theoretical framework we proposed in Figure 1.

Construct analysis: We began with a construct analysis of skills involved in public speaking and a review of existing assessments and rubrics. We decided to use the construct model proposed in (Schreiber et al., 2012), in which public speaking skills are grouped to two aspects, *content* vs. *delivery*. The content aspect is related to organizing presentation content and language usage while the delivery aspect refers to verbal and nonverbal communication behaviors.

Task design: In order to provide a more comprehensive coverage on different presentation types in real life, when developing the tasks in our public speaking assessment, we created two different tasks. The first task is an informative talk for measuring a presenter's ability to convey factual ideas. The second task is an impromptu talk measuring a presenter's spontaneous responses.

Rubric design: The final rubric was based on the Public Speaking Competence Rubric (PSCR) (Schreiber et al., 2012) due to its psychometric properties. Using the PCSR as tailored to our tasks, human raters scored these presentation videos on 10 dimensions, such as vocal expression, which measures the efficiency of using vocal expression and paralinguistic cues to engage the audience, and nonverbal behavior, which measures the effect of using nonverbal behaviors to reinforce verbal messages, as well as overall holistic scores.

Data capture: We took advantage of the rapid progress on multimodal sensing devices in recent years when designing our data capture system. In particular, a Microsoft Kinect for Windows Version 1 device was used to record 3D body motions. Brekel Pro Body Kinect tracking software (v1.30 64-bit version) was used to record 48 body joints and store the motions in the Biovision hierarchical data format (BVH). Using such a consumer product level depth camera, we are able to obtain body movement observations in a very economical and relatively accurate way. In addition, to better track rich behaviors on the face and head, we also used a digital camcorder for audio/video recording. The camcorder was mounted together with the Kinect on a tripod. Both the Kinect and camcorder were placed 6ft away from the front of the speaking zone, which was marked on the ground.

Data collection: 17 speakers participated in our data collection. Each speaker was required to present two prepared talks (4 to 5 minutes long for each) based on the provided slides (displayed on a smart board behind the speaker) and two impromptu speeches (2 to 3 minutes long for each). In total, we obtained 56 presentations with complete multimodal recordings.

Human rating: Then, a human rating process was conducted according to the standard practice in educational assessment. By using a set of practices to improve human rating reliability, e.g., including expert raters, rater-training, and double-scoring, we rated all 56 presentations. The following adjudication process was used to generate final scores. If two raters agreed with each other, the score was used as the final score. Otherwise, a third rater (expert) was brought in to make the final judgment. Note that in our model-building experiments described below, we focused on the final holistic scores after the adjudication process.

Feature extraction: based on transcripts, speech, video and the motion traces recorded by Kinect, we extracted the three types of features, namely speech delivery, speech content, and non-verbal behaviors. In particular, following the feature extraction method described in (Chen et al., 2009), we used speech and transcription to generate a series of features on the multiple dimensions of speaking skills, e.g., speaking rate, prosodic variation, pausing profile, and pronunciation. The lexical features were extracted using a syntactic complexity analyzer tool (Lu, 2010) on the speech transcripts. Note that this is a crude first attempt because speech transcripts differ from written texts in important ways. With respect to visual features, we focused on the amount of locomotion, hand movements, and head orientation. The locomotion was indexed by the hip movement from the Kinect data. A basic feature set was extracted based on the mean and standard deviation (SD) of the hip and hand movement speeds and the log-transformed values. The orientation of the head is an approximation of the speaker's attention to the audience. Head poses were tracked by using GAVAM head tracker (Morency et al., 2010) from the video data. The head features consist the mean, SD, as well as mean/SD on the log scale, of the horizontal and vertical head movements.

Automatic scoring model: After obtaining the above multimodal features, we ran a standard machine learning experiment of predicting human-judged holistic scores. In particular, we ran a leave-one-out cross-validation among all subjects ($n = 17$). In each fold, presentations from 16 subjects were used to train a regression model and then the trained model was applied to the presentations from the remaining subject. Two regression approaches were utilized, including Random Forest (RF) and Support Vector Machine (SVM) using a polynomial kernel. We used the implementations provided by the R caret package (Kuhn, 2008). Hyper parameters of these machine learning models were automatically tuned by using a 5-fold cross-validation on the training set. The whole process was repeated 17 times to obtain the machine-predicted scores for all of the presentations. Table 1 reports on the correlation between the human-rated holistic scores and machine predicted ones. By using the lexical, speech, and visual features that were extracted from videos and motion traces tracked by Kinect, respectively, we find that each modality provides information for predicting overall presentation performance. Among the three modalities, the speech channel provides the most information. On the basis of the verbal model (using both lexical and speech features), adding visual features makes the multimodal model achieve the highest performance. The correlation has been increased from 0.383 to 0.416 for the RF model and from 0.377 to 0.447 for the SVM model.

Table 1. Regression models of using multimodal features to predict final holistic scores

Feature set	Random Forest (RF)	SVM
lexical	0.283	0.220
speech	0.312	0.372
visual	0.202	0.132
lexical + speech	0.383	0.377
multimodal	0.416	0.447

5. Discussion

Echoing (Blikstein, 2013), we propose that MLA may be used to expand and enhance assessments of communicative competence as originally conceptualized by (Hymes, 1972) and (Canale & Swain, 1980). We presented results from one case study (Chen et al., 2014) illustrating how MLA technologies can be used to extract features relevant to a variety of key constructs in communicative competence in presentation tasks.

It is encouraging that recent advancement of MLA makes many challenging tasks tractable. For example, the Kinect depth sensor provides quite accurate motion tracking at a very low cost and with great efficiency. Our preliminary results suggest that MLA technologies, even using some rudimentary features, e.g., speaking rate, is already useful for measuring communicative competence. More importantly, successful application of MLA technologies to automatically assess communicative competence will make running large scale and continuous observations possible.

The research is nascent, and much remains unexplored. Nevertheless, there are several lessons learned from our experience. First, compared to the capability of recording low-level communication signals, e.g., body motions, spoken words, pitch and intensity, the capability to derive meaningful higher-level cues relevant to humans' understanding is still under-developed. Second, the quantity and quality of human rating done on such performance-based assessments are still limited compared to other traditional assessment tasks. Finally, the machine learning modeling approaches utilized in these previous trials needs an overhaul in order to fully use the rich information expressed in the process of human activity.

References

Batrinca, Ligia, Stratou, Giota, Shapiro, Ari, Morency, Louis-Philippe, and Scherer, Stefan. Cicero-Towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents*, pp. 116–128, 2013.

Baur, Tobias, Damian, Ionut, Gebhard, Patrick, Porayska-Pomsta, Kaska, and Andr, Elisabeth. A job interview simulation: Social cue-based interaction with a virtual character. In *Social Computing (SocialCom), 2013 International Conference on*, pp. 220–227. IEEE, 2013.

Blikstein, Paulo. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 102–106. ACM, 2013.

Canale, Michael and Swain, Merrill. Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47, 1980.

Chen, L., Zechner, K., and Xi, X. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *NAACL-HLT*, 2009.

Chen, Lei, Feng, Gary, Joe, Jilliam, Leong, Chee Wee, Kitchen, Christopher, and Lee, Chong Min. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th international conference on multimodal interaction (ICMI)*. ACM, 2014.

Hoque, Mohammed Ehsan and Picard, Rosalind W. Rich nonverbal sensing technology for automated social skills training. *Computer*, 47(4):28–35, 2014.

Hymes, Dell. On communicative competence. *Sociolinguistics*, pp. 269–293, 1972.

Jenkins, Susan and Parra, Isabel. Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal*, 87 (1):90–107, 2003.

Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

Lu, Xiaofei. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496, 2010.

Morency, Louis-Philippe, Whitehill, Jacob, and Movellan, Javier. Monocular head pose estimation using generalized adaptive view-based appearance model. *Image and Vision Computing*, 28(5):754–761, 2010.

Morency, Louis-Philippe, Oviatt, Sharon, Scherer, Stefan, Weibel, Nadir, and Worsley, Marcelo. ICMI 2013 grand challenge workshop on multimodal learning analytics. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 373–378. ACM, 2013.

Nguyen, Anh-Tuan, Chen, Wei, and Rauterberg, Matthias. Online feedback system for public speakers. In *IEEE Symp. e-Learning, e-Management and e-Services*. Citeseer, 2012.

Oviatt, Sharon. Problem Solving, Domain Expertise and Learning: Ground-truth Performance Results for Math Data Corpus. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pp. 569–574, New York, NY, USA, 2013. ACM.

Oviatt, Sharon, Cohen, Adrienne, and Weibel, Nadir. Multimodal learning analytics: Description of math data corpus for ICMI grand challenge workshop. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 563–568. ACM, 2013.

Pennycook, Alastair. Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly*, 19(2):259–282, 1985.

Scherer, Stefan, Weibel, Nadir, Morency, Louis-Philippe, and Oviatt, Sharon. Multimodal prediction of expertise and leadership in learning groups. In *Proceedings of the 1st International Workshop on Multimodal Learning Analytics*, pp. 1. ACM, 2012.

Scherer, Stefan, Stratou, Giota, and Morency, Louis-Philippe. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 135–140. ACM, 2013.

Schreiber, Lisa M., Paul, Gregory D., and Shibley, Lisa R. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233, 2012.

Verhelst, N, Van Avermaet, Piet, Takala, S, Figueras, N, and North, B. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2009.

Worsley, Marcelo and Blikstein, Paulo. Towards the development of multimodal action based assessment. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 94–101. ACM, 2013.