Conveying Empathy in Multiple Modalities

Sachi Patel, Eileen Zalavarria, Umme Raisah, Barnard College
Mentors: Dr. Julia Hirschberg, Run Chen, Columbia University

Introduction

Background & Motivation

- 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– and how we can reproduce that. [1]

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 corpora** 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 (ex. 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

1. Collected over 300 YouTube videos in a wide range of contexts (ex. shows, films, practice therapy sessions, etc.) consisting of almost *53 hours of audio*
2. Parsed dialog files into TextGrid transcripts and diarized audios using ***PyAnnotate diarization model***
3. Utilized ***Praat*** software to manually correct audio-text alignments through adjusting timestamps, realigning boundaries for utterances, and addressing speaker overlap
4. Labeled audio segments: empathetic, neutral, or anti-empathetic

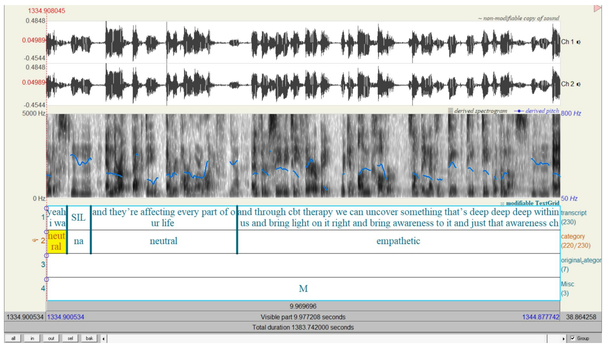

Figure 1. A section of a transcript in the Praat interface after ***automatic alignment & manual correction***. Praat allows users to view the waveform/spectrogram of the speech signal and the TextGrid transcript simultaneously.

Acoustic-Prosodic Analysis

- Utilized ***Praat*** & Python script written using ***Parselmouth*** to extract features from audio segments (intensity, pitch, speaking rate)

Lexical Analysis

- Used ***LIWC dictionary*** to identify the changes in frequencies of word categories (ex. psychological processes, personal concerns, etc.) in empathetic speech segments

Limitations & Challenges

- Inter-rater reliability due to subjective opinions on how to properly classify empathetic speech
- Variability in audio quality can affect the accuracy of acoustic-prosodic feature extraction

Results & Discussion

Acoustic-Prosodic Features

- Analyzed 400+ segments and conducted statistical analysis to compare features between empathetic and neutral segments

Acoustic-Prosodic Feature Comparison

Feature	t-score	p-value
Min Pitch	-2.896	0.004
Max Pitch	-0.222	0.008
Mean Pitch	-7.450	5.433E-12
Min Intensity	-4.802	3.371E-06
Max Intensity	-4.247	3.452E-05
Mean Intensity	-2.112	0.016
Jitter	-0.656	0.025
Shimmer	-4.643	6.783E-06
HNR	2.767	0.006

A positive t-score indicates that empathetic segments had a higher average value in the given category. ***All categories have a negative t-score except Harmonic to Noise Ratio***

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

- For empathetic speech, speaking rate is also lower when compared to neutral speech by the same speaker, as shown by the KDE plot on the right

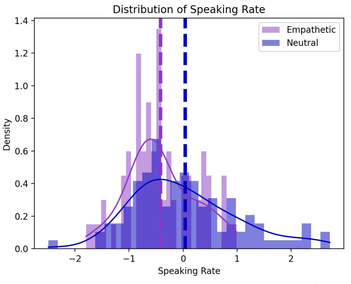

Figure 3. Kernel density estimate plot overlaid with histogram to show difference in average speaking rate.

Lexical Features

- Empathetic speakers generally express **agreement**, discuss **negative emotions**, and highlight the other person's **perception of their feelings** [1]
- Empathetic speech has **lower lexical diversity** (word variation)
- The frequency of **hedge phrases** (phrases that express hesitation or uncertainty) is very similar in both neutral and empathic speech segments
  - The most common hedges found in empathetic speech are: like, about, really, kind of, feel, think
- Empathetic speech is more complex with **lower readability scores**

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

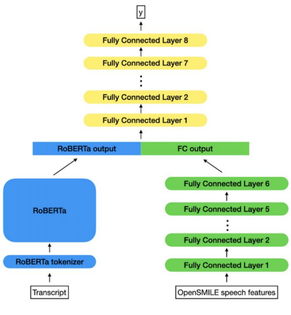

Figure 4. RoBERTa+openSMILE multimodal model architecture

References & Acknowledgments

[1] Chen, R., et al. (2024). Detecting Empathy in Speech. Proceedings of Interspeech 2024. https://www.cs.columbia.edu/speech/PaperFiles/2024/interspeech24_empathy_paper.pdf

[2] Zhi-Jiang, Y. A. N., S. U. Jin-Long, and S. U. Pan-Cha. "From human empathy to artificial empathy." *Journal of Psychological Science* 2 (2019): 299.

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