Exploring the Promise of Speech Technology for Education Research

Cynthia D’Angelo, SRI International
Chad Dorsey, The Concord Consortium

This material is based upon work supported by the National Science Foundation under Grant #IIS-1550800. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Some background

• Capacity Building grant from NSF in Cyberlearning and Future Learning Technologies program:
  
  • CAP: Building Partnerships for Education and Speech Research

• How do current speech technology capabilities match up with big problems in education research?
Focus Groups in 2016

• We held a series of focus group meetings in 2016 with education researchers and speech researchers

• Gathering information about interests and experiences in this type of cross-disciplinary work

• June 2016 - education researchers

• September 2016 - speech researchers
Some questions for this webinar:

• How do we define this intersection of speech and education research?

• What are current attempts at doing work in this area?

• What are the funding avenues for this type of work?

• What should education researchers know when starting to work in this area?

• What are the realities of this kind of interdisciplinary work?

• What are the characteristics of good partners when working in this area?
Current state of speech technology
Some definitions of speech terms

• ASR - automatic speech recognition

• prosody - qualities of speech (e.g., pitch, energy)

• speaker diarization - identifying individual speakers within a group (and separating them)

• SAD - speech activity detection

• NLP - natural language processing
LENA device

- Designed for use with children—durable/light
- Up to 16 hours of audio
- Currently $399 per unit
Microphone Arrays and Beamforming

- Combining multiple microphone signals
- Reduces noise and echo
- Ambient and unobtrusive collection

The three Kaizen Rooms feature a beam forming microphone hanging from the ceiling that incorporates 24 elements to pick up comments from participants without visual intrusion.
Audio Captured

Secondary

Secondary

Primary

Primary

Secondary

Secondary

Primary

Primary

Speech

Speech

Background

Background

Background

Speech

Speech

Speech

Speech

Primary Speech

Secondary Speech

Primary Speech

Secondary Speech

Primary Speech

Secondary Speech

Primary Speech

Secondary Speech

Speech Activity Detection (SAD)

Speech Retained

Background Eliminated

Time

Time
Speaker Diarization

- Diarization algorithms help identify and separate individual speakers in a single audio track
- Error rates can be as low as 15% for naturalistic settings
Speech Activity Detection (SAD)

- Separates speech from acoustic background
- Can process large amounts of data efficiently; error rates around 10%
Speech Signal Analysis

Audio Pre-Processing
- Speech Activity Detection (SAD)
- Speaker Diarization

Speech Signal Analysis
- Word Counting
- Non-verbal event detection
- Overlap Speech Detection

Lexical Analysis
- Automatic Speech Recognition (ASR)
- Keyword Spotting (KWS)

Audio Data

Metrics
High potential from general speech signal data

- Word Counts, Turn Counts and Sharing
- Sentiment detection (uses prosody)
- Keyword-based systems
- Can detect sentiment even under high error rates
- Detection of social signals
- Laughter
- Filled vs. unfilled pauses
- Overlapped speech
- Stress detection
Participation

- Turn Taking
- Turns Initiated
- Total Turns
- Total Words Spoken

Behavior

- Overlapped Speech
- Non-verbal events
- Sentiment
- Agreeing/Disagreeing

Keywords
Lexical Analysis

Audio Data → Audio Pre-Processing
- Speech Activity Detection (SAD)
- Speaker Diarization

Speech Signal Analysis
- Word Counting
- Non-verbal event detection
- Overlap Speech Detection

Lexical Analysis
- Automatic Speech Recognition (ASR)
- Keyword Spotting (KWS)

Metrics
Fusion

• In many cases, the combination of multiple signals from “down the chain” can result in intriguing metrics

• Different low-hanging fruit may combine to create valuable and reliable signals for analysis

• Promise may lie in matching these combinations to problems of interest for education research and learning sciences
Conversation count
+ Prosody/Sentiment analysis
↓ Engaged??

Word count
+ Turn-taking
+ Keyword analysis
↓ Argumented??

Overlapped speech
+ Speaker diarization
+ Question detection
↓ Student-centered classroom culture??
Realities of working with microphones in classrooms

- Privacy issues
- Data collection difficulties (individual microphones vs array microphones, who gets one, power cords)
- Speaker registration
- Data storage difficulties (especially if you use wireless microphones)
What are some educational contexts that are low-hanging fruit for speech integration?

- Language learning
- Reading fluency
- Tutoring systems
- Collaboration and group work settings
- Assistive technologies for hard-of-hearing and visually impaired communities
- Early childhood (using speech instead of typing for input)
Current projects in this area

• Speech-Based Learning Analytics for Collaboration (SBLAC) (PI: D’Angelo)

• UT - Dallas, early childhood classroom settings

• Kristy Boyer

• others
SBLAC

• Studying collaboration of middle schoolers working on math word problems

• Groups of 3, shared iPad screen

• Each student has an individual close-talking microphone and the group has a microphone for comparison

• Also video taping for qualitative coding of collaboration indicators/behaviors to feed into machine learning on audio
Importance of Cross-Disciplinary Partnerships

• It is very important that you find a strong partner in the other discipline that you can partner with to do this work

• Speech data collection, especially in educational contexts, is very technical and difficult

• Competitive proposals will likely include research questions that are relevant to both disciplines
Funding

• Bad news: no easy/obvious program right now

• CISE is a good target, but you need a strong partnership and compelling computer science-related research questions

• SBLAC was funded through NSF EHR Core Research, which could be an option for some potential projects
Barriers

• Most work has been done on adults. Few data (corpora) of child speech exist

• Error rates still significantly high for Automated Speech Recognition of speech signals

• Youth speech development an issue

• Linguistic variation highly significant, both from youth to adults and across ages

• Naturalistic speech patterns mostly unexplored

• Multiple-speaker interaction still a frontier

• Naturalistic acoustic environments challenging
Now and near future

• The frontier is getting closer, but innovation in this space will be slower than you might expect

• Many tools have been released recently supporting various parts of this chain
  
  • Tools from Google and Facebook recently offer unexplored language/syntax promise
  
  • New open tools coming available from additional groups and efforts

• Data collection (and data quality) is still a HUGE issue and needs to be done with speech researcher partners
Next Steps

• White paper(s)

• What information do you think you need to do this kind of work? (i.e., what should be included in these white papers?)

• Best venues?

• Other outreach
Questions?

Contact us:

Cynthia D’Angelo  cynthia.dangelo@sri.com

Chad Dorsey  cdorsey@concord.org

SRI Education

The Concord Consortium