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@columbia.edu, ajk2256@barnard.edu, eml2221@columbia.edu, sp4049@columbia.edu, dt2760@columbia.edu, js5810@columbia.edu, julia@cs.columbia.edu

Abstract

Empathy is the ability to understand another's feelings as if we were having those feelings ourselves. It has been shown to increase to people's trust and likability. Much research has been done on creating empathetic responses in text in conversational systems, yet little work has been done to identify the acoustic-prosodic speech features that can create an empathetic-sounding voice. Our contributions include 1) collection of a new empathy speech dataset, 2) identifying interpretable acoustic-prosodic features that contribute to empathy expression and 3) benchmarking the empathy detection task.

Index Terms: empathy, computational paralinguistics, speech processing

1. Introduction

Much research has been done in the past 15 years on creating empathetic responses in text, facial expressions and gestures in conversational systems. However, little has been done to identify the speech features that can create an empathetic *sounding* voice. *Empathy* is the ability to understand another's feelings as if we were having those ourselves [1]. It can take several forms: *cognitive empathy* or "perspective-taking", being able to put yourself in another's place – a particularly useful skill for managers; *emotional empathy*, actually feeling another person's emotion, also called "emotional contagion", which can be overwhelming; and *compassionate empathy* – understanding another's pain as if we are having it ourselves and taking action to mitigate problems producing it [2, 3]. This third category has been found especially useful in dialogue systems and robots, 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. Compassionate empathy can also be used to improve success in health-care advice-giving, as well as in negotiations and conflict resolution. Even when humans know that they are dealing with a computer system, if that system behaves empathetically, users will still like and trust it more [4].

Producing empathetic responses requires first identifying a user's emotions to understand the need for such responses as well as the type of emotion the user is expressing and the reason for that emotion. Much research has been done to recognize the user's emotion and its cause from the user's words and sometimes from their speech. Much has also been done to create the appropriate emotional content of the system response — in some research projects also to provide appropriate facial expressions and gestures. But very little work has been done to discover what vocal cues can be used to create an empathetic-sounding voice. For empathy is more than simple agent emotional responses: to encourage users to connect with a conversational agent, that agent must present itself as empathetic even before the user expresses a need.

Our goal is to identify the acoustic-prosodic as well as lexical aspects of speech that convey empathy — beyond merely producing appropriate emotion to address a user's particular issue or entraining to the user. We collect a new dataset of empathetic videos. We compare empathetic speech segments with neutral ones for changes in pitch, intensity, voice quality and speaking rate. We report empathetic lexical categories, specificity and readability levels. We also study how the speech features interact with the lexical content through ML modeling.

2. Related work

Previous work focused on developing multimodal avatars to produce feelings of engagement with the user, using different forms of listening behavior: backchannels, turn-ending identification, gestures, eyebrow raising, and other facial expressions. These include [5]'s Rea, a conversational agent; [6] and [7]'s Virtual Laboratory Exercise Agents, created to improve daily exercise interactions; [8] and [9]'s Rapport Agents with human-like listening behavior including backchannels, turn-ending identification, smiles and nods; [10]'s Greta, used to evaluate different methods of combining emotional facial expressions to produce empathy; [11]'s Jade Semantics Agents which added more empathetic emotions beyond happy and sad in email messages to the mix; [12]'s development of more rapport-building strategies using non-verbal behavior.

While text-based empathetic chatbots have been created to detect and address users' negative emotions [13] and generate empathetic responses [14], little work has been done focusing on the speech aspect of empathy. Multimodal approaches incorporating text, audio, and speaker information have proven effective in predicting session-level empathy ratings [15, 16]. For turn-level empathy, [17] discovered that both pitch and intensity (loudness) were lower for both male and female speakers in empathetic speech than in neutral speech on their collected corpus of empathy and emotion labels on Italian call center conversations. More recent studies have investigated empathy in Cantonese [18], and Japanese [19], yet no publicly available speech dataset in *English* has been released.

In this work, we aim to identify empathetic speech using both acoustic-prosodic and lexical features of English YouTube video data. We have collected a large number of these and annotated segments in them as empathetic, neutral, or anti-empathetic in what is said and how it is said, as well as many other features of the videos to identify which are most watched and liked as well as different *stages* of empathetic speech. In contrast to previous empathy studies where training data were confidential, our dataset is sourced from publicly available video platform and will be made accessible for future research.

3. Dataset

3.1. Data Collection

We have collected an empathetic dataset consisting of 346 English videos and about 53 hours in total.[1] The key dataset statistics are summarized in Table 1. These were manually collected from Youtube through keyword searches, such as “empathy” and “empathetic training” from 2020 to 2022. They include empathy training videos, acted therapy sessions, TV shows, movies, interviews and TED Talks. The videos comprise 38% spontaneous and 62% acted speech. We identify metadata from video platform APIs, including video and channel information and viewer likes and comments. We also annotate additional information such as video category, speaker number and gender, language, intended audience, and emotions expressed. Each video is rated by at least three expert annotators as “empathetic”, “neutral” or ”anti-empathetic” and taken the majority vote.

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

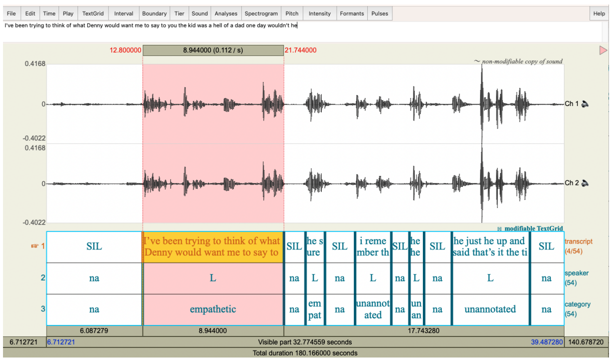
Table 1: *Empathetic Dataset Summary*

3.2. Data Annotations

To gain better understanding of the empathetic speech, we select a subset of 65 videos for diarization and annotation for further analysis. Using the audio we obtain from the Youtube API, we first transcribe and diarize using `pyannote`[2] diarization model. However, as the quality of the transcripts and alignments require further manual correction, we re-align the transcripts using the Praat[3] interface shown in Figure 1. These videos were annotated between 2023 and 2024 by 10 annotators and verified by least one different annotator.

The manual re-alignment and annotation resulted in 1718 segments with time stamps, transcripts, speakers, empathetic labels (“empathetic” or“neutral”) and empathetic stages. We define a segment as a natural sentence uttered by a speaker, potentially shorter than a speaker turn. Each segment was sampled at 16k Hz, and we excluded any with music or noisy backgrounds to ensure audio quality.

We also annotate four stages of empathetic behavior, a simplification of empathy practices in therapy such as therapeutic empathy system of empathetic attunement, attitude/stance, communication, and technical/conceptual knowledge [20, 21]



I've been trying to think of what Denny would want me to say to you the kid was a hell of a dad one day wouldn't he

Figure 1: *Manual Re-alignment And Annotation with Praat*

The four stages are defined below with examples shown in Table 2.

Stage 1: Make the other person feel comfortable. This stage intends to establish connections between speakers and a sense of resonance.

Stage 2: Asking questions. This stage is intended to gain information about the other’s personal situation, corresponding to the part of “feeling someone’s pain.”

Stage 3: Reframing and acknowledging the other person’s experience or situation. This stage often entails repeating or paraphrasing what the other has said. This stage intends to make the other feel heard (like stage 1 but more specific to personal situation).

Stage 4: Proposing solutions. This stage provides new information to the other: some problem solution or new insights which can help the other.

Stage	Examples
1	”Hey, we all do.”
2	”When does Katherine come out in play?”
3	”Katherine who has a lot of hurt and unevolved feelings, I’m taking your words.”
4	”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: *Examples of four stages of empathy at the segment-level annotations from an interview between a therapist and Katy Perry.*

Segment-level annotation yields 771 empathetic and 947 neutral segments. The average length of a segment is 3.01 seconds (empathetic 3.74 sec and neutral 2.43 sec). We use these manually annotated segments for developing empathy classification models.

4. Empathy Analysis

To investigate the role of text and speech in conveying empathy, we employ significance tests on interpretable lexical and speech features. Specifically, we conduct unpaired t-tests on these features extracted from 771 empathetic segments and 947 neutral

[1] Code and data will be released at: `https://github.com/run-chen-nlp/empathy`

[2] https://github.com/pyannote/pyannote-audio

[3] https://www.fon.hum.uva.nl/praat/

segments to identify the features that exhibit significant differences in the empathy segments.

4.1. Speech Analysis

We extract a set of 12 acoustic-prosodic features representing the pitch, energy, voice quality and speaking rate of speakers: pitch mean, minimum, maximum, and standard deviation, intensity mean, minimum, maximum, and standard deviation, jitter, shimmer, harmonics-to-noise ratio (HNR), and speaking rate. These features are extracted with praat [22] and parselmouth [23] tools on default parameter settings. The speaking rate is measured in words per second from human-annotated transcripts. Additionally, we obtain 384 low-level features identified using the Interspeech 2009 (IS09) ComParE Challenge OpenSMILE baseline feature set, a standard benchmark feature set for many computational paralinguistic tasks [24, 25]. The OpenSMILE feature size is comparable to RoBERTa textual embeddings dimensions, preventing the model from ignoring speech information in the training.

We run independent t-tests for speech features extracted from the empathetic segments against those from the neutral segments. We apply Bonferroni correction to the p-values to control for errors in multiple testings. In Table 3, most acoustic-prosodic features are significantly different.

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.*

The empathetic speech is significantly lower in pitch minimum, mean and standard deviation, consistent with our expectation of a typical lower, flatter “therapist tone”. The empathetic speech also has significantly lower minimum, mean and maximum intensities but higher standard deviations in intensities. This corresponds to a quieter, softer but more varied speech. Higher jitter and shimmer are usually associated with the breathiness of a calming voice. A lower speaking rate can also help to convey the empathetic message that one hears and understands the other. These results all align with our expectation that a comforting and soothing empathetic speaker typically features a lower, softer and slower voice.

We train a random forest (RF) classifier[4] using the 12 acoustic-prosodic features, to distinguish between empathetic and neutral speech segments. The empathetic segments are downsampled create a balanced set. With an 80/20 train/test split, the RF model achieves 0.540 accuracy and 0.587 F1 score. After model fitting, the Gini importance for the classifier identifies pitch mean and intensity standard deviation as the most crucial features (both about 0.11), although overall the normalized importance scores distribute approximately uniformly. These findings are consistent with our earlier t-test results, which highlighted pitch and intensity as the most significant indicator of empathy.

4.2. Lexical Analysis

We further investigate lexical features associated with empathy, including significant LIWC dictionary categories [26], lexical diversity [27], concreteness scores [28], hedging frequencies [29, 30] and readability scores [31, 32]. Although we are able to identify a few lexical features that are characteristic of our empathetic dataset, the textual content itself alone may not be sufficient for us to understand or convey empathy, in contrast to the speech analysis in the previous section 4.1.

The LIWC dictionary categories [26] range from linguistic dimensions, psychological processes, personal concerns to spoken categories. In our analysis, we pinpoint specific LIWC lexical categories that exhibit notable frequency changes in empathetic speech, including `assent`, `informal`, `anx`, `feel`, `tentat`, `negemo`, `cause`. This suggests that when expressing empathy, individuals tend to express agreement, speak informally, emphasize the perceptual process of feeling, utilize vocabulary associated with tentative and causation cognitive processes, and discuss negative emotions like anxiety more frequently. These linguistic choices align with the empathetic goal of understanding and connecting with the other person’s feelings.

The empathetic text has slightly lower lexical diversity. The averaged type to text ratio for empathetic and neutral segments are 0.141 and 0.170, respectively. [27]’s Measure of Textual Lexical Diversity (MTLD) for empathetic and neutral segments are 43.04 and 49.37, respectively. This could be attributed to the fact that empathy is typically manifested through the process of generalization and abstraction from the specific circumstances that give rise to the emotions of the other speaker, a phenomenon often observed in Stage 3 of empathetic responses.

The hedge phrase frequencies are very similar between empathetic and neutral speech, though empathetic segments have slightly lower frequencies. *Relational hedges* [29], which distance the speaker’s relation to the propositional content, occur with a frequency of 0.00686 in empathetic speech and 0.00701 in neutral speech. The most common relational hedges for empathy include words in the LIWC cognitive processes category, such as “know”, “feel” and “think”. *Propositional hedges*, which introduce uncertainty into the propositional content itself, appear at a rate of 0.00456 in empathetic speech and 0.00509 in neutral speech. The most common propositional hedges for empathy are “like ”, “about”, “really” and “kind of”. We speculate that empathetic speakers may employ clearer and less ambiguous language when presenting their advice to their interlocutors, a strategy we observe in Stage 4.

Similarly, the concreteness scores [28] are comparable for empathetic and neutral speech. The empathetic segments averaged unigram score is 1.81 (std 0.68) and bigram score 3.18 (std 0.79), whereas the neutral segments averaged unigram score is 1.87 (std 0.72) and bigram score 3.11 (std 0.96). However, as the difference is minimal, such similarity in concreteness, as well as the frequency of hedge words between empathetic and neutral speech highlights the crucial role of acoustic-prosodic cues in conveying empathy.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Lower readability scores indicate the complexity of empathetic speech. The Flesch Reading Ease scores [31] for empathetic and neutral transcripts are 29.97 and 63.06, respectively, indicating that empathetic speech is significantly more challenging to read and comprehend. The Dale-Chall Readability score [32] for empathetic and segments are 8.35, which corresponds to a text level understandable by 11th or 12th-grade students. In contrast, neutral segments have a higher score of 6.98, matching a 7th or 8th-grade student level. This suggests that empathy utterances are more difficult to understand, as empathetic speakers often demonstrate their understanding by deepening or adding complexity to their interlocutors' experience, a strategy we observe in Stage 3.

5. Empathy Classification and Results

To assess the impact of speech cues, we conduct an ablation study by comparing model performance with and without textual and speech information. We fine-tune a pretrained `roberta-base` model on our dataset [33]. Addressing the class imbalance between empathetic and neutral, we downsample empathetic data, resulting in a balanced dataset of equal number of empathetic and neutral segments. We then divide the data into training and validation sets with a 80/20 split ratio with `StratifiedGroupKFold(n_splits=5)` [5].

Baseline "RoBERTa": The baseline textual model is a pre-trained `roberta-base` RobertaForSequenceClassification model, with tokenized transcripts as input, finetuned for binary classification with lr= 2e-5, batch_size=16, for 20 epochs.

"RoBERTa+openSMILE": The multimodal model combines signals from both text and speech (Figure 2). Each segment transcript is encoded with a pretrained `roberta-base` encoder and passed through a pretrained `roberta-base` model with frozen parameters. The 384 dimensional IS09 openSMILE feature vector representing the speech signal goes through 6 fully connected layers, each followed by a ReLU activation and 0.1 dropout. Then outputs from both text and speech are concatenated and fed into 8 fully connected layers, each followed by a ReLU activation and 0.1 dropout, except the last output layer. The model is trained with AdamW optimizer (lr=2e-5, eps = 1e-8), batch size = 8, epochs = 10.

All models are trained on Tesla T4 GPU.

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

Table 4: *Model performance on the empathetic/neutral binary classification task. Accuracy and F1 score on the held-out validation set.*

The classification results are summarized in Table 4. The RoBERTa text-only model achieves 0.528 accuracy and 0.603 F1 score, with empathy class accuracy of 0.545 and neutral class accuracy of 0.496. The RoBERTa + openSMILE model performance peaked at epoch 3 with accuracy 0.781 and 0.840 F1 score, among with the empathy class accuracy 0.881 and neutral 0.591. The RandomForest results are copied from Section 4.1 for comparison.

We observe a huge performance improvement in a model utlizing both text and speech. This experiment demonstrates that speech features play a valuable role in enhancing the model's ability to predict empathy. It underscores that text alone may not be sufficient to convey empathy, emphasizing the need of integrating acoustic-prosodic information into conversational agents, as misalignment between text and speech expression often leads to ineffective or even sarcastic responses.

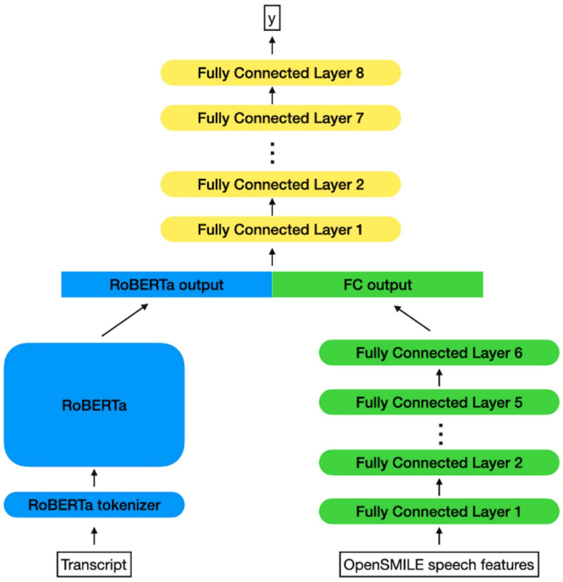


Figure 2: *RoBERTa+openSMILE multimodal model architecture. Each fully connected layer is followed by a ReLU activation and 0.1 dropout, except the last fully connected layer 8.*

6. Conclusions

We have collected a new empathy corpus of English empathetic videos. Our analysis on this dataset reveals distinctive characteristics of empathetic voices and texts. Empathetic voices tend to be lower, softer and slower, compared to neutral speech; and empathetic texts are emotion-based, less diverse and slightly more complex. These results are useful in guiding the development of empathetic conversational agents. We benchmark the empathy classification task with the RoBERTa model. The classification results underlines the importance of speech in conveying empathy beyond the text. As we are releasing the dataset to the public, the research community can use our collected data to train their own models for tasks such as empathy detection and empathetic text-to-speech synthesis.

In the future, we plan to identify acoustic-prosodic and lexical features associated with different stages of empathy for more fine-grained analysis that could enhance training empathetic chatbots as well as therapists in their practice. We also plan to incorporate other modalities which are currently not utilized in our models to investigate how facial expressions and gestures in the videos cooperate with speech to convey empathy.

Furthermore, we have been collecting and annotating additional empathy data in Mandarin, with the aim of conducting similar analyses as we have with our English dataset. As one may speculate that empathy expression may vary with different language and cultures, this expansion will enable us to explore cross-linguistic and cross-cultural dimensions of empathy expression.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html

7. Acknowledgements

This research is being developed with funding from the Columbia Center of AI Technology (CAIT) research award and the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

8. References

[1] R. F. Baumeister and K. D. Vohs, *Encyclopedia of social psychology*. Sage, 2007, vol. 1.

[2] M. L. Healey and M. Grossman, "Cognitive and affective perspective-taking: evidence for shared and dissociable anatomical substrates," *Frontiers in neurology*, vol. 9, p. 491, 2018.

[3] J. L. Goetz, D. Keltner, and E. Simon-Thomas, "Compassion: an evolutionary analysis and empirical review." *Psychological bulletin*, vol. 136, no. 3, p. 351, 2010.

[4] G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano, "Getting to know each other: The role of social dialogue in recovery from errors in social robots," in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018, pp. 344–351.

[5] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interfaces," *AI Mag.*, vol. 22, no. 4, p. 67–83, oct 2001.

[6] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Trans. Comput.-Hum. Interact.*, vol. 12, no. 2, p. 293–327, jun 2005. [Online]. Available: https://doi.org/10.1145/1067860.1067867

[7] T. Bickmore, D. Schulman, and L. Yin, "Maintaining engagement in long-term interventions with relational agents," *Applied artificial intelligence : AAI*, vol. 24, pp. 648–666, 07 2010.

[8] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *Intelligent Virtual Agents*, C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 125–138.

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

[10] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: An interactive expressive eca system," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, ser. AAMAS '09. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2009, p. 1399–1400.

[11] M. Ochs, C. Pelachaud, and D. Sadek, "An empathic virtual dialog agent to improve human-machine interaction," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, ser. AAMAS '08. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2008, p. 89–96.

[12] R. Zhao, T. Sinha, A. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in *Intelligent Virtual Agents*, 2016.

[13] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: An emotion-aware wellbeing chatbot," in *International Conference on Affective Computing and Intelligent Interaction*, September 2019. [Online]. Available: https://www.microsoft.com/en-us/research/publication/emma-an-emotion-aware-wellbeing-chatbot/

[14] 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: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381. [Online]. Available: https://aclanthology.org/P19-1534

[15] J. Gibson, D. Can, B. Xiao, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan, "A Deep Learning Approach to Modeling Empathy in Addiction Counseling," in *Proc. Interspeech 2016*, 2016, pp. 1447–1451.

[16] 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*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 406–415. [Online]. Available: https://doi.org/10.1145/3577190.3614105

[17] F. Alam, M. Danieli, and G. Riccardi, "Annotating and modeling empathy in spoken conversations," *Comput. Speech Lang.*, vol. 50, no. C, p. 40–61, jul 2018. [Online]. Available: https://doi.org/10.1016/j.csl.2017.12.003

[18] D. Tao, T. Lee, H. Chui, and S. Luk, "Characterizing Therapist's Speaking Style in Relation to Empathy in Psychotherapy," in *Proc. Interspeech 2022*, 2022, pp. 2003–2007.

[19] 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*, 2022, pp. 5155–5159.

[20] M. Fuller, E. Kamans, M. van Vuuren, M. Wolfensberger, and M. D. de Jong, "Conceptualizing empathy competence: a professional communication perspective," *Journal of business and technical communication*, vol. 35, no. 3, pp. 333–368, 2021.

[21] B. D. Jani, D. N. Blane, and S. W. Mercer, "The role of empathy in therapy and the physician-patient relationship," *Complementary Medicine Research*, vol. 19, no. 5, pp. 252–257, 2012.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.2.14, retrieved 24 May 2022 http://www.praat.org/, 1992-2022.

[23] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[24] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech 2009*, 2009.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[26] J. W. Pennebaker, R. Booth, R. L. Boyd, and M. Francis, *Linguistic Inquiry and Word Count: LIWC2015*, Austin, TX, Sep. 2015.

[27] P. M. McCarthy and S. Jarvis, "Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment," *Behavior research methods*, vol. 42, no. 2, pp. 381–392, 2010.

[28] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior research methods*, vol. 46, pp. 904–911, 2014.

[29] A. Prokofieva and J. Hirschberg, "Hedging and speaker commitment," in *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*, 2014.

[30] E. F. Prince, J. Frader, C. Bosk *et al.*, "On hedging in physician-physician discourse," *Linguistics and the Professions*, vol. 8, no. 1, pp. 83–97, 1982.

[31] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.

[32] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.