Detecting Empathy in Speech

Run Chen[1] Haozhe Chen[1] Anushka Kulkarni[1] Eleanor Lin[1]
Linda Pang[1] Divya Tadimeti[1] Jun Shin[1] Julia Hirschberg[1]

[1]Columbia University, USA

runchen@cs.columbia.edu, {hc3295,ajk2256,eml2221,sp4049,dt2760,js5810}@columbia.edu, julia@cs.columbia.edu

Summary

1. *Create a new dataset* **EmpatheticVideos**
2. *Identify interpretable acoustic-prosodic features for empathy expression - a lower, softer and slower voice*
3. *Benchmark the empathy detection task*

Introduction

Compassionate empathy – understanding another's pain as if we are having it ourselves and taking action to mitigate problems producing it – has been found useful in dialogue systems, since empathetic behavior can encourage users to like a speaker more, to believe the speaker is more intelligent, to actually take the speaker's advice, and to want to speak with the speaker longer and more often.

Related Work

- **Multimodal Avatars** produce feelings of engagement with the user through backchannels, turn-ending identification, gestures, eyebrow raising, and other facial expressions [1].
- **Textual empathetic chatbots** have been created to detect and address users' emotions and generate empathetic responses [2] - little work focused on the **speech aspect of empathy**.
- **Empathy Detection** studies show that incorporating text, audio, and speaker information are effective in predicting session-level empathy ratings [3].
- **Empathy in Different Languages** such as Italian [4] and Japanese [5], yet **no publicly available speech dataset in English** has been released.

Our goal is to identify the **acoustic-prosodic** as well as lexical aspects of speech that convey empathy — beyond merely producing appropriate emotion. In contrast to previous empathy studies where training data were often confidential, our dataset consists of **publicly available videos**.

Dataset

Data Collection

Language	English
Count	346
Length	3s to 1.5h
Category	79.2% Empathetic 17.0% Anti-empathetic 2.2% Neutral
Speakers	38.0% Female 34.4% Male 27.6% Both
Topics	Social Work, Relationship, Therapy, Interview, Parenting, Workplace
Emotions	Anger, Stress, Confusion, Frustration, Happy

Table 1. Empathetic Dataset Summary

Figure 1. Example Video of an Interview Between a Therapist and Katy Perry

Data Annotation

We diarize and annotate a subset of 65 videos for analysis. Manual re-alignment and annotation yields 1718 segments (771 empathetic and 947 neutral).

The average length of a segment is 3.01 seconds (empathetic 3.74 sec and neutral 2.43 sec).

- Audio:
  - Youtube API
  - sampling rate 16k Hz
- Diarize:
  - `pyannote` diarization model
- Re-align:
  - manually via Praat
- Annotate:
  - empathetic labels
  - 4 empathetic stages

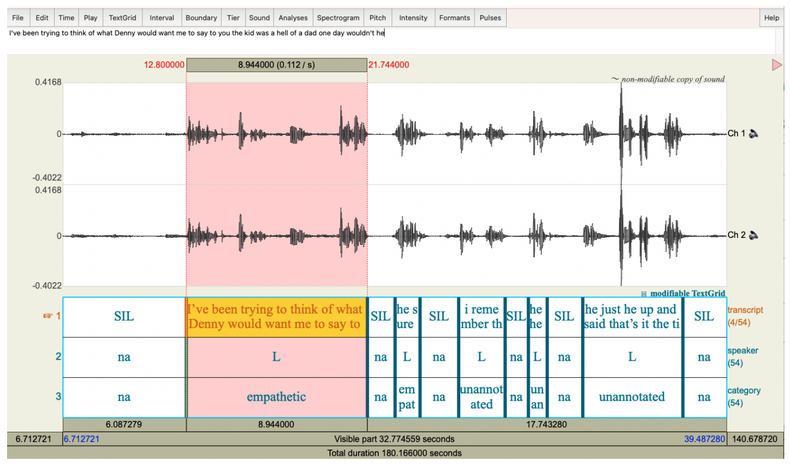

Figure 2. Manual Re-alignment and Annotation with Praat

Four Stages of Empathy

Stage	Examples
1. Establish Connection	*"Hey, we all do."*
2. Gather Information	*"When does Katherine come out in play?"*
3. Reframe & Acknowledge	*"Katherine who has a lot of hurt and unevolved feelings, I'm taking your words."*
4. Propose Solutions	*"There's a kahuna principle, it's all about where we get right energy to and our attention to ...so Katie is bigger than life but Katherine gets a little bit of time, so she can be just as evolved and happy and content."*

Table 2. Four Stages of Empathy at the Segment-Level Annotations and Examples from an Interview Between a Therapist and Katy Perry

Speech Analysis of Empathy

We extract 12 acoustic-prosodic features representing the **pitch**, **energy**, **voice quality** and **speaking rate** with praat and parselmouth tools on default parameter settings.

Feature	*t* statistics	*p*-values
min pitch	-7.476999	1.4562e-12**
max pitch	-2.222450	0.3166
mean pitch	-11.613545	5.6166e-29**
sd pitch	-3.071652	2.5952e-02**
min intensity	-4.868858	1.4707e-05**
max intensity	-5.087848	4.8222e-06**
mean intensity	-10.464473	8.3186e-24**
sd intensity	5.767524	1.1427e-07**
jitter	4.426121	1.2248e-04**
shimmer	3.379457	8.9135e-03**
hnr	0.486188	1.0
speaking rate	-3.583394	4.1835e-03**

Table 3. t-Test Statistics on Acoustic-Prosodic Features for Empathetic and Neutral Speech. ** for $p < 0.05$ after Bonferroni correction.

An empathetic voice is lower, softer and slower.

Lexical Analysis of Empathy

- Significant Linguistic Inquiry and Word Count (**LIWC**) dictionary categories:

  `assent informal anx feel tentat negemo cause`

- Slightly **Lower Lexical Diversity**, measured by averaged type to text ratio (TTR) and Measure of Textual Lexical Diversity (MTLD)
- **Lower Readability**, measured by Flesch Reading Ease scores and Dale-Chall Readability score
- Concreteness Scores and Hedging Frequencies are similar

Feature	Empathetic	Neutral
TTR	0.141	0.170
MTLD	43.04	49.37
Flesch Reading Ease Score	29.97	63.06
Dale-Chall Readability Score	8.35 (11/12th-grade)	6.98 (7/8th-grade)
Relational Hedges Freq.	6.86e-3	7.01e-3
Propositional Hedges Freq.	4.56e-3	5.09e-3
Unigram Concreteness	1.81 (± 0.68)	1.87 (± 0.72)
Bigram Concreteness	3.18 (± 0.79)	3.11 (± 0.96)

Table 4. Lexical Features for Empathetic and Neutral Segments. Concreteness Scores are in mean ± standard deviation.

Empathy Classification and Results

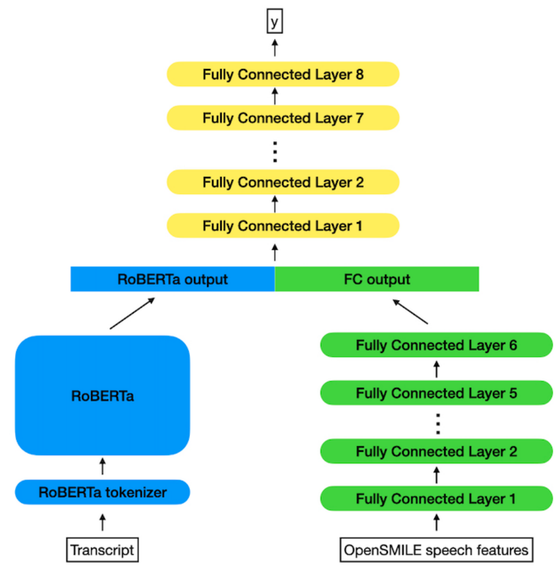

Figure 3. RoBERTa+openSMILE Multimodal Model Architecture. Each fully connected layer is followed by a ReLU activation and 0.1 dropout, except the last fully connected layer 8.

- Data Balanacing: downsample neutral data to 771 segments, the same as empathetic
- Training and validation splits: 80/20
- **Baseline RoBERTa**: `roberta-base` RobertaForSequenceClassification model, finetuned with lr= 2e-5, batch_size=16, epochs = 20
- **RoBERTa+openSMILE**: AdamW optimizer (lr=2e-5, eps = 1e-8), batch size = 8, epochs = 10

Model	Val. Acc	F1 score
RoBERTa	0.528	0.603
RoBERTa + openSMILE	0.781	0.840
RandomForest	0.540	0.587

Table 5. Model Performance on the Empathetic/Neutral Binary Classification Task. Accuracy and F1 score on the held-out validation set.

Conclusions and Future Work

- We have collected a new publicly available empathy corpus of English empathetic videos
- Empathetic voices tend to be lower, softer and slower, compared to neutral speech; and empathetic texts are emotion-based, less diverse and slightly less readable
- The classification results underlines the importance of speech in conveying empathy beyond the text
- We have been collecting and annotating additional empathy data in Mandarin
- Empathy as a positive communication change

References

[1] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," *Intelligent Virtual Agents*, pp. 68–79, 2011.

[2] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 5370–5381, Association for Computational Linguistics, July 2019.

[3] T. Tran, Y. Yin, L. Tavabi, J. Delacruz, B. Borsari, J. D. Woolley, S. Scherer, and M. Soleymani, "Multimodal analysis and assessment of therapist empathy in motivational interviews," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, (New York, NY, USA), p. 406–415, Association for Computing Machinery, 2023.

[4] F. Alam, M. Danieli, and G. Riccardi, "Annotating and modeling empathy in spoken conversations," *Comput. Speech Lang.*, vol. 50, p. 40–61, jul 2018.

[5] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent," in *Proc. Interspeech 2022*, pp. 5155–5159, 2022.