Much research has been conducted on creating empathetic responses through text, facial expressions, and gestures. Limited research on identifying acoustic-prosodic speech features—what makes a voice sound empathetic—has focused on developing a comprehensive system for empathetic speech. This can enhance human-AI interactions through creating effective empathetic agents in areas such as customer service, healthcare, and more. [2]

Our Research
- Focus on developing a comprehensive system for detecting & conveying empathy in multiple modalities.
- This can enhance human-AI interactions through creating effective empathetic agents in areas such as customer service, healthcare, and more. [2]

Goals & Hypotheses
1. Contribute to the development of a publicly available empathy speech corpus for the dialogue research community, with accurate annotations and transcript alignment.
2. Use our data to identify both the acoustic-prosodic and lexical features that distinguish empathetic, anti-empathetic, and neutral speech (e.g., change in rhythm, intonation, and cadence).
3. Develop machine learning models that can effectively detect and generate empathetic speech.

We hypothesize that taking a multi-modal approach to understanding empathy will improve both empathy classification & generation in models.

Methods
- Data Collection & Annotation
  - Collected over 300 YouTube videos in a wide range of contexts (e.g., shows, films, practice therapy sessions, etc.) consisting of almost 53 hours of audio
  -_parsed dialo&q files into TextGrid transcripts and diarized audios using Parselmouth diarization model
- Acoustic-Prosodic Analysis
  - Utilized Praat software to manually correct audio-text alignments through adjusting timestamps, realigning boundaries for utterances, and addressing speaker overlap
  - Labeled audio segments: empathetic, neutral, or anti-empathetic

Conclusions & Next Steps
Conclusions
- **Acoustic-prosodic features** indicate that empathetic voices talk in a lower pitch, in a softer tone, and at a slower pace in comparison to neutral speech.
- **Lexical features** suggest that empathetic speech is emotion-based and more complex in comparison to neutral speech.
- Lexical features alone are not sufficient in distinguishing empathetic speech.

Next Steps
- Collect and annotate more data to expand the empathy speech corpus
- Complete data pre-processing to test improvements in classification with re-aligned data within multimodal architecture
- Classify specific stage of empathy for each empathetic speech segment to help with further classification

References & Acknowledgments

We would like to thank Bridgeswater Associates, the New York State Education Department HEOP, Laura and Lloyd Blankfein and the National Science Foundation for supporting and funding our research.

Figure 1. A section of a transcript in the Praat interface after automatic alignment & manual correction. Praat allows users to view the waveform/spectrogram of the speech signal and the TextGrid transcript simultaneously.

Figure 2. Results of statistical analysis, indicating lower average pitch, intensity, jitter, and shimmer in empathetic segments.

Figure 3. Kernel density estimate plot overlaid with histogram to show difference in average speaking rate.

Figure 4. RoBERTa+openSMILE multimodal model architecture