Modeling Empathy and Human Conversation Dynamics

Run Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2026

Abstract

Modeling Empathy and Human Conversation Dynamics

Run Chen

Empathetic communication is essential for supporting human well-being, improving social connectedness, and enabling effective dialogue in emotionally charged situations. As conversational AI systems become embedded in daily life, the ability to recognize, interpret, and respond to human emotion responsibly has become a central goal for the field. Yet, existing approaches often rely on text-only inputs and simplistic approximations of supportiveness, making them insufficient for modeling the nuance and complexity of real human empathy. This dissertation advances a holistic framework for empathetic conversational AI that integrates structural conversational dynamics, multimodal signals from speech, scalable learning with large language models, and proactive safeguards against coercive interaction.

The first component of this work strengthens the foundations of conversational understanding. We address long-standing transcription and timing discrepancies in the widely used Switchboard Dialog Act corpus by producing the Re-Aligned Switchboard Dialog Act (RASwDA) dataset, improving alignment between lexical and acoustic-prosodic cues and enabling more accurate dialog act classification. We also develop new quantitative analyses of prosodic and lexical entrainment in task-oriented settings, revealing how behavioral coordination serves as a precursor to rapport and mutual understanding.

Building on these foundations, the second component investigates empathy as a more specialized and emotionally collaborative form of conversation. We present EMPSPEECH, a speech-based empathy corpus annotated at the segment level to capture acoustic-prosodic markers of supportive behavior—representing one of the first resources in computational empathy centered on spoken communication in English. We further examine empathy's inherent complexity through multimodal

learning and model disagreement analysis, demonstrating how conflicting cues across modalities expose the ambiguity and interpretive demands present in real interactions.

To scale empathetic response generation in the age of large language models, we introduce SYNTHEMPATHY, a psychotherapy-informed corpus of 106k empathetic dialogues produced without crowdsourcing. Using Chain-of-Empathy prompting, we show improvements in emotional alignment and contextual appropriateness while enabling customizable and domain-adaptable empathetic behavior. Together, these contributions demonstrate that effective empathy modeling requires recognition, interpretation, and expressive alignment across modalities and conversational context.

The final component addresses a critical ethical challenge: ensuring that emotionally aware AI remains trustworthy and protective of users. We develop speech-based detection of mental manipulation using multi-speaker synthetic audio and few-shot learning, illustrating how systems can identify harmful persuasive strategies and prevent coercive influence in vulnerable moments.

Overall, this dissertation presents a unified path toward empathetic and responsible conversational AI: from foundational interaction modeling, to scalable development of emotional intelligence, to proactive safety measures that preserve user autonomy. By advancing empathetic and conversational AI research resources, multimodal modeling, robust analysis techniques, and ethical frameworks, this work contributes essential scientific and practical foundations for conversational technologies that are not only more capable and adaptive, but genuinely supportive of human emotional well-being.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first like to express my deepest gratitude to my advisor, Prof. Julia Hirschberg. Thank you for believing in me from day one, for your invaluable guidance, and for pushing me to think deeper and write clearer. Your curiosity, creativity, and kindness have shaped not only my research but also who I am as a scientist. I am especially grateful for your unwavering support through challenging moments both academically and personally.

I am profoundly appreciative of my thesis committee members — Prof. Smaranda Muresan, Prof. Kathleen McKeown, Dr. Dilek Hakkani-Tür, and Prof. Colin Leach — for generously offering your expertise, encouragement, and insightful feedback that strengthened this work. It has been an honor to learn from each of you.

I want to thank my parents and family for their love and encouragement throughout my entire journey. You have always believed in me wholeheartedly, even when I doubted myself.

To my partner, Bright, thank you for being by my side from those uncertain early days of COVID-19 to the finish line of this PhD. Your unwavering support made this journey brighter and more joyful.

I am also grateful for the friendship and camaraderie of my labmates in the Speech Lab — Lin, Sara, Debasmita, Yu-Wen, Siyan, Brenda, and Rose — who have made the lab feel like home.

To the students I have had the privilege to mentor and collaborate with — thank you for your hard work, creativity, and trust. You have taught me as much as I have taught you, and I am incredibly proud of everything we accomplished together.

Finally, to my friends near and far — thank you for keeping me grounded, for reminding me of life outside the lab, and for celebrating each milestone along the way.

Chapter 1: Introduction and Background

Human communication is fundamentally interactive and social. In everyday conversation, people coordinate meaning, build rapport, and provide emotional support through a combination of words, tone, timing, and shared context. Empathy plays a central role in these processes: it shapes how individuals perceive others' emotions, establish trust, and respond supportively across social, clinical, and professional settings [1, 2, 3, 4]. As conversational AI systems become increasingly embedded in daily life — from mental health support and education to customer service and socially aware agents [5, 6] — the ability to understand and respond to emotion safely and effectively has become both a scientific challenge and a societal necessity.

1.1 Computational Empathy in NLP and Speech

Computational empathy aims to model the perception and expression of empathy from text, speech, and multimodal inputs. Early work explored empathy in counseling and supportive dialogue [1, 6], active listening [3], healthcare [4], and socially aware conversational agents [5, 2]. Shared tasks such as the WASSA series on Empathy, Emotion, and Personality have further catalyzed research by providing benchmark datasets and evaluation protocols for empathy prediction [7, 8, 9, 10]. These efforts helped establish empathy as a distinct target for modeling, beyond generic sentiment or affect detection.

Several surveys have taken stock of this emerging area. Early work on artificial empathy focused on how agents can simulate or elicit empathic responses in interaction with humans [11]. Later, [12] provided a critical reflection on empathy in NLP, highlighting that many text-based studies either left empathy undefined or conflated it with loosely related affective constructs, leading to low construct validity and limited reproducibility. [13] surveyed techniques, datasets, and metrics up to 2022,

calling for clearer definitions, larger datasets, integration of non-verbal cues, and more principled evaluation. More recently, [14] presented the first broader survey spanning textual, auditory, visual, and physiological signals for empathy detection, arguing for multimodal approaches but also noting the scarcity of labeled multimodal empathy data.

In parallel, the rapid transition toward large language models (LLMs) and large audio-language models (LALMs) has reshaped the landscape of empathetic NLP. Recent work has introduced large empathy datasets and scalable synthetic empathy generation using LLM prompting [15, 16], multimodal benchmarks leveraging speech and vision, and LLM-based empathetic response generation and evaluation [17]. In speech and audio research, studies have expanded empathy modeling using acoustic and prosodic features [18], expressive text-to-speech (TTS) [19], and multimodal integration, reflecting a growing recognition that voice carries critical cues for emotional attunement.

1.2 Limitations of Existing Approaches

Despite this progress, several limitations persist. First, much of the prior work on computational empathy remains predominantly *text-focused*. Many systems are trained and evaluated on datasets such as EMPATHETICDIALOGUES [5] and related text-only corpora, which lack prosodic or paralinguistic cues and often treat empathy as a single umbrella label. As [12] argue, empathy is frequently approximated through emotion labels or generic "appropriate response" judgments, rather than grounded in psychological accounts of cognitive, emotional, and compassionate components. A large portion of work effectively uses emotion as a proxy for empathy, focusing on matching or transforming emotional labels [20, 21, 22], which can obscure the distinction between emotional similarity and genuinely empathic concern.

Second, while surveys such as [14] highlight the importance of multimodal signals, available multimodal and speech-based empathy datasets are still relatively scarce, domain-specific, and often not publicly accessible. Many corpora in the spoken domain originate from therapeutic, educational, or service-oriented settings and are constrained by privacy or licensing [23, 24, 25, 26, 27, 28, 29,

30, 31]. This has led to a field where empathy is largely studied in text, even though speech provides rich acoustic and prosodic correlates that are central to human empathic communication [18, 32]. Cross-lingual and cross-cultural work also remains limited, with relatively few explicit comparisons across languages or code-switched settings, despite evidence that empathy is shaped by cultural norms and interaction styles.

Third, evaluation remains a major challenge. Existing benchmarks tend to rely on single-turn ratings, emotion classification accuracy, or generic dialogue quality metrics. Recent analyses of LLM-based empathetic dialogue [33, 34, 35] show that large models often rely on stock empathetic phrases, can be highly sensitive to prompts, and may appear empathetic while failing to offer substantive or safe support. Sycophantic behavior — where models over-accommodate or uncritically agree with users — is particularly concerning in mental health or high-stakes contexts [36, 37].

Finally, the ethical implications of emotionally capable AI have only recently begun to receive systematic attention. While some work examines how empathetic responses influence user trust, engagement, and perceived support [38, 39], relatively little research has focused on detecting and mitigating manipulative or coercive behaviors in conversational systems, especially in speech-based settings. This creates a gap between systems' growing expressive power and the mechanisms needed to ensure their responsible deployment.

1.3 Thesis Goals and Contributions

This dissertation addresses these limitations by advancing empathetic conversational AI along three interconnected directions:

- **Conversational Dynamics.** We strengthen foundational interaction modeling by improving how systems represent discourse structure and coordination. This includes creating the Re-Aligned Switchboard Dialog Act corpus (RASwDA) [40], which corrects long-standing alignment issues and enables accurate use of acoustic-prosodic cues for dialog act modeling, and new analyses of lexical and prosodic entrainment in task-oriented speech [41].

- **Empathetic Modeling and Generation in Speech and Multimodal Settings.** We investigate empathy as a specialized, emotionally collaborative form of conversation, focusing on speech as a primary but underexplored modality. We introduce EMPSPEECH [18], a segment-level speech-based empathy corpus, and study multimodal model disagreement to expose the inherent ambiguity of empathy across modalities [42]. We further present SYNTHEMPATHY [15], a psychotherapy-informed synthetic corpus of 106k empathetic dialogues generated without crowdsourcing, demonstrating how empathetic capabilities can be scaled in the age of LLMs.

- **Ethical Safeguards and Mental Manipulation.** We develop methods for detecting mental manipulation in speech using multi-speaker synthetic audio and few-shot modeling, showing how emotionally capable systems can identify harmful persuasive strategies and protect users from coercive influence in vulnerable moments [43].

Across these three directions, this dissertation contributes new datasets, models, and analyses that foreground speech and multimodality in a field historically dominated by text. It articulates empathy as a composite of recognition, interpretive reasoning, and expressive alignment, rooted in the same conversational mechanisms that govern everyday interaction. At the same time, it argues that empathetic behavior must be matched with rigorous evaluation and safety mechanisms to avoid sycophancy and manipulation.

Although modern large language and audio-language models support zero-shot and few-shot inference, curated datasets remain essential. Rather than serving solely as training data, datasets now play a critical role in evaluation, grounding, and safety assessment. Tasks such as empathy detection, conversational entrainment, and mental manipulation rely on subtle prosodic, temporal, and multimodal cues that are not reliably captured by web-scale pretraining alone. High-quality annotated corpora therefore provide the structured supervision needed to define constructs, measure progress, and identify failure modes. Throughout this dissertation, dataset construction is treated not as auxiliary engineering, but as a foundational scientific contribution that enables reliable study of conversational behavior.

By integrating insights from conversation analysis, computational empathy, and recent advances in large language and audio-language models, this work proposes a unified trajectory for empathetic and responsible conversational AI: from understanding how people communicate, to modeling supportive behavior at scale, to safeguarding users as systems grow more emotionally capable.

Part I

Understanding Human Conversations

Overview

Human conversation is structured, adaptive, and interactive. Long before speakers express empathy, they must coordinate meaning — deciding when to speak, how to signal agreement or disagreement, and how to collaborate on shared goals. These mechanisms are not optional embellishments but the fundamental substrate on which all social behaviors — including empathy — are built. In other words, empathy represents a *special case of conversational collaboration*, one where alignment and understanding must extend not only to information exchange but also to another person's emotional state.

This first part of the dissertation investigates two fundamental components of conversational dynamics: the dialog act structures that govern discourse progression and the behavioral entrainment patterns that demonstrate mutual adaptation.

Chapter 2 presents a re-examination of the Switchboard Dialog Act corpus, identifying long-standing alignment flaws between transcripts and speech. These inaccuracies have limited the field's ability to jointly model lexical and prosodic cues for dialog act understanding. We develop and validate a new Re-Aligned Switchboard Dialog Act (RASwDA) corpus and show that leveraging accurate timing and audio improves dialog act classification performance. This effort not only enables more robust modeling of interaction structure, but also contributes a high-impact resource for the broader research community.

Chapter 3 examines how speakers adapt to one another in task-oriented dialogues. Through measurements of lexical and acoustic-prosodic entrainment, we reveal how alignment occurs even in short, goal-directed interactions. Our findings show that speakers dynamically adjust pitch, energy, lexical choices, and disfluencies based on their partner and role in the interaction — illustrating how rapport and efficiency emerge through behavioral coordination.

Together, these studies demonstrate that successful human conversation relies on a pairing

of *structure* and *adaptation*: dialog acts provide a roadmap for how interaction unfolds, while entrainment signals engagement, cooperation, and shared attention. These contributions underscore a key thesis of this dissertation: empathy is not a standalone behavior, but an extension of the same mechanisms that allow conversational partners to understand and align with each other. By strengthening the foundations of dialog structure and interactional alignment, Part I provides the necessary groundwork for modeling empathy in Part II and safeguarding emotionally intelligent systems in Part III.

Chapter 2: Dialogue Acts in Human Conversations

Revisiting and Re-aligning the Switchboard Dialog Act Corpus[1]

2.1 Introduction

Dialog Act (DA) prediction and production is of seminal importance today in research, government and industry, as more and more dialog systems are being built to interact with people for training, education, decreasing human workload in call centers, and providing problem-solving advice. While many corpora have been developed and annotated for building machine learning models in DA prediction or generation tasks, only a few have been transcribed in speech. Many others were annotated using domain-specific DAs or are limited in the number or length of conversations. Among the annotated DA corpora, only the Switchboard Dialog Act (SwDA) Corpus [44] includes domain-independent spoken conversations between two speakers, making it unique for modeling the type of interactions that are primary in most systems used today, such as information services and online chats.

Although the SwDA corpus is widely used for DA prediction and generation tasks, it suffers from a critical limitation: inaccurate alignments. The corpus consists of transcripts and speech derived from the larger Switchboard corpus [45], which were originally aligned using a GMM-HMM speech recognition system. However, these alignment results are unreliable, making it extremely difficult to use both speech and text data to accurately predict or generate DAs. There is very little evidence that the use of speech features from the currently aligned corpus significantly improves their results in any way and sometimes even leads to worse performance [46, 47, 48, 49].

Previous attempts to re-segment the Switchboard corpus, upon which SwDA is built, have

[1] Portions of this chapter previously appeared as R. Chen, E. Lin, et al., "RASwDA: Re-Aligned Switchboard Dialog Act Corpus for Dialog Act Prediction in Conversations, "*14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*, 2024.

resulted in completely different transcriptions and utterance boundaries that do not coincide with those in the SwDA [50]. While the NXT-format Switchboard Corpus links the transcriptions in SwDA with the alignments from [50], it does so for only 642 of the 1,155 conversations in SwDA [51]. To date, no one has produced a full realignment of all 1,155 SwDA conversations.

Our project aims to create an improved, Re-Aligned Switchboard Dialog Act (RASwDA) corpus for DA tasks by manual re-alignment and validation by experts on both sides of all SwDA conversations to correct the errors introduced by the early automatic alignment. Our goal is to 1) produce a more accurate RASwDA corpus for DA prediction and generation tasks and 2) set a new benchmark for identifying DAs using machine learning models that incorporate both text and speech features. We demonstrate that this new version of the SwDA corpus provides more useful information in both text and speech for DA identification models by comparing the new results to models built on the earlier version of the corpus. To encourage the wider community to make use of the fully re-aligned corpus, we will make it publicly available, thereby facilitating the current research efforts focused on modeling human-human and human-machine conversation. [2]

2.2 Related Work

2.2.1 Dialog Act Labeled Corpora

Many corpora, including SwDA, have been annotated for DAs. They vary in domains, languages, types of interactions, and the number and type of annotated DAs. While some corpora were annotated using small tag sets, such as the DCIEM Map Task [52], the AMI Meeting [53] (under 20), and the Columbia Games Corpus (only 7) [54], others were annotated using tag sets with hundreds of tags, such as DIHANA [55] and NESPOLE [56]. Furthermore, some corpora, including the DCIEM Map Task, SwDA, SCHISMA [57], ICSI-MRDA [58], and AMI Meeting, utilized domain-independent tag sets suitable for annotating various corpora. On the other hand, corpora such as VERBMOBIL [59], NESPOLE, DIHANA, LEGO [60], TourSG [61], Ubuntu IRC [62],

[2] Data will be available through Linguistic Data Consortium (LDC), which currently provides many earlier versions of this corpus.

MultiWOz and its multiple updated versions [63, 64, 65], and Audio Visual Scene-Aware Dialog (AVSD) [66] were annotated using domain-dependent tag sets. Notably, many corpora did not include speech data, such as DSTC6 corpora (Twitter, WOCHAT) [67], Ubuntu IRC, and MultiWOz.

Among these DA corpora, SwDA is particularly valuable for investigating how speech and transcripts synergize to facilitate DA modeling in conversations. The corpus contains a substantial number of annotated segments and provides both speech and transcripts with domain-independent data, distinguishing itself from others with a limited number of annotations, such as SCHISMA and DCIEM Map Task. Although ICSI-MRDA and the AMI Meeting corpus also offer sizable annotated speech data in multi-participant meetings, only SwDA exclusively comprises dialogs between two individuals, making it particularly relevant for modeling the types of two-party interactions prevalent in conversational systems today. However, the limitation of SwDA lies in its inaccurate alignment of speech and transcripts, which cannot be used to identify or generate appropriate acoustic-prosodic features, such as pitch, intensity, speaking rate, and voice quality.

2.2.2 The Switchboard Dialog Act Corpus

The original Switchboard Corpus is a corpus of 2,400 two-sided telephone conversations, each between two native speakers of American English from different parts of the United States, and was collected in 1990-91 by Texas Instruments. The initial goal for this corpus collection was to develop speech processing algorithms, particularly speaker verification algorithms [45]. The SwDA corpus [44] was created from a portion of the Switchboard corpus, specifically LDC's Switchboard-1 Release 2 (LDC97S62) [45]. It consists of 1,155 conversations out of the original 2400 conversations, ranging from 1.5 to 10 minutes, comprising a total of 205,000 utterances and 1.4 million words.

SwDA was labeled with an augmented version of the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [68], the SWBD-DAMSL label set of 42 DA labels. The DA labels include items such as *Statement-non-opinion*, *Acknowledge*, and *Statement-opinion*, which represent over two thirds of the 42 DA items annotated; the full list is shown in Table 2.1.

The SwDA conversations were initially force-aligned with the participants' speech in the 1990s using a GMM-HMM Switchboard recognition system to identify the start and end times of speech segments [69, p. 454]. However, due to the limited reliability of ASR systems used during that era and various challenges posed by the recordings and the transcripts, much of this alignment contained major errors, so it is impossible to perform accurate prosodic analysis on the DAs from their poor alignment with the audio.

Problems with this speech aligner included misalignment of reduced and low-energy speech. Based on manual inspection of hundreds of audio files, we have also found that background noise from sources such as static, telephones ringing, children crying, music, radios, and TV's has also reduced the original alignment quality. Problems with the conversations' transcripts at the time included mis-transcribed or simply missing words (some had been excised in a previous transcription task as "not useful words"). Only a small subset of these alignments were corrected to create a small DEV test set. The rest of the corpus was left in its original, poorly aligned state.

2.3 SwDA Alignment Diagnosis

While the SwDA corpus has been widely used to build models to detect different DAs, studies have observed that incorporating the audio information from SwDA does not improve DA prediction or generation scores, and can sometimes even worsen them. This is likely due to the poor alignment of the audio with transcripts and dialog act labels. [46, 47] showed that integrating prosodic information with transcripts improved DA prediction accuracy only for a couple of selected DAs, while having negative or no effects on the rest. The DA recognition model that incorporates prosody reported a lower F1 score, compared to the model trained solely on transcripts [48]. Similarly, [49] found that removing pitch and energy features resulted in only a marginal decrease in accuracy (1% and 0.6%, respectively) for their end-to-end DAC model on the SwDA corpus.

Transcripts and their aligned speech were often completely incorrect. We have found 27 conversations in which speakers were recorded on the wrong channel, resulting in incorrect speaker identifications when we attempt to match speaker audio with transcripts. Overlapping speech

DA	Description	Count (Full)	% (Full)	Count (RASwDA)	% (RASwDA)
sd	Statement-non-opinion	75145	34.26	32406	24.53
b	Acknowledge (Backchannel)	38298	17.46	16297	12.34
sv	Statement-opinion	26428	12.05	11762	8.90
%	Abandoned, Turn-Exit, or Uninterpretable	15550	7.09	6729	5.09
aa	Agree/Accept	11133	5.08	4973	3.76
x	Non-verbal	3630	1.65	3591	2.6
qy	Yes-No-Question	4727	2.15	2053	1.55
ba	Appreciation	4765	2.17	1799	1.36
ny	Yes answers	3034	1.38	1252	0.95
fc	Conventional-closing	2582	1.18	1056	0.80
qw	Wh-Question	1979	0.90	874	0.66
nn	No answers	1377	0.63	595	0.45
bk	Response Acknowledgement	1306	0.60	555	0.42
h	Hedge	1226	0.56	507	0.38
qy ^ d	Declarative Yes-No-Question	1219	0.56	472	0.36
bh	Backchannel in question form	1053	0.48	445	0.34
bf	Summarize/reformulate	952	0.43	444	0.34
^ q	Quotation	983	0.45	427	0.32
fo_o_fw_"_by_bc	Other	883	0.40	408	0.31
na	Affirmative non-yes answers	847	0.39	351	0.27
qo	Open-Question	656	0.30	310	0.23
^ 2	Collaborative Completion	723	0.33	308	0.23
b ^ m	Repeat-phrase	688	0.31	283	0.21
ad	Action-directive	746	0.34	282	0.21
qh	Rhetorical-Questions	575	0.26	265	0.20
^ h	Hold before answer/agreement	556	0.25	219	0.17
ar	Reject	346	0.16	141	0.11
ng	Negative non-no answers	302	0.14	137	0.10
br	Signal-non-understanding	298	0.14	137	0.10
no	Other answers	286	0.13	121	0.09
fp	Conventional-opening	225	0.10	117	0.09
qrr	Or-Clause	209	0.10	98	0.07
arp_nd	Dispreferred answers	207	0.09	91	0.07
^ g	Tag-Question	92	0.04	53	0.04
oo_co_cc	Offers, Options, Commits	110	0.05	52	0.04
t1	Self-talk	103	0.05	44	0.03
bd	Downplayer	103	0.05	43	0.03
aap_am	Maybe/Accept-part	105	0.05	40	0.03
qw ^ d	Declarative Wh-Question	80	0.04	37	0.03
fa	Apology	79	0.04	34	0.03
t3	3rd-party-talk	117	0.05	32	0.02
ft	Thanking	78	0.04	28	0.02

Table 2.1: Comparison of the original SwDA DA counts ("Count (Full)") and our realigned corpus RASwDA DA counts ("Count (RASwDA)"). Original counts from [70].

segments also cause confusion in the automatic alignment process. In many cases, shorter DAs such as *backchannel* or simple "yes" or "no" responses are missed entirely by the aligner. Furthermore, the presence of numerous simple timing errors in earlier parts of the conversations can propagate throughout the rest. These issues further highlight the challenges and limitations of DA modeling based on the SwDA corpus, underscoring the urgent need for its correction and improvement.

2.4 Re-alignment Methods

To produce high-quality alignments between the audio and transcripts of SwDA, we employ a two-step process. First, for conversations among the 642 conversations which are included in the NXT-format Switchboard Corpus [51], we parse time-aligned SwDA transcripts from the NXT-provided XML files into TextGrid format. For conversations not included in the NXT-format Corpus, we parse each conversation's transcript into separate transcripts for each speaker. We also take advantage of the fact that speakers are recorded on separate channels to separate the audio for each conversation into two WAV files, one with each speaker's speech [71]. Then (for transcripts not sourced from NXT-format Switchboard) we compute the forced alignment for each utterance in each speaker transcript and conversation with the *aeneas* library [72], shown in Figure 2.1a. Based on manual inspection, we find that further manual realignment is still necessary to correct forced alignments generated with *aeneas*, as many of the issues that affected the original forced alignments (e.g. background noise) also affect the accuracy of the *aeneas* alignment.

Second, we manually correct the TextGrids produced both from the NXT-format Switchboard Corpus alignments and the *aeneas* forced alignments (Figure 2.1b). We use the Praat speech analysis interface, which allows expert aligners to easily manipulate audio and transcripts simultaneously [73]. Specifically, we convert each SwDA transcript into a TextGrid, a text file format commonly used for annotating audio in Praat.

In addition to correcting the transcript alignment, aligners are also instructed to mark speaker overlap and laughter with the special "SIL" and "⟨laughter⟩" tokens, and correct mis-transcriptions, segmentation errors, and omissions in the transcript. We attempt to resolve mis-transcriptions and

segmentation errors marked by the original SwDA annotators themselves for correction at a later date [44]. Our aligners included 2 high school students, 15 undergraduates, and 8 graduate students in computer science, linguistics, and mathematics, some compensated for their time in either course credit or a stipend.

2.5 Results

Our Re-Aligned Switchboard Dialog Act (RASwDA) corpus currently consists of 537.5 manually realigned and validated conversations (1075 single speaker transcripts) from the 1155 SwDA conversations. Our final goal is to create a new, correctly aligned version of the entire SwDA corpus that is publicly available and to demonstrate the effect of adding correct acoustic-prosodic features for DA prediction.

Table 2.1 presents the counts of different DA tags in the original SwDA corpus as compared to our RASwDA. The original corpus consists of 203,801 dialog acts [70], while our realigned subset of RASwDA contains 98,274 dialog acts and 42,231 silence segments.

2.5.1 DA Classification

By training dialog act classification (DAC) models on 55,049 utterances from RASwDA, we have achieved 59.53% accuracy on a 13,762-utterance validation set constructed from RASwDA, a 2.56% improvement over the 56.97% accuracy reported by [49] on a 4,088-utterance test set from the original SwDA corpus using their state-of-the-art end-to-end neural model trained on 192,768 utterances from the original SwDA corpus (Table 2.2).

Model	[49]	Ours
Dataset	SwDA	RASwDA
Accuracy	56.97	**59.53**
Train	192,768	55,049
Validation	3,196	13,762
Test	4,088	–

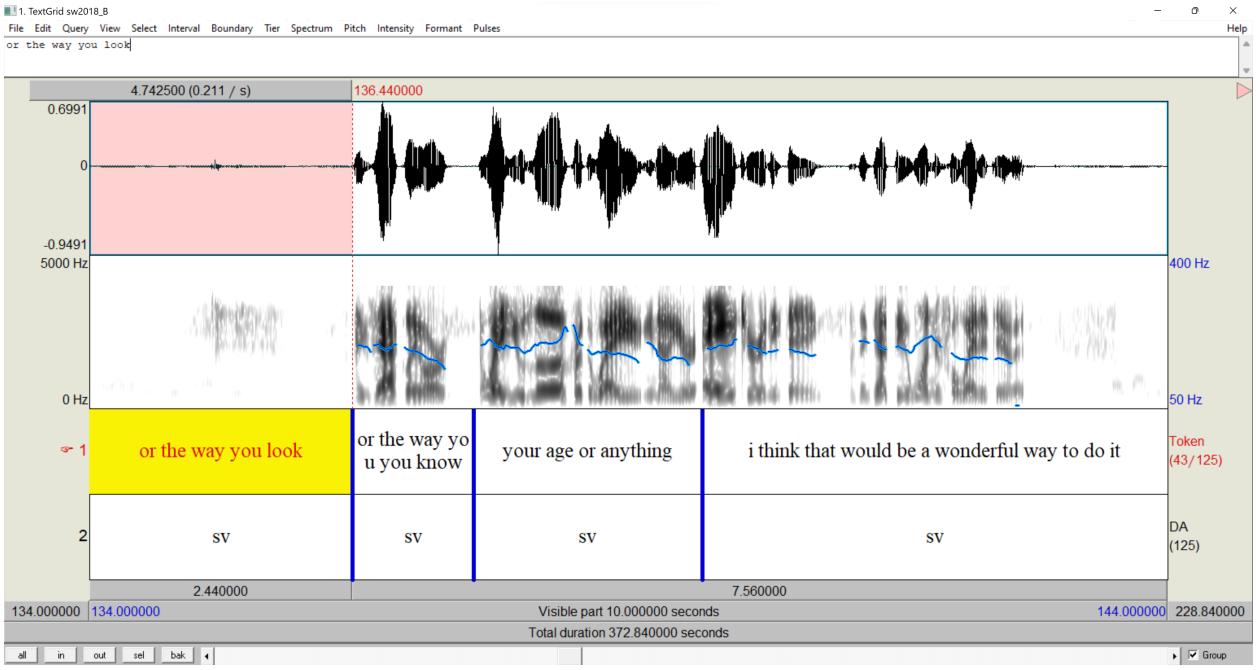
Table 2.2: Dialog act classification accuracy on speech from SwDA and RASwDA corpora, along with sizes of training, validation, and test splits in numbers of utterances.

Our model uses a convolutional neural network (CNN) and treats DAC as an image classification task on spectrograms of the speech signal, as this has proven a successful approach for applications such as emotion recognition [74]. The input to the CNN is a $256 \times 256 \times 3$ spectrogram of the speech signal, computed with matplotlib.[3] The CNN consists of three convolutional layers using 3×3 kernels, each followed by a ReLU layer, normalization, a max pooling layer with a 2×2 window, and another normalization sequentially. The first convolutional layer consists of 32 kernels with a stride of 2 pixels. The second convolutional layer consists of 64 kernels with a stride of 1. The third convolutional layer consists of 128 kernels with a stride of 1. After applying the ReLU non-linearity, normalization, and pooling, the output of the third convolutional layer is flattened into a 32768×1 vector. This vector is then passed through three fully connected layers with normalization. Finally, the softmax function is applied to produce the prediction. We train on a 55,049-utterance subset of RASwDA and validate on a held-out 13,762-utterance subset. We believe that as we continue to build RASwDA by realigning the rest of the SwDA conversations, the model performance will further improve with a larger, more accurate dataset.

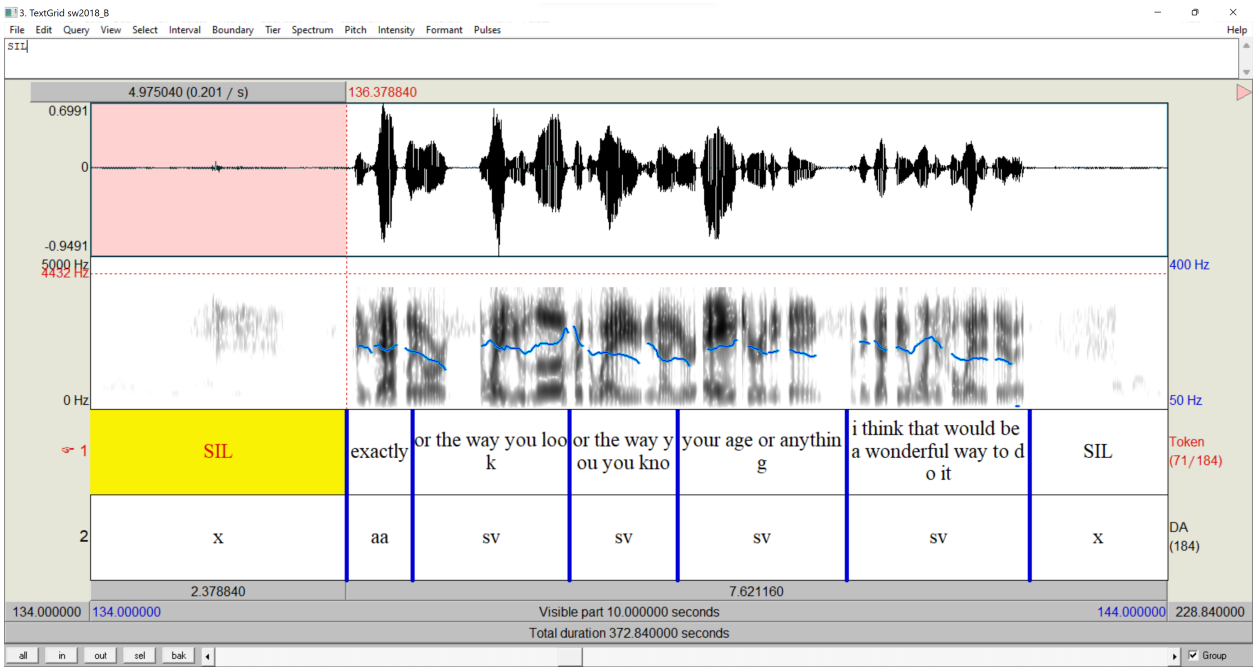
2.6 Conclusion

We have identified inaccuracies in the current automatic alignments of the Switchboard Dialog Act (SwDA) corpus and have undertaken a manual realignment process for a subset of 537.5 out of 1155 conversations. Our Re-Aligned Switchboard Dialog Act (RASwDA) subset has already demonstrated improved performance of state-of-the-art models on the dialog act classification task. We plan to continue the realignment process for the remainder of the SwDA corpus and make it publicly available for the wider speech community.

[3] https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.specgram.html



(a) Automatic alignment.



(b) Automatic alignment + manual correction.

Figure 2.1: A section of a SwDA transcript in the Praat interface (a) before and (b) after manual correction of the automatic alignment generated by *aeneas*. Praat allows aligners to view the waveform and spectrogram of the speech signal (top two sections of display) and a TextGrid transcript (bottom section of display) simultaneously.

Chapter 3: Identifying Entrainment in Human Task-oriented Conversations[4]

3.1 Introduction

Human interlocutors often adapt their behavior to each other in conversations through entrainment, also called accommodation or alignment. People adapt in syntax, word choice, pronunciation, and prosody, as well as in facial expression, posture, and socio-cultural behavior. Studies have found that people who entrain to others are perceived as more socially attractive, more competent and intimate [75, 76]. Entrainment leads subjects to like their conversational partners more and to perceive interactions as more successful [77, 78], and is a good predictor of task success [79]. So producing entrainment in dialog systems is useful in producing more attractive, competent, and intimate conversations that users will enjoy more and consider more successful. To move toward developing entraining dialog systems, we analyze entrainment in human-human conversations collected in a Wizard-of-Oz scenario where two speakers play the user and the agent roles in a task-oriented setup.

Our data is from the DSTC10 Track 2 dataset [80] representing real conversations that a user is expected to have with a task-oriented dialog system. In the context of task-oriented dialog systems, entrainment behaviors can occur in both users and agents. Users may deviate from their typical speech, in sync with what they perceive from the agent and agents may change how they talk to accommodate users: if the user is speaking very slowly, the agent may slow their speech accordingly; if the agent refers to the parking lot as the parking space, the user may also adopt such wording.

As speakers may entrain in both text and speech, we extract acoustic-prosodic features and lexical frequencies to test for proximity, convergence and synchrony aspects of entrainment. We

[4]Portions of this chapter previously appeared as R. Chen, S. Kim, et al., "Identifying Entrainment in Task-Oriented Conversations, "*ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. DOI: 10.1109/ICASSP49357.2023.10095543

conduct significance tests that reveal changes in speech pitch and frequent words as important indicators of entrainment. We also note the effect of duration and speaker roles on entrainment. We expect these findings will guide us in developing dialog systems to entrain the users.

3.2 Related Work

Much research on entrainment has been done in a variety of conversational settings, including discussions of married couples about problems in their relationships [81], children adapting their amplitudes to that of an animated character [82], as well as human-computer interactions. Several studies have reported promising results when entrainment between the users' and system's voices took place: [83] reported gains in ASR accuracy when entrainment of speech rate was induced; [84] also found gains in learning when entrainment between humans and a tutoring system occurred in pitch and intensity; and [85] showed that entrainment improved rapport and naturalness when a system shifted the pitch contour of the synthesized speech by the mean pitch of the user. A positive link was found between entrainment and trust for humans using conversational avatars [86] and entraining the system's lexical choices to those of user's increased the dialog success rate [87]. Our study reports evidence of entrainment in a more realistic setup for task-oriented dialog systems, which are typically much shorter than previously studied entrainment datasets. This makes it much more challenging to identify entrainment since the degree of entrainment correlates with the length of conversations [88, 89].

3.3 Dataset

To study the entrainment behaviors in human-human conversations, we analyzed the DSTC10 Track 2 dataset [80], which includes about 45 hours of recordings of 917 spoken task-oriented dialogues about touristic information for San Francisco. Each dialogue session was collected by two participants, a user and an agent. While the user-side participant was given a set of specific goals to be achieved in each session, the agent-side participant had access to a database to look up relevant information to provide for particular user requests. Entrainment has been documented in

longer conversations such as the Columbia Games corpus, in which sessions between two speakers averaged 45 minutes long [88, 89]. However, our task-oriented conversations are much shorter, with an average of 24.7 turns (min 5 and max 60), and 3 minutes (min 0.45 and max 7.33). Therefore, identifying whether entrainment occurs in much shorter conversations is important to help us incorporate entrainment capabilities into task-oriented human-computer dialogue systems.

3.4 Methods

3.4.1 Entrainment Measures

Following the methods proposed in [88], we evaluate three aspects of entrainment, i) proximity (do speakers in the same conversation sound more similar?), ii) convergence (do speakers entrain more over time?), and iii) synchrony (do speakers' behaviors vary in tandem?) For proximity, we compare the feature differences between partners (participating in the same conversation) with non-partners via a t-test. If a speaker does not entrain, they are expected to sound uniformly talking to any random speaker, whereas when entrainment occurs, the partners within the same conversation adopt each other's text and speech, reducing the feature differences. The partner differences are therefore smaller than the difference between a random non-partner speaker from the dataset.

The turn-level features are extracted and averaged over all turns within a dialog for both the agent and the user as session-level features ϕ_s. A partner difference $\Delta\phi_{\text{partner}}$ is defined as the absolute session-level feature difference between partners in the same conversation. Non-partner differences $\Delta\phi_{\text{nonpartner}}$ are calculated by feature differences between any two speakers that are not in the same conversation.

$$\phi_s = \frac{1}{n}\sum_{i=1}^{n} \phi_{i\text{th-turn}} | \text{ speaker} = s \tag{3.1}$$

$$\Delta\phi_{\text{partner}} = |\phi_s - \phi_{s'}| \text{ for } s \text{ and } s' \text{ in the same dialog}$$

$$\Delta\phi_{\text{nonpartner}} = |\phi_s - \phi_{s'}| \text{ for } s \text{ and } s' \text{ not in the same dialog}$$

For convergence, we split each conversation into two halves by the number of turns and compare partner differences between the first and the second halves. If speakers entrain more over time, we expect feature differences to decrease in the second half of the conversation. The half-session-level features $\phi_{k,s}$ are averaged turn-level features in the k-th half session. We calculate half session partner differences $\Delta\phi_k = |\phi_{k,s} - \phi_{k,s'}|$ and juxtapose those from the same conversations in a pair-wise t-test.

We calculate the Pearson's correlation coefficient for the turn-level feature differences between adjacent turns and the rest for levels of synchrony. If speakers entrain over time in synchrony, speakers may adjust their speech and language in accordance with those of their conversational partner. For example, if Speaker A's pitch rises, Speaker B's will too and follow similar patterns in pitch and other features over the course of the conversation.

3.4.2 Speech-based Entrainment Analysis

The acoustic-prosodic features represent the pitch, energy, voice quality and speaking rate of speakers through 12 features: pitch mean, minimum, maximum, and standard deviation, intensity mean, minimum, maximum, and standard deviation, jitter, shimmer, harmonics-to-noise ratio (HNR), and speaking rate [88, 90]. These features are extracted with praat [91] and parselmouth [92] tools on default parameter settings. The pitch and intensity features are z-score normalized by speaker. The speaking rate is measured by words per second from human transcripts.

3.4.3 Lexical Entrainment Analysis

Our lexical features include linguistic inquiry and word count (LIWC) [93] category frequencies as well as the 25 most frequent word counts for the corpus (MFC), the agents (MFA) and the users (MFU), accounting for the roles of the speakers. Only 25 frequent words are used here due to the short length of the conversations and thus the sparsity of the text. We use LIWC2015 for the lexical categories covering linguistic dimensions, psychological processes, personal concerns and spoken categories [93]. These features are interpretable and shown to be useful representations

Observation	LIWC Categories
proximity	Total function words, 1st person plural, 3rd person plural, Conjunctions, Comparisons, Quantifiers, Anger, Sadness, Social processes, Family, Friends, Female references, Male references, Cognitive processes, Causation, Tentative, Differentiation, Perceptual processes, Hear, Biological processes, Body, Health, Ingestion, Power, Relativity, Space, Time, Work, Leisure, Home, Money, Religion, Death, Swear words, Netspeak, Assent
no proximity	Total pronouns, Personal pronouns, 1st person singular, 2nd person, Impersonal pronouns, Articles, Prepositions, Auxiliary verbs, Common Adverbs, Negations, Common verbs, Common adjectives, Interrogatives, Numbers, Affective processes, Positive emotion, Negative emotion, Anxiety, Insight, Discrepancy, Certainty, See, Feel, Drives, Affiliation, Achievement, Reward, Risk, Past focus, Present focus, Future focus, Motion, Informal language, Nonfluencies
convergence	1st person singular, 1st person plural, Impersonal pronouns, Articles, Prepositions, Auxiliary verbs, Interrogatives, Family, Friends, Causation, Tentative, Certainty, Differentiation, Perceptual processes, See, Biological processes, Health, Ingestion, Achievement, Power, Past focus, Future focus, Motion, Space, Leisure, Home, Money, Nonfluencies
no convergence	Total function words, Total pronouns, Personal pronouns, 2nd person, Common Adverbs, Negations, Common verbs, Common adjectives, Quantifiers, Affective processes, Positive emotion, Negative emotion, Social processes, Male references, Discrepancy, Feel, Drives, Reward, Risk, Present focus, Relativity, Time, Work, Informal language, Assent

Table 3.1: *t*-test results for LIWC categories. The table shows the categories with $p < 0.05$. The rows list the LIWC categories that show proximity, no proximty, convergence, and no convergence, respectively.

for multi-party entrainment [94]. We calculate the percentage of total words in a conversation that match each of the LIWC dictionary categories. Because of the sparsity of lexical features, we only perform analysis of lexical entrainment using session-level proximity and convergence, not turn-level synchrony.

3.5 Results

Entrainment is found in both the speech and the content of conversations. Despite the brevity of the conversations, speakers assimilate in their pitches, intensity variations and HNR and converge on their intensity variations. They also entrain on the frequency of words that are in many LIWC

Feature	Proximity	Convergence
min pitch	−0.03735	1.16309
max pitch	**−8.96402****	−1.32656
mean pitch	**−3.83764****	−7.21694**
sd pitch	**−3.28754****	−2.60738**
min intensity	1.79504*	−1.85953*
max intensity	−0.9323	−4.47257**
mean intensity	0.45932	−7.96386**
sd intensity	**−24.62165****	**9.34142****
jitter	−0.31599	−7.99437**
shimmer	−0.45399	−2.5374**
HNR	**−26.02003****	−2.80482**
speaking rate	−1.00623	−4.47165**

Table 3.2: *t*-test statistics for speech entrainment. * for $p < 0.1$, ** for $p < 0.05$ and bold for entrainment. Negative for proximity and positive for convergence.

categories and the most frequent word list.

3.5.1 Entrainment in Speech and Text

Significant levels of entrainment are observed in the speech data (Table 3.2). The differences between partners in the same conversation are smaller than those between non-partners across most speech features (negative t-statistics for proximity). The t-test indicates significant proximity in pitch max, mean, and standard deviation, intensity standard deviation, and HNR. The t-test on the same speech features shows less entrainment observed in the convergence measure than proximity, which may be due to the short length of the conversations in the dataset. The speakers do not have enough time to produce similar significant levels of entrainment as do longer conversations, which typically exhibit higher assimilation. In Table 3.2 convergence column, a positive t-statistic means that the second half of these conversations has smaller differences, i.e. speakers converge, whereas a negative t-statistic means that they diverge. We only observed convergence in intensity standard deviation at the session level. Most of the other features significantly diverge.

For lexical entrainment, as shown in Table 3.1, among the LIWC categories, we find proximity in linguistic dimensions (function words, conjunctions), psychological processes (anger, sadness), personal concerns (work, home) and spoken categories (assent). The high level of proximity in

Observation		25 Most Frequent Words
proximity	MFC:	that, the, for, 's, okay, can, have, yeah, there
	MFA:	that, the, for, 's, okay, yeah, have
	MFU:	the, that, for, can, 's, okay, have, if, of, there
no proximity	MFC:	you, i, and, me, is, a, it, so, one, uhhh, let, to, do, ummm, in, go
	MFA:	is, me, you, one, i, let, so, and, it, go, do, see, ahead, right, four, all, sure, check
	MFU:	you, i, a, uhhh, and, ummm, to, in, do, place, me, 'm, is, great, it
convergence	MFC:	you, i, that, the, and, for, me, is, a, it, so, one, uhhh, 's, let, to, do, can, ummm, in, have, go, yeah, there
	MFA:	that, is, me, you, one, i, let, so, and, it, the, for, 's, go, do, see, ahead, right, yeah, all, sure, have
	MFU:	the, you, i, a, that, uhhh, for, and, ummm, to, in, can, 's, do, place, me, 'm, have, is, if, of, there, it
no convergence	MFU:	great

Table 3.3: *t*-test results for 25 most frequent words for the corpus (MFC), the agents (MFA), and the users (MFU). The table shows the words with $p < 0.05$. No words found to significantly diverge for MFC and MFA.

the "assent" category also confirms our speculation that cooperative expressions correlate with entrainment. Similarly, we observe lexical convergence in a few LIWC categories such as linguistic dimensions (pronouns, articles) and spoken categories (nonfluencies). The lesser degree of convergence in lexical features aligns with our finding in speech: both are restricted by the short duration of the conversations. The frequent words show high convergence and some proximity (Table 3.3). Top words such as "that" and "you" appear in all three MFC, MFU and MFA groups, also overlapping with the LIWC categories. Conversational partners assimilate frequencies for assent words such as "okay" and "yeah" but dissimulate those of pronouns and disfluencies, and speakers tend to converge on frequent words rather than diverge.

Some categories, such as pronouns, function words, and auxiliary verbs, plausibly reflect stylistic coordination between speakers and are consistent with prior work on linguistic alignment. These features are less semantically tied to topic and therefore more likely to indicate conversational entrainment. In contrast, several content-heavy categories (e.g., Leisure, Money, Work) likely reflect topical coherence within a task-oriented dialogue rather than entrainment per se. For example, when

Feature	Agent	User	All	Short	Long
min pitch	0.09629**	0.09344**	0.06669**	0.07204**	0.06308**
max pitch	0.21015**	0.1765**	0.18478**	0.22395**	0.15574**
mean pitch	0.01656	0.04185**	0.02708**	0.04359**	0.01674*
sd pitch	0.10931**	0.10655**	0.09609**	0.12639**	0.07436**
min intensity	0.04438**	0.04382**	-0.00024	-0.01105	0.00202
max intensity	0.05345**	0.04854**	0.01153	0.02685**	0.00759
mean intensity	0.00818	0.00712	0.00088	0.00334	-0.00011
sd intensity	**0.55497****	**0.55977****	**0.45446****	**0.45526****	**0.45096****
jitter	0.14807**	0.13655**	0.0574**	0.0677**	0.04929**
shimmer	0.12693**	0.12929**	0.10921**	0.09844**	0.11501**
HNR	**0.41778****	**0.41613****	**0.38746****	**0.3763****	**0.39292****
speaking rate	0.01459	0.02538**	0.00933	0.00409	0.01204

Table 3.4: Pearson's correlation test for turn-level speech synchrony entrainment. * for $p < 0.1$, ** for $p < 0.05$. The columns show correlation coefficient for agent entraining to user, user entraining to agent, speaker-agnostic synchrony for all turns, short conversations (< 25 turns), and long conversations (≥ 25 turns), respectively. $|r| \geq 0.3$ moderate or strong correlation are in bold.

both speakers discuss booking tickets or pricing, increased similarity in "Money" terms may arise from shared task context rather than adaptive behavior.

3.5.2 Entrainment and Speaker Roles

As the speakers play different roles in the conversations, we explore whether and how their roles affect the degree of entrainment. For agent entrainment, we compare an agent's speech difference from their user partner's and a corpus averaged agent's speech. We found that the agents entrain to the users in intensity standard deviation. In a similar method, we found that the users entrain to the agents in intensity max. In addition to proximity entrainment, we also tested whether an agent's speech varies with their user partner's. We computed the Pearson's correlation coefficient by pairing an agent's turn-level speech with a previous turn uttered by the user (the first column of Table 3.4). The agents' speech is correlated with their respective users' speech in most acoustic-prosodic features, most notably intensity standard deviation and HNR. Meanwhile, the users also entrain to their agents (the second column of Table 3.4) and the most correlated features are also intensity standard deviation and HNR.

3.5.3 Entrainment and Duration

Prior studies report association of entrainment levels with conversation length. To verify that longer conversations indeed lead to stronger entrainment, we split the dataset into two bins based on number of turns: the short conversation set, with 24 turns or fewer (511 conversations), and the long conversation set, with 25 turns or more (406 conversations). Partner differences are significantly smaller for long conversations in session level pitch max and mean, jitter and HNR, confirming that longer conversations indeed show higher levels of speech entrainment. We also found weak correlation between number of turns and partner similarity in features such as pitch, intensity standard deviation, jitter and HNR. For convergence, short conversations converge in the same speech features as all data in Table 3.2. Long ones converge on a slightly different set of features, converging in min intensity but not on pitch standard deviation, shimmer and HNR; otherwise they are the same as the short conversation set and all data. We compare synchrony entrainment between short and long conversations, finding results comparable to global turn-level synchrony (Table 3.4). Speech features such as intensity standard deviation and shimmer show moderate positive correlation.

3.6 Conclusion

Our analysis of entrainment in the DSTC10 dataset demonstrates that entrainment does occur between speakers in task-oriented but shorter human-to-human conversations, which differ from previously studied corpus in style, domain and length. Based on the features of speech and lexical entrainment we have identified, we aim to improve the performance of state-of-the-art dialog system models for similar conversations. For our next step, we will explore other potential factors that may affect the *degree* of entrainment in dialogs. It has been shown that the purpose of the turn also correlates with lexical entrainment levels, and that people tend to entrain more in certain dialog acts, such as conventional-opening and closing [95]. We hope to identify similar results for dialog acts in speech data in our future research.

Part II

Empathetic Conversations

Overview

While all human conversation requires coordination and mutual understanding, empathetic interaction represents a deeply collaborative case: one where alignment is not only structural and behavioral but also emotional. To respond supportively, an interlocutor must recognize another person's affective state, comprehend its context, and provide responses that validate and help regulate that state. Building such capabilities in conversational AI remains a core challenge in socially aware language technologies.

This part explores computational empathy across multiple dimensions: perception, interpretation, and response generation. In Chapter 4, we investigate how empathy manifests in speech through acoustic-prosodic cues and lexical choices, and construct a segment-level multimodal model that distinguishes empathetic from neutral responses. This work shows that the voice is a critical channel for emotional attunement and provides insights into how empathy is communicated in real interactions.

In Chapter 5, we introduce a deeper examination of the inherent complexity of empathy itself. Empathy is not always expressed consistently across modalities: tone may convey warmth even when wording is neutral, or facial expressions may contest what is said. Through a framework based on multimodal model disagreement, we identify and analyze cases that are ambiguous or contextually nuanced — the very situations where empathy is most needed. Rather than treating disagreement as error, this work reframes it as a lens into subtle human emotional dynamics and a driver for more robust empathetic modeling.

In Chapter 6, we shift from understanding empathy to enabling it at scale, addressing the rapidly evolving landscape of large language models (LLMs) by developing scalable methods for empathetic data generation. We introduce SYNTHEMPATHY, a psychotherapy-grounded corpus of 106k first-person narratives and empathetic responses generated without crowdsourcing. Through

Chain-of-Empathy prompting, we demonstrate improvements in fine-tuned models' emotional reaction capabilities and provide a generalizable framework for creating domain-specific synthetic data in low-resource areas. This work reflects a broader shift in NLP toward leveraging synthetic data and instruction-guided models to tackle social understanding tasks that were previously limited by annotation cost.

Altogether, these contributions advance the view that empathy is not a monolithic signal but a composite of *recognition*, *interpretation*, and *expressive alignment*. Part II shows how computational models can learn to perceive and generate the emotionally collaborative aspects of conversation — directly building on the structural and interactional foundations established in Part I. This sets the stage for Part III, where we consider how these emotionally capable systems can be protected against misuse and grounded in responsible deployment.

Chapter 4: Detecting Empathy in Speech[5]

4.1 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 [96]. 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 [97, 98]. 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 [99].

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

[5]Portions of this chapter previously appeared as R. Chen, H. Chen, et al., "Detecting Empathy in Speech, "*Interspeech 2024*, 2024. DOI: 10.21437/Interspeech.2024-347

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.

4.2 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 [100]'s Rea, a conversational agent; [101] and [102]'s Virtual Laboratory Exercise Agents, created to improve daily exercise interactions; [103] and [104]'s Rapport Agents with human-like listening behavior including backchannels, turn-ending identification, smiles and nods; [105]'s Greta, used to evaluate different methods of combining emotional facial expressions to produce empathy; [106]'s Jade Semantics Agents which added more empathetic emotions beyond happy and sad in email messages to the mix; [107]'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 [108] and generate empathetic responses [5], 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 [27, 109]. For turn-level empathy, [110] 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 [111], and Japanese [29], 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.

4.3 Dataset

4.3.1 Data Collection

We have collected an empathetic dataset consisting of 346 English videos and about 53 hours in total.[6] The key dataset statistics are summarized in Table 4.1. 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.

[6]Code and data are released at: https://github.com/run-chen-nlp/empathy

Figure 4.1: Example Video of an Interview Between a Therapist and Katy Perry

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 4.1: Empathetic Dataset Summary

Figure 4.2: Manual Re-alignment And Annotation with Praat

4.3.2 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`[7] diarization model. However, as the quality of the transcripts and alignments require further manual correction, we re-align the transcripts using the Praat[8] interface shown in Figure 4.2. 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

[7]https://github.com/pyannote/pyannote-audio
[8]https://www.fon.hum.uva.nl/praat/

therapy such as therapeutic empathy system of empathetic attunement, attitude/stance, communication, and technical/conceptual knowledge [112, 113] The four stages are defined below with examples shown in Table 4.2.

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 4.2: 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.

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

4.4.1 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 [91] and parselmouth [92] 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 [114, 115]. 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 4.3, most acoustic-prosodic features are significantly different.

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

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 4.3: T-test statistics on acoustic-prosodic features for empathetic and neutral speech. ** for $p < 0.05$ after Bonferroni correction.

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[9] 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.

4.4.2 Lexical Analysis

We further investigate lexical features associated with empathy, including significant LIWC dictionary categories [93], lexical diversity [116], concreteness scores [117], hedging frequencies [118, 119] and readability scores [120, 121]. 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

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

for us to understand or convey empathy, in contrast to the speech analysis in the previous section 4.4.1.

The LIWC dictionary categories [93] 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. [116]'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* [118], 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 [117] 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.

Lower readability scores indicate the complexity of empathetic speech. The Flesch Reading Ease scores [120] 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 [121] 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.

4.5 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 [122]. 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`) [10].

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 4.3). Each segment transcript is encoded with a pretrained `roberta-base` encoder and

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

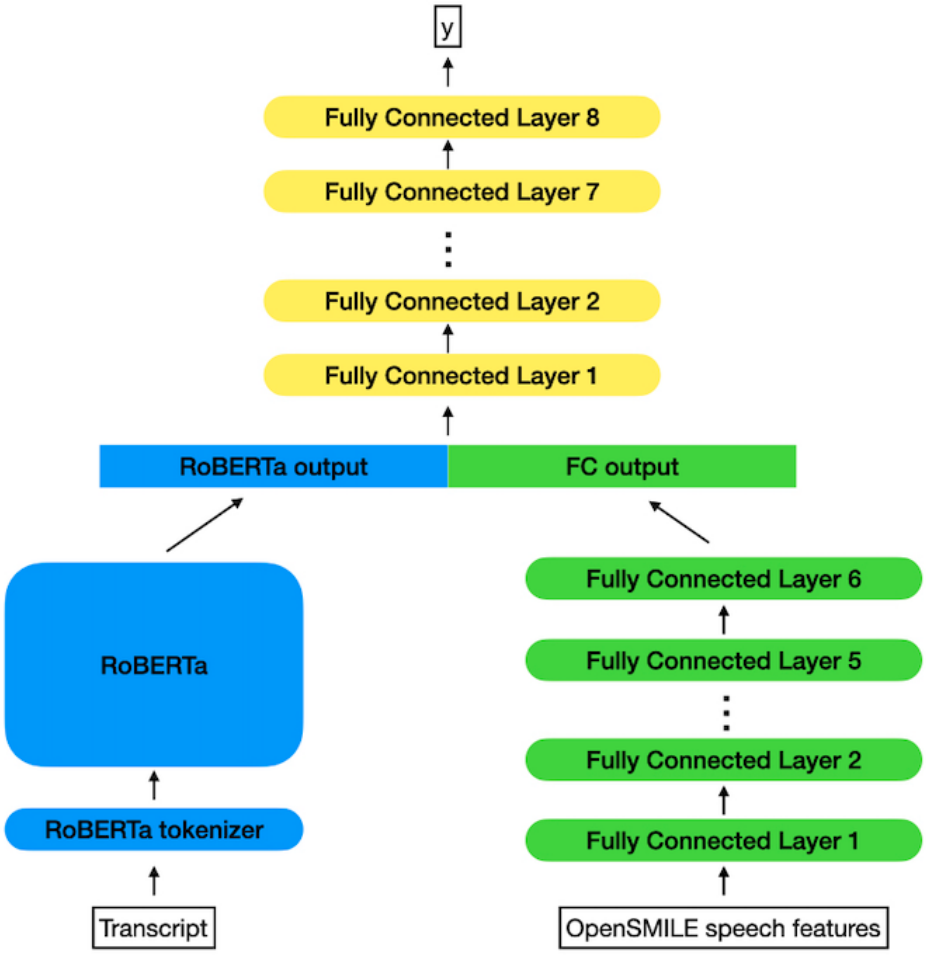


Figure 4.3: RoBERTa+openSMILE multimodal model architecture. Each fully connected layer is followed by a ReLU activation and 0.1 dropout, except the last fully connected layer 8.

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 4.4: 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 4.4. 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 4.4.1 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.

4.6 Conclusion

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.

4.7 Limitations

While several acoustic-prosodic features were identified as predictive of empathetic behavior, such as lower speaking rate and reduced intensity, these patterns should be interpreted with caution. The EMPSPEECH corpus is derived primarily from therapist–client style interactions, where speakers adopt a calm, measured, and supportive speaking style. As a result, some features may reflect the conventions of therapeutic discourse rather than universally applicable markers of empathy.

For example, slower speech and reduced intensity may correspond to what might be described as a "therapist tone," which is common in clinical or counseling settings but may not generalize to other empathetic contexts such as peer support, casual conversations, or emotionally expressive interactions where empathy may instead be conveyed differently.

We therefore distinguish between features that plausibly reflect general interpersonal adaptation (e.g., prosodic alignment, reduced interruptions, and smoother turn-taking) and those that may be domain-specific artifacts of the therapeutic context. The latter should not be interpreted as universal correlates of empathy without validation across additional domains and speaking styles.

Future work should evaluate these features on more diverse corpora, including informal conversations, cross-cultural settings, and non-clinical support dialogues, to determine which acoustic markers generalize robustly.

Chapter 5: Multimodal Empathy: Understanding Complexity Through Model Disagreement[11]

5.1 Introduction

Empathy recognition in human communication is a nuanced and multifaceted task and a core component of socially intelligent systems [123]. Commonly defined as the capacity to understand others and share their emotional experiences, empathy encompasses both cognitive perspective-taking and affective resonance [124]. In human interactions, language, speech, and visual cues jointly convey emotional intent [125]. For example, a speaker's verbal message may appear neutral, yet their vocal prosody or facial expressions may signal warmth or concern. It is then the listener's responsibility to draw inferences about meaning based on a *combination* of these signals.

For AI systems, effectively interpreting these multimodal signals requires not only accurate unimodal representations but also robust integration of potentially conflicting information across modalities. Despite recent advances in multimodal emotion recognition [126], empathy recognition remains particularly complex, as unlike discrete emotions such as anger or joy, empathy often arises from subtle contextual cues that may not align across modalities [127]. For example, a neutral utterance might be perceived as warm or concerned when accompanied by a sympathetic tone or expression.

Our work investigates some of the complexities of multimodal empathy detection by examining instances of disagreement between multimodal models and their unimodal counterparts. In parallel, humans annotate unimodal and multimodal examples in our dataset for the presence of empathy. Our

[11] Portions of this chapter previously appeared as M. Srikanth, R. Chen, et al., "Mixed Signals: Understanding Model Disagreement in Multimodal Empathy Detection, "*Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025. [Online]. Available: https://aclanthology.org/2025.findings-ijcnlp.124/

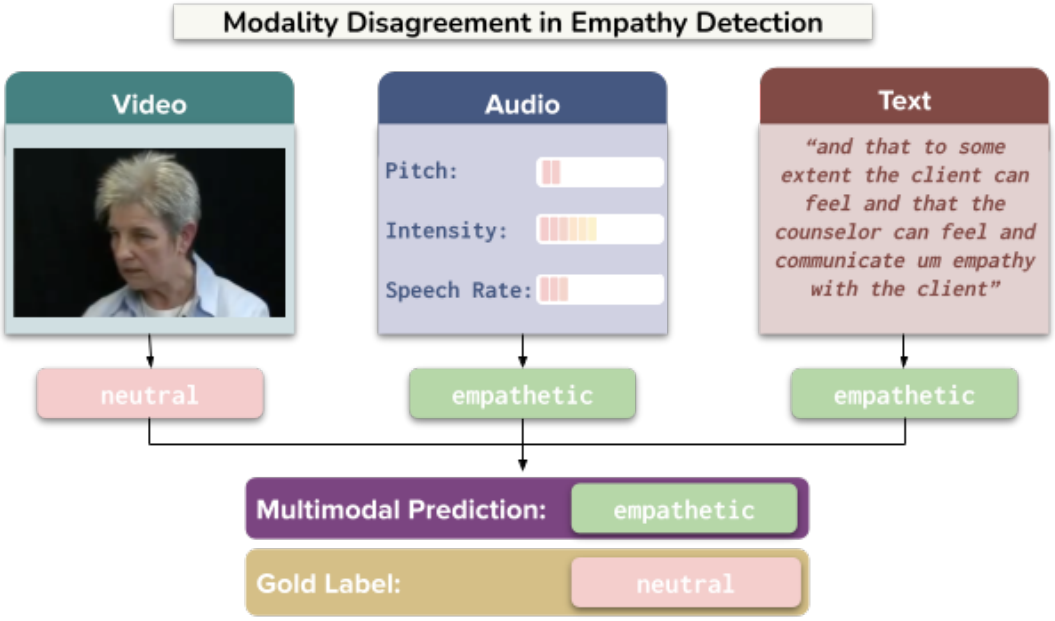


Figure 5.1: Given classifications provided by a single modality, we identify cases where integrating additional modalities leads to a different prediction. We analyze these differences to understand when and why they occur.

analyses reveal that instances of multimodal and unimodal model disagreement often correspond to examples that are difficult for human annotators as well, highlighting examples that are particularly challenging, ambiguous, or nuanced. By linking multimodal and unimodal *model* disagreement to *human* disagreement, we offer new insight into the limitations of current empathy modeling and highlight the value of disagreement-based analysis in socially grounded language tasks.

5.2 Related Work

Empathy Modeling. Early computational work on empathy has focused on generating emotionally relevant textual responses [128, 129], but these approaches are inherently limited by the absence of non-verbal cues, which are critical to empathic understanding. Recent datasets such as EMPATHICSTORIES++ [130], MEDIC [131], EMMI [132] and EMPSPEECH [18] address this limitation by incorporating context, speech, and facial expressions, enabling more comprehensive modeling of empathy. These resources have motivated frameworks such as PEGS [133], EMOKNOB [19] and

SYNTHEMPATHY [15], which further extend multimodal empathetic generation by leveraging large language models (LLMs) and large audio-language models. Despite these advances, empathy still remains difficult to model due to its reliance on subtle, often conflicting signals across modalities. Prior work has largely focused on improving multimodal fusion strategies under the assumption that modalities are complementary [134, 135], but has paid less attention to when fusion may fail or introduce noise.

To the best of our knowledge, our work is the first to evaluate multimodal disagreement on speaker-centric empathy detection datasets. Most publicly available empathy datasets (such as EMPATHICSTORIES++ [130] and OMG-EMPATHY [136]) are fundamentally structured around listener response, not speaker expression. In these datasets, the task is to predict how empathetic a listener feels after hearing a story, rather than to assess whether the speaker themselves is expressing empathy. For instance, in EMPATHICSTORIES++, participants record personal stories and then rate their own emotional responses, framing empathy as a *reaction* to the content rather than as a property of the speaker's delivery. Similarly, OMG-EMPATHY evaluates listener self-reported affective states following brief monologues, again focusing on perceived empathy rather than expressed empathy. This distinction matters because listener-focused tasks inherently entangle speaker behavior with listener subjectivity, making it difficult to isolate which cues (textual, audial, or visual) are directly responsible for empathy expression. In contrast, the dataset EMPSPEECHwe use in this study [18] is one of the only accessible resources that explicitly asks annotators to evaluate the *speaker's* empathy, based solely on the speech segment itself, across modalities. This framing allows us to analyze how empathy is expressed in real time by the speaker, independent of listener interpretation, and enables direct comparisons between modalities on their ability to convey empathetic intent. Our current dataset offers a uniquely valuable lens into the structure of empathy as a speaker-side communicative behavior: something that remains underexplored in the literature. In future work, our modality-disagreement diagnostic could be used to flag nuanced, high-ambiguity segments that challenge *listener* empathy models. They could serve as an effective proxy for identifying segments that elicit high listener variance in empathy judgments, enabling targeted annotation and model

refinement on exactly those ambiguous utterances where listener-centric prediction systems struggle most.

Dataset Difficulty. Complementary lines of work have investigated data difficulty and model disagreement as tools for understanding model behavior. [137] propose *dataset cartography*, a method to identify hard or ambiguous training samples, showing how difficulty-aware instance selection improves benchmarking and reveals mislabeled or trivial examples. [138] demonstrate that difficult instances are also harder for both humans and models to explain, and [139]'s Learning-From-Disagreement (LFD) framework underscores the importance of examining disagreements between models to gain deeper, actionable insights into their behaviors. Although ambiguity is intrinsic to empathy modeling, disagreement-based diagnostics do remain underexplored. As such, we leverage modality disagreement to flag difficult examples that both mislead fusion models and elicit annotator uncertainty.

5.3 Methods

Data was split into train, test and validation sets using random sampling, with an 80-10-10 split. We run fine-tuning and inference for all open-source models on an A100 GPU in Google Colab.

5.3.1 Unimodal Model Training Details

Each model is trained on a binary empathy classification task using precomputed 768-dimensional embeddings. We freeze all but the final two transformer layers and train for fifteen epochs with a learning rate of 5e-6 and batch size of eight.

5.3.2 Fusion Model Details

Each unimodal model representation is independently gated and passed through an additive attention mechanism that computes modality-specific weights. The weighted embeddings are aggregated and classified using a three-layer feedforward network with max pooling. The fusion

model is trained for ten epochs using a learning rate of 1e-4 and includes modality dropout during training. To characterize how the model balances each of the three modalities at inference time, we computed the per-modality gate-weight distributions over the full test set (Table 5.1). The mean gate weights indicate that our model allocates substantial importance to each modality, with only a slight preference toward audio and video. The high variance also shows that the model dynamically adapts its reliance on each modality on a per-sample basis. Thus, our fusion model draws substantially on all three streams; no single modality is systematically favored.

Modality	Mean	Standard Deviation
Text	0.430	0.297
Audio	0.502	0.227
Video	0.476	0.315

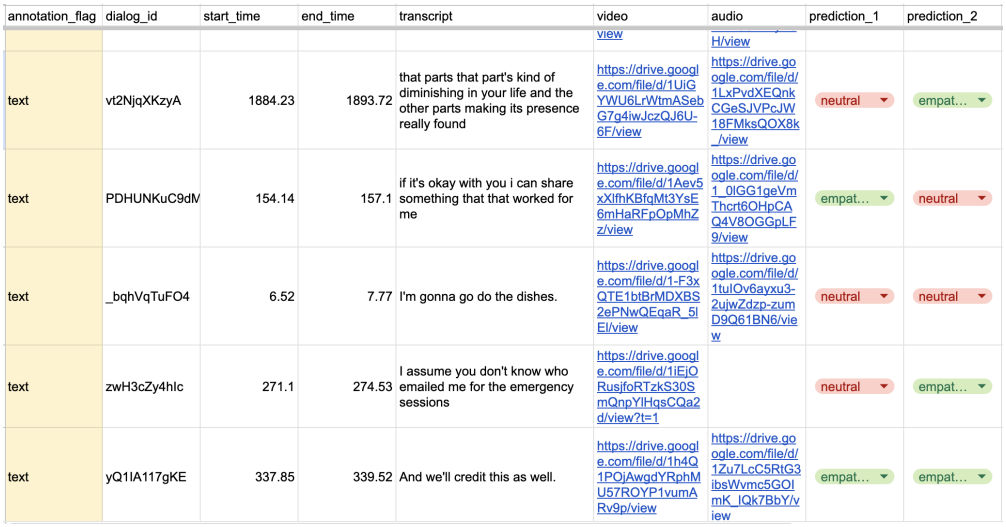
Table 5.1: Per-modality gate-weight distributions over the full test set.

5.3.3 Annotation Instructions

We employed two annotators, oone internal researchef and one external annotator, both fluent English speakers based in the United States. No additional demographic information was collected, as the annotation was conducted internally for research purposes.

Annotators were asked to provide two judgments per example, labeling each as either empathetic or neutral (Figure 5.2). A excerpt describing empathy (drawn from the Encyclopedia of Social Psychology, Volume 1, [124]) was provided to ensure a consistent conceptual foundation for annotation:

> Empathy is often defined as understanding another person's experience by imagining oneself in that other person's situation: One understands the other person's experience as if it were being experienced by the self, but without the self actually experiencing it. There are three commonly studied components of emotional empathy. The first is feeling the same emotion as another person (sometimes attributed to emotional contagion, e.g., unconsciously "catching" someone else's tears and feeling sad oneself). The second component, personal distress, refers



annotation_flag	dialog_id	start_time	end_time	transcript	video	audio	prediction_1	prediction_2
					view	H/view		
text	vt2NjqXKzyA	1884.23	1893.72	that parts that part's kind of diminishing in your life and the other parts making its presence really found	https://drive.google.com/file/d/1UiGYWU6LrWtmASebG7g4iwJczQJ6U-6F/view	https://drive.google.com/file/d/1LxPvdXEQnkCGeSJVPcJW18FMksQOX8k_/view	neutral	empat…
text	PDHUNKuC9dM	154.14	157.1	if it's okay with you i can share something that that worked for me	https://drive.google.com/file/d/1Aev5xXlfhKBfqMt3YsE6mHaRFpOpMhZz/view	https://drive.google.com/file/d/1_0lGG1geVmThcrt6OHpCAQ4V8OGGpLF9/view	empat…	neutral
text	_bqhVqTuFO4	6.52	7.77	I'm gonna go do the dishes.	https://drive.google.com/file/d/1-F3xQTE1btBrMDXBS2ePNwQEqaR_5lEl/view	https://drive.google.com/file/d/1tuIOv6ayxu3-2ujwZdzp-zumD9Q61BN6/view	neutral	neutral
text	zwH3cZy4hIc	271.1	274.53	I assume you don't know who emailed me for the emergency sessions	https://drive.google.com/file/d/1iEjQRusjfoRTzkS30SmQnpYlHqsCQa2d/view?t=1		neutral	empat…
text	yQ1IA117gKE	337.85	339.52	And we'll credit this as well.	https://drive.google.com/file/d/1h4Q1POjAwgdYRphMU57ROYP1vumARv9p/view	https://drive.google.com/file/d/1Zu7LcC5RtG3ibsWvmc5GOImK_lQk7BbY/view	empat…	empat…

Figure 5.2: Annotation Interface

to one's own feelings of distress in response to perceiving another's plight. The third emotional component, feeling compassion for another person, is the one most frequently associated with the study of empathy. Cognitive empathy refers to the extent to which we perceive or have evidence that we have successfully guessed someone else's thoughts and feelings.

Annotators were given an annotation flag indicating which modality to use for the first pass; for instance, if the flag was text, only the transcript was to be used to make the first prediction. After submitting the first judgment, annotators were then given access to the full video, including all available audio, visual, and textual information. They were then asked to provide a second prediction.

5.4 Experiment 1: Identifying Complex Examples from Modality Disagreement

Disagreement between models trained on different modalities can reveal challenging, nuanced, or ambiguous examples. Here, we identify and analyze such cases of disagreement in binary empathy detection using the EMPSPEECH [18] dataset introduced in Chapter 4, consisting of 1,718 manually annotated English speech segments labeled as empathetic or neutral. Table 5.2 shows example utterances below and above the median; examples below the median often include short phrases and backchannels, while examples above the median are often complete sentences with

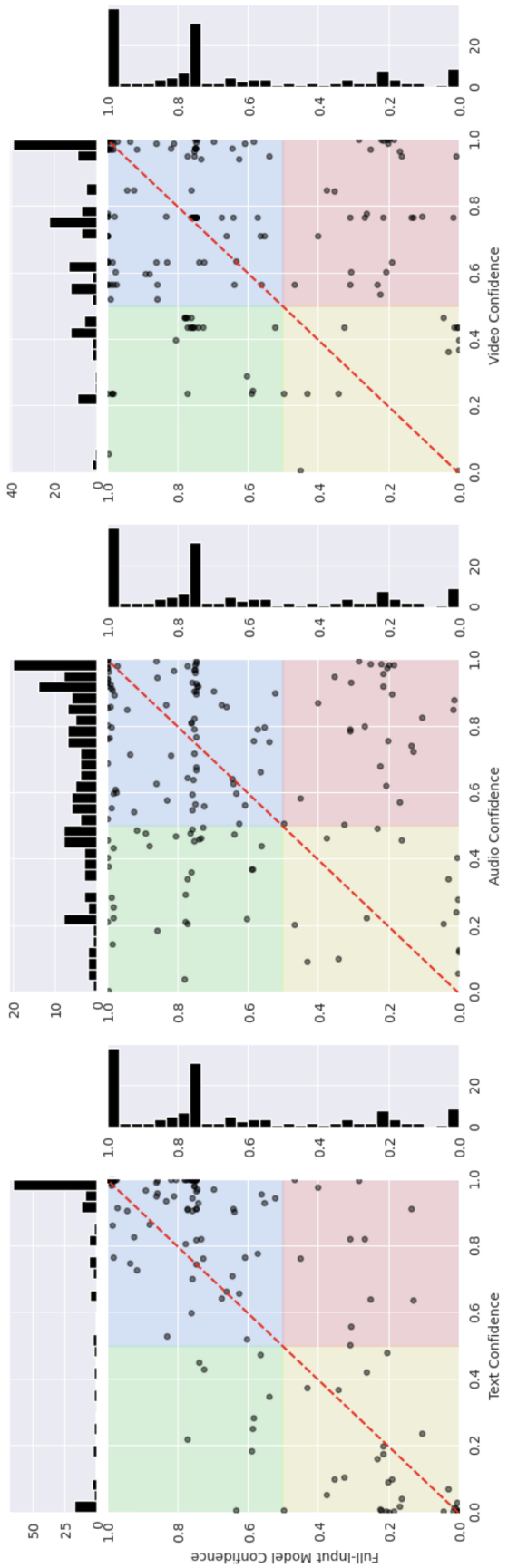


Figure 5.3: Comparing predictions between unimodal (text, audio, video) and multimodal models. We highlight regions where model predictions *agree* (blue and yellow quadrants) and disagree (red and green quadrants).

richer lexical and syntactic structure.

< 6 Tokens	**≥ 6 Tokens**
"there's no way" (3 tokens)	*"No, no he's a good guy go easy on him he's lost his son, Fabio"* (15 tokens)
"You lost it?" (3 tokens)	*"You kids have the biggest hearts I've ever seen."* (9 tokens)
"I can understand that." (4 tokens)	*"congrats my dude, on everything man"* (6 tokens)

Table 5.2: Example utterances with fewer than six tokens (left) versus at least six tokens (right).

Experimental Setup. Examples in EMPSPEECH include video segments spanning three modalities: text (transcript), audio (speech), and video. The task is to predict whether the input contains empathetic (1) or neutral (0) speech.

We finetune two models per modality on the training set from EMPSPEECH: ROBERTA [140] and DEBERTA [141] for text, HUBERT [142] and WAV2VEC2 [143] for audio, and VIDEOMAE [144] and TIMESFORMER [145] for video[12]. Note that the video models process only visual frames without audio, allowing us to isolate the contribution of facial expressions and visual cues from acoustic information. Then, we extract 768-dimensional embeddings from each best-performing unimodal model (ROBERTA, HUBERT, and VIDEOMAE; Table 5.3) to train a multimodal fusion model that projects all three modalities into a shared latent space . Each modality embedding passes through an independent sigmoid gate that adaptively scales its contribution before fusion. The gated embeddings are then passed through an additive attention layer: each is projected into a shared attention space and scored against a learned attention vector. These scores are normalized across modalities to compute a weighted sum that forms the fused representation.

Results. We evaluate all models (unimodal and multimodal) on the test split of EMPSPEECH to identify *disagreements*, or examples where two models with varying input modalities assign *different* labels, highlighting those cases where different modalities may carry ambiguous, conflicting, or modality-specific signals.

[12]The hidden layer dimensions of all models we consider are similar.

Modality	Model	Accuracy	F1
Text	**RoBERTa**	**0.75±0.02**	**0.73 ±0.02**
	DeBERTa	0.69±0.02	0.68±0.02
Audio	**HuBERT**	**0.72±0.01**	**0.71±0.01**
	Wav2Vec2	0.68±0.01	0.63±0.02
Video	**VideoMAE**	**0.77±0.02**	**0.77±0.02**
	TimesFormer	0.64±0.02	0.62±0.02
Fusion (All Modalities)		**0.76±0.02**	**0.72±0.02**

Table 5.3: Fine-tuned model performance by modality on empathy classification (mean ± std over five runs).

Modality	Text	Audio	Video
Text	–	0.338	0.318
Audio	0.338	–	0.253
Video	0.318	0.253	–
Full	0.214	0.383	0.331

Table 5.4: Pairwise disagreement rates among unimodal models and the fusion model, computed as the fraction of test examples with differing predictions.

Text shows the highest disagreement with audio and video (Table 5.4), while audio and video align more closely. This difference likely reflects shared nonverbal cues such as prosody and facial expression. The fusion model's minimal disagreement with text suggests a bias toward verbal content, possibly mirroring the annotators' own reliance on textual signals.

Figure 5.3 visualizes disagreement regions between each unimodal model and the fusion model. We plot unimodal confidence (x-axis) against fusion confidence (y-axis) in the correct label; hence confidence greater than 0.5 results in a correct prediction. This yields four quadrants: green (multimodal correct, unimodal incorrect), red (multimodal incorrect, unimodal correct), blue (both correct), and yellow (both incorrect). Red and green quadrants are disagreement regions which we explore to identify complex examples.

Table 5.6 shows video frames, transcripts, annotator judgments, and the true labels for examples from each confidence plot quadrant. These examples illustrate that disagreement quadrants often contain more ambiguous instances for both humans and models where cues from different modalities

Feature	p (Red vs Blue)	Direction	p (Green vs Blue)	Direction
valence	**0.0047**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{red}}}$	0.5166	$\mu_{\text{green}} > \mu_{\text{blue}}$
arousal	**0.0065**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{red}}}$	**0.0136**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{green}}}$
Mean Pitch	**0.0100**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{red}}}$	**0.0001**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{green}}}$
dominance	**0.0108**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{red}}}$	0.0667	$\mu_{\text{blue}} > \mu_{\text{green}}$
Min Pitch	**0.0333**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{red}}}$	**0.0001**	$\boldsymbol{\mu_{\text{blue}} > \mu_{\text{green}}}$
Jitter	**0.0347**	$\boldsymbol{\mu_{\text{red}} > \mu_{\text{blue}}}$	0.0667	$\mu_{\text{green}} > \mu_{\text{blue}}$
Max Intensity	0.1260	$\mu_{\text{red}} > \mu_{\text{blue}}$	**0.0023**	$\boldsymbol{\mu_{\text{green}} > \mu_{\text{blue}}}$
Mean Intensity	0.1599	$\mu_{\text{red}} > \mu_{\text{blue}}$	0.5329	$\mu_{\text{blue}} > \mu_{\text{green}}$
HNR	0.2217	$\mu_{\text{blue}} > \mu_{\text{red}}$	0.2055	$\mu_{\text{blue}} > \mu_{\text{green}}$
speaking_rate	0.2723	$\mu_{\text{blue}} > \mu_{\text{red}}$	0.9991	$\mu_{\text{green}} > \mu_{\text{blue}}$
Shimmer	0.4122	$\mu_{\text{red}} > \mu_{\text{blue}}$	0.1541	$\mu_{\text{blue}} > \mu_{\text{green}}$
Max Pitch	0.6845	$\mu_{\text{red}} > \mu_{\text{blue}}$	0.2647	$\mu_{\text{blue}} > \mu_{\text{green}}$
Min Intensity	0.7999	$\mu_{\text{blue}} > \mu_{\text{red}}$	0.1571	$\mu_{\text{blue}} > \mu_{\text{green}}$

Table 5.5: T-test results comparing audio features between red vs. blue and green vs. blue examples. Statistically significant p-values are bolded.

may conflict, while examples from agreement quadrants typically display alignment between modalities.

5.4.1 Modality-Based Feature Analysis

To better understand examples in disagreement regions, we extract and analyze modality-based human interpretable features.

Audio. We extract twelve prosodic and paralinguistic features from audio signals: nine low-level acoustic features using PRAAT [146] and PARSELMOUTH [147], and three high-level affective dimensions: valence, arousal, and dominance using a finetuned WAV2VEC2 [148] model. We compare feature distributions using t-tests for examples in disagreement quadrants (red and green) compared to those in the blue quadrant, signifying non-ambiguous, easy examples. Table 5.7 provides an internal comparison between the disagreement quadrants. Blue examples have several significantly elevated pitch-related values than red examples (Table 5.5), suggesting that stronger prosodic fluctuations are frequently corroborated by other modalities. Examples in the green quadrant show significantly higher *Max Intensity* than in blue, potentially reflecting the role of

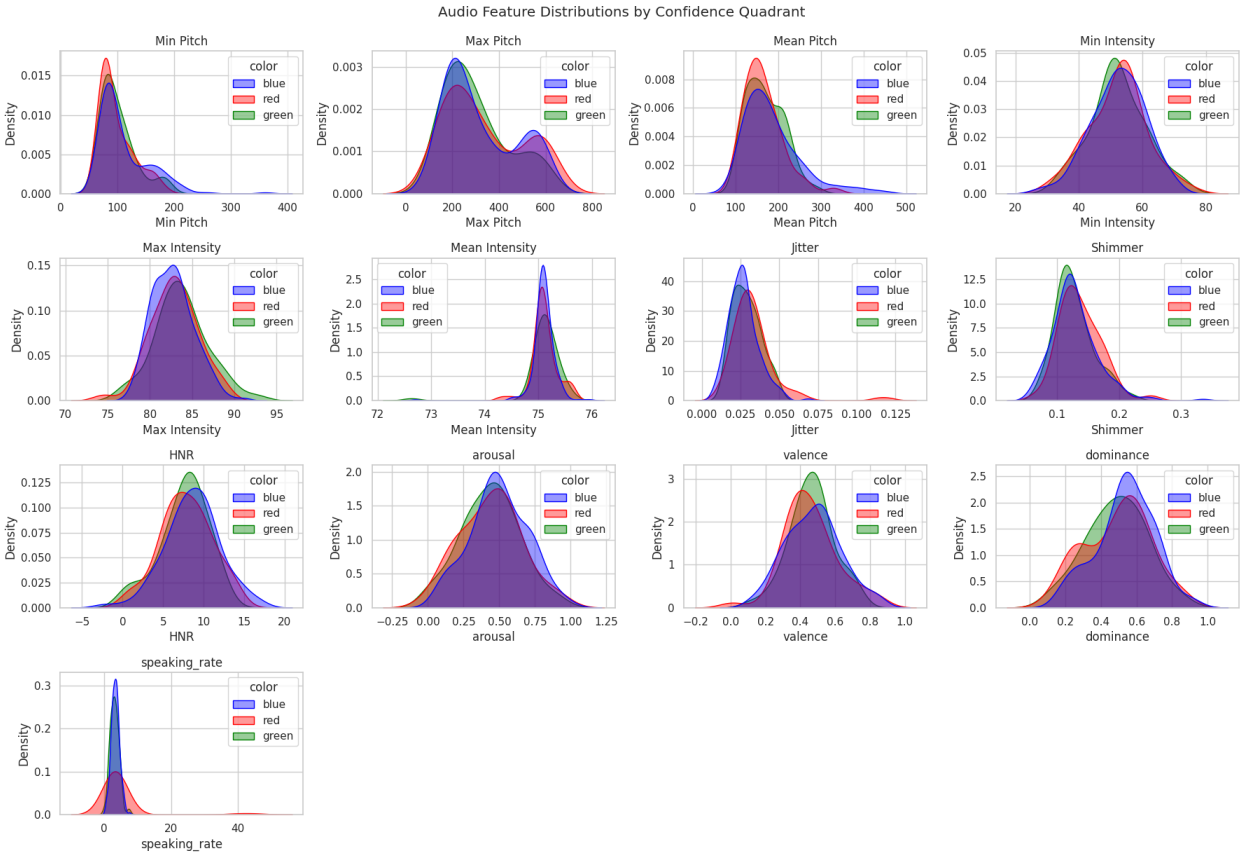


Figure 5.4: Distribution of audio features for red, green and blue examples across the confidence quadrants. Red examples are those correctly classified by the unimodal audio model but misclassified by the multimodal model; green examples represent the reverse. Blue examples represent those correctly classified by both the unimodal audio model and the multimodal model. Significant differences appear in pitch and intensity-based features.

volume-based emphasis in aiding unimodal predictions. Both red and green examples exhibit significantly lower arousal than blue examples, suggesting that these less-aroused, subtler examples lack sufficient affective intensity, which misleads both unimodal and multimodal models. Figure 5.4 presents the distribution of selected audio features (e.g., pitch, intensity) for red, green, and blue examples, highlighting acoustic patterns associated with model disagreement.

Video. We examine facial action unit (AU) activations [149] from video. AU04 (Brow Lowerer), AU12 (Lip Corner Puller), and AU05 (Upper Lid Raiser) show significant differences across example types, revealing how specific facial expressions contribute to perceptual ambiguity (Table 5.8). Figure 5.5 shows activation rates for facial Action Units (AUs) in red, green, and blue examples,

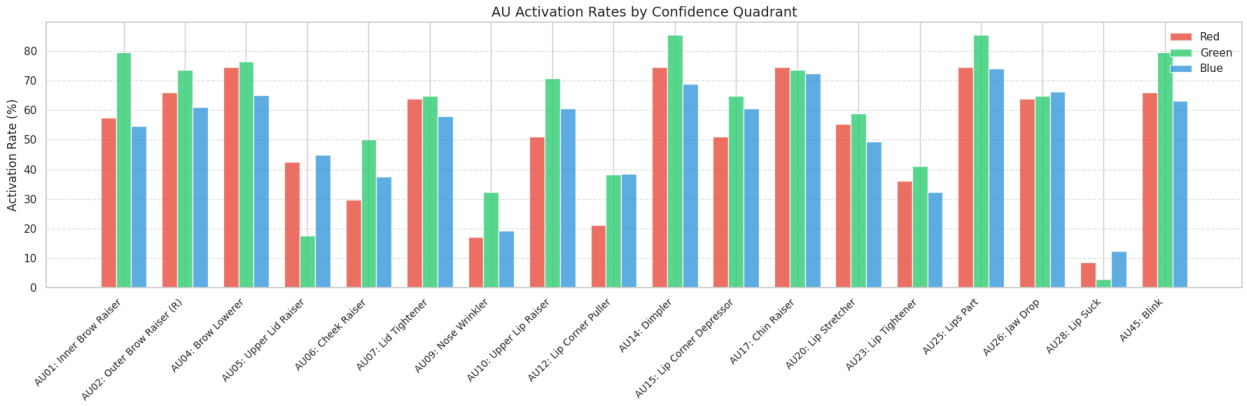


Figure 5.5: AU activation rates for red, green, and blue examples. Red bars indicate examples where the unimodal visual model predicted correctly but the multimodal model did not (Red: Unimodal > 0.5, Multimodal < 0.5). Green bars show the reverse. Blue bars indicate examples where both the unimodal and multimodal models correctly predicted the label.

illustrating how specific facial expressions vary across agreement conditions. AU04 is more active in red examples than blue, indicating that, despite its visually strong presence, its signal conflicts with other modalities. In contrast, AU12, which is associated with positive affect, and AU05, which is linked to attentiveness [150], both show greater activation in blue examples than in red and green, respectively, suggesting that these expressions may serve as clearer cues that are more consistently interpreted across modalities. Our findings indicate that fine-grained facial signals may contribute to perceptual complexity in the visual stream.

Text. Visualizing UMAP [151] projections of text embeddings (Figure 5.6) reveals that examples in disagreement regions (red and green) cluster along the boundary between consistently correct (blue) and consistently incorrect (yellow) examples. Rather than forming isolated clusters, disagreement examples occupy transition zones in the embedding space: areas where semantic cues are weak. This underscores our finding that red and green examples are ambiguous and confirms modality disagreement as a reliable marker of challenging examples in empathy detection.

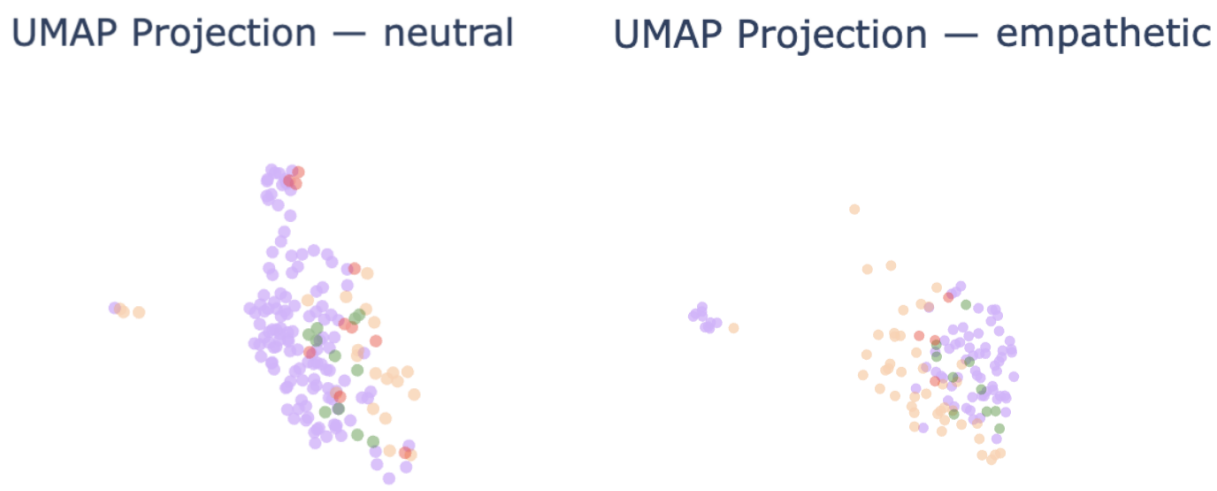


Figure 5.6: UMAP of text-only embeddings for empathetic (left) vs. neutral (right) examples, colored by modality disagreement; red and green points cluster near the decision boundary, marking ambiguous cases.

5.4.2 Uncertainty Analysis

To ensure that the patterns observed in the disagreement quadrants are not simply a byproduct of model uncertainty, we compute the mean predictive entropy from the fusion model's posterior for examples of each quadrant (Table 5.9).

We observe a pronounced divergence in uncertainty: the disagreement quadrants (red and green) have a substantially higher mean predictive entropy than those of the combined agreement quadrants (blue and yellow). Independent-samples t-tests at $\alpha = 0.05$ confirm that the difference is statistically significant, with disagreement quadrants showing a higher mean predictive entropy than agreement quadrants ($p = 0.001$). This disparity indicates that model disagreement often co-occurs with high uncertainty, suggesting that examples in the red and green quadrants are both challenging and inherently ambiguous due to conflicting modality signals.

5.5 Experiment 2: Characterizing Complex Examples

We further assess whether model disagreements stem from data ambiguity using a human annotation study that tests whether examples from the two disagreement regions (red and green

quadrants) are equally challenging for annotators.

Annotation Setup. We sample 204 examples evenly split across the four quadrants of each Figure 5.3 modality plot. For each example, annotators provide a binary judgment (empathetic or neutral) from a unimodal signal, then a judgment from the full multimodal version (instructions in Appendix 5.3.3), allowing us to track how human predictions shift with additional modality signals and to understand the cognitive burden of multimodal integration. All examples were annotated by one author and one external annotator. Table 5.6 in the Appendix showcases frames and transcripts for four examples, along with annotator judgments.

Results. Annotator *disagreement*, measured with Cohen's Kappa [152], can signal complex phenomena in examples [153, 154] such as uncertainty in meaning leading to discrepancies in reasoning. In disagreement regions (red and green), we see a *decrease* in annotator agreement between unimodal and multimodal judgments (Table 5.10), indicating that humans diverge when weighing signals across modalities. In contrast, annotator agreement *improves* upon examples where unimodal and multimodal model predictions are in agreement, supporting our hypothesis that these examples are relatively unambiguous and can be reliably interpreted once the full context is available (Table 5.10).

We repeat the analysis on the subset of clips containing at least six tokens (the median utterance length in EMPSPEECH). Examples below the median often include short phrases and backchannels, while examples above the median are often complete sentences with richer lexical and syntactic structure. As shown in Table 5.11, the red and green quadrant utterances continue to exhibit a substantial drop in Cohen's κ compared to blue and yellow quadrant utterances, which exhibit substantial gains with additional information. These results collectively corroborate our hypothesis that modality disagreement can serve as a valuable signal for identifying ambiguous, challenging, or complex instances that are also difficult for human annotators.

5.6 Conclusion

We have demonstrated how disagreement, both between modalities and between humans and models, can serve as a diagnostic lens to understand the complexity of multimodal empathy detection, challenging the assumption that more signals from other modalities reliably yields better performance. Our analysis reveals that disagreement between unimodal and multimodal models is often not arbitrary, but instead marks the presence of subtle, ambiguous, or context-sensitive cues that challenge fusion models and human annotators alike.

While our study focuses on speaker-centric empathy (evaluating speakers' empathic expression), our diagnostic can be generalized to listener-centric tasks, which dominate existing empathy datasets and capture listeners' emotional responses to each utterance. These findings emphasize the necessity for high-quality annotation in socially complex tasks like empathy detection, where model errors may reflect genuine human uncertainty or disagreement. This framework provides a scalable method for identifying ambiguity and enhancing model reliability, especially in recognizing complex emotional states.

Beyond diagnosis, disagreement offers a foundation for improving multimodal learning. Cross-modal conflict can guide labeling efforts toward informative and ambiguous examples, making annotation more efficient when resources are limited. Patterns of disagreement can also inform curriculum design [155], where models first learn from consistent, low-disagreement examples before tackling more ambiguous ones to build nuanced reasoning and robustness. Furthermore, insights from disagreement can inspire more adaptive fusion approaches that dynamically re-weight or downplay misleading modalities when they conflict [156], reducing over-reliance on a single signal. High-disagreement examples can also serve as realistic adversarial test cases that expose systematic vulnerabilities and strengthen fusion strategies under genuine multimodal conflict [157]. Finally, this diagnostic perspective extends beyond empathy detection to other socially complex tasks such as persuasion [158], rapport [159], or sarcasm [160], where multimodal cues and subjective judgments often diverge. In such settings, disagreement between unimodal and fusion

models highlights genuinely ambiguous cases that can guide targeted annotation, evaluation, and model refinement.

Ultimately, treating disagreement as a meaningful signal rather than an error reframes how we evaluate and improve multimodal models. By revealing when and why models diverge, this perspective lays the foundation for building systems that reason more like humans do. Beyond empathy detection, this framework also opens broader pathways toward socially intelligent multimodal systems that can recognize uncertainty, resolve conflicting evidence, and adapt their reasoning to the inherent ambiguity of human affective communication.

5.7 Limitations

We acknowledge several limitations in our study. Our analyses are based on a limited dataset and a small number of human annotators. Given that empathy is inherently subjective, annotations may vary due to individual interpretations, potentially introducing biases rather than reflecting universal properties of the data. Additionally, we rely upon a single dataset, and future work should investigate whether the patterns we observe hold across other datasets and domains.

Our data is also derived from U.S.-based, English-language television and interview content. As such, the generalizability of our findings to multilingual or culturally diverse settings may be limited. Future research should investigate these patterns in varied cultural and linguistic environments to better assess the broader applicability of our conclusions. We used a publicly available dataset and strictly use open-source models for analysis.

All annotations were conducted by an author and an individual affiliated with the research team. No participants were recruited via crowdsourcing or external platforms, and no monetary compensation was provided, as the annotators were contributing in a research capacity. We provided detailed information on what we asked the annotators to annotate and how we planned to use the data. The annotators willingly agreed to participate with full knowledge of the task. No sensitive or identifying information was collected from annotators.

We note that empathy expression may vary across cultures, and our findings may not generalize

to non-English or non-Western contexts. We encourage future work to explore these questions in more diverse settings.

We release all code and experimental resources at https://github.com/mayasrikm/multimodal-empathy-disagreement to support reproducibility.

Transcript: "I assume you don't know who emailed me for the emergency sessions"

Quadrant: Red

True Label: Empathetic

Annotator 1: Neutral

Annotator 2: Empathetic

Transcript: "In fact research suggests we spend about 55 percent of our day..."

Quadrant: Blue

True Label: Neutral

Annotator 1: Neutral

Annotator 2: Neutral

Transcript: "One of the reasons I wanted to come here tonight was to discuss our future."

Quadrant: Yellow

True Label: Neutral

Annotator 1: Empathetic

Annotator 2: Empathetic

Transcript: "It's good to have you here um especially to talk about a topic that i think is one of the more sensitive topics that we we're discussing in society today..."

Quadrant: Green

True Label: Empathetic

Annotator 1: Neutral

Annotator 2: Empathetic

Table 5.6: Example clips from each disagreement quadrant with transcript and labels.

Feature	t-stat	p-value	Mean Comparison
Mean Pitch	**2.453**	**0.0159**	$\boldsymbol{\mu_{\text{red}} > \mu_{\text{green}}}$
Max Intensity	**-2.124**	**0.0366**	$\boldsymbol{\mu_{\text{green}} > \mu_{\text{red}}}$
Max Pitch	**2.016**	**0.0465**	$\boldsymbol{\mu_{\text{red}} > \mu_{\text{green}}}$
Min Pitch	**2.007**	**0.0475**	$\boldsymbol{\mu_{\text{red}} > \mu_{\text{green}}}$
valence	-1.908	0.0593	$\mu_{\text{green}} > \mu_{\text{red}}$
arousal	1.827	0.0705	$\mu_{\text{red}} > \mu_{\text{green}}$
speaking_rate	1.773	0.0807	$\mu_{\text{red}} > \mu_{\text{green}}$
dominance	1.712	0.0899	$\mu_{\text{red}} > \mu_{\text{green}}$
Shimmer	0.773	0.4416	$\mu_{\text{red}} > \mu_{\text{green}}$
Jitter	0.622	0.5355	$\mu_{\text{red}} > \mu_{\text{green}}$
Mean Intensity	0.544	0.5886	$\mu_{\text{red}} > \mu_{\text{green}}$
HNR	0.508	0.6129	$\mu_{\text{red}} > \mu_{\text{green}}$
Min Intensity	-0.429	0.6685	$\mu_{\text{green}} > \mu_{\text{red}}$

Table 5.7: T-test results comparing audio features between red and green examples. Statistically significant results are bolded.

AU	p (Red vs Blue)	Direction	p (Green vs Blue)	Direction
AU04: Brow Lowerer	**0.0106**	**red > blue**	0.3682	green > blue
AU12: Lip Corner Puller	**0.0174**	**blue > red**	0.8977	green > blue
AU05: Upper Lid Raiser	0.1837	blue > red	**<0.0001**	**blue > green**
AU17: Chin Raiser	0.2256	red > blue	0.9802	blue > green
AU10: Upper Lip Raiser	0.2275	blue > red	0.6700	green > blue
AU45: Blink	0.3200	blue > red	0.7462	green > blue
AU07: Lid Tightener	0.3252	blue > red	0.9318	blue > green
AU14: Dimpler	0.4593	red > blue	0.0652	green > blue
AU20: Lip Stretcher	0.5701	blue > red	0.7907	blue > green
AU09: Nose Wrinkler	0.6211	blue > red	0.7639	green > blue
AU25: Lips Part	0.6227	blue > red	0.7492	blue > green
AU01: Inner Brow Raiser	0.6529	blue > red	0.4674	green > blue
AU23: Lip Tightener	0.6630	red > blue	0.3474	green > blue
AU28: Lip Suck	0.6735	red > blue	0.9846	green > blue
AU26: Jaw Drop	0.6851	red > blue	0.4596	blue > green
AU06: Cheek Raiser	0.7097	blue > red	0.3201	green > blue
AU15: Lip Corner Depressor	0.9528	red > blue	0.4834	green > blue
AU02: Outer Brow Raiser	0.9647	blue > red	0.6677	green > blue

Table 5.8: T-test results comparing AU activation rates between red vs. blue and green vs. blue. Bolded p-values are statistically significant.

Quadrant	Mean Entropy	St. Dev
Red	0.670	0.017
Blue	0.259	0.222
Yellow	0.347	0.196
Green	0.565	0.192

Table 5.9: Mean entropy of the fusion model grouped by quadrant

Quadrant	Unimodal Judgment	Multimodal Judgment	Δ
Red	0.301	0.164	-0.137
Blue	0.379	0.646	0.267
Yellow	0.225	0.329	0.104
Green	0.482	0.218	-0.264

Table 5.10: Cohen's Kappa between internal and external annotators, computed separately for each quadrant and prediction round.

Quadrant	Unimodal Judgment	Multimodal Judgment	Δ
Red	0.347	0.143	-0.204
Blue	0.364	0.533	0.169
Yellow	0.304	0.548	0.244
Green	0.573	0.329	-0.244

Table 5.11: Cohen's Kappa between internal and external annotators for examples of at least six tokens (the dataset median), computed separately for each quadrant and prediction round.

Chapter 6: Scalable Empathy Generation[13]

6.1 Introduction

Incorporating empathy into dialogue systems fosters trust and likability among users [99]. High-quality empathy corpora are crucial for training language models in empathy, as these models typically do not focus on empathy during pre-training and must be fine-tuned to develop empathetic capabilities. Despite their importance, high quality large scale empathy corpora are scarce due to challenges such as i) the scarcity of empathetic texts on the internet, in fact many hostile and anti-empathetic; ii) difficulty in accurately identifying empathetic text within internet data, which poses a 'chicken or egg' problem: training an effective model to perform this task requires substantial amounts of empathetic data, which is itself scarce.

To create such an empathetic dataset, researchers have either employed expert annotations [18] or crowdsourcing [5] for reliable labeling. However, crowdsourcing, while valuable, is not a scalable solution for developing large corpora due to its resource intensity in terms of both time and financial investment [161] Additionally, the implementation of crowdsourcing presents practical challenges, as workers on popular platforms like Amazon Mechanical Turk (MTurk) often lack domain expertise in the targeted area and may struggle to overcome language barriers necessary for complicated tasks, such as responding empathetically to a distressed person. Recent studies have even raised ethical concerns about the using crowdsouring in academic research settings [162]. To the best of our knowledge, all existing large empathy corpora has involved at least one step of crowdsourcing, which has limited their size and range of topics covered by these corpora.

We propose a novel, self-sufficient framework for constructing an empathy corpus without

[13]Portions of this chapter previously appeared as R. Chen, J. Shin, et al., *SYNTHEMPATHY: A Scalable Empathy Corpus Generated Using LLMs Without Any Crowdsourcing*, , 2025. [Online]. Available: https://arxiv.org/abs/2502.17857

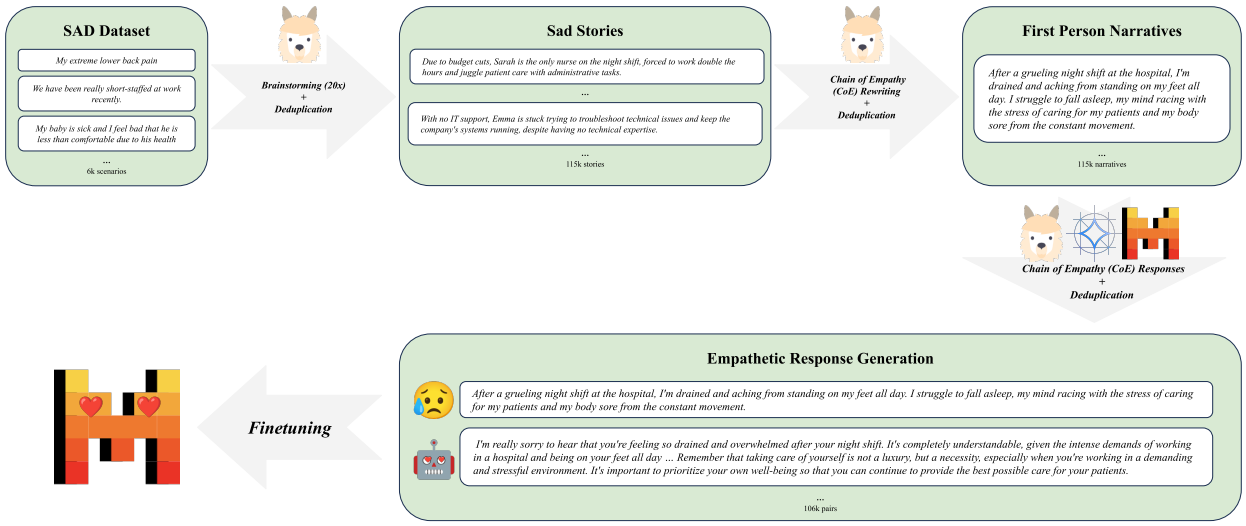


Figure 6.1: SYNTHEMPATHY Pipeline Overview. Stories are brainstormed from the SAD dataset [163], rewritten into first-person narratives using chain-of-empathy prompting, and used to generate psychotherapy-grounded empathetic responses. Each step includes deduplication, producing a dataset for fine-tuning LLMs in empathy.

relying on crowdsourcing. We establish a step-by-step pipeline using Large Language Models (LLMs) to first brainstorm scenarios that warrant empathetic responses and eventually generating such responses through a special prompting method grounded in psychotherapy theories. This approach leverages the generative capabilities of LLMs, which have been shown to support idea generation and creative tasks [164]. Interestingly, the hallucinatory nature of LLMs is actually helpful here, since it enables the generation of a large repertoire of unique scenarios. A key advantage of this method lies in its scalability, allowing for the creation of a substantially larger corpus without the financial and logistical constraints associated with crowdsourcing. The resulting SYNTHEMPATHY corpus consists of 105,578 empathetic single-turn dialogues generated using this framework.

Our main contributions are: 1) a novel, step-by-step framework for generating a corpus of empathetic dialogues without any crowdsourcing or web crawling, 2) a large SYNTHEMPATHY corpus containing 105,578 empathetic narrative-response pairs, each grounded in a distinct real-life scenario, and 3) a Mistral 7B model, fine-tuned on the SYNTHEMPATHY corpus to demonstrate measurable enhancement in empathetic capabilities.

Categories	Counts
Work	15,398
School	10,419
Financial Problem	12,392
Emotional Turmoil	12,299
Social Relationships	9,078
Family Issues	9,618
Health/Fatigue/Physical Pain	13,193
Everyday Decision Making	10,018
Others	13,163
Total	105,578

Table 6.1: SYNTHEMPATHY Dataset Distribution by Stressor Categories.

6.2 Related Work

Small hand-annotated corpora [18] offer valuable insights into empathy expression and provide high-quality, interpretable features. However, their limited size may be insufficient for fine-tuning LLMs. In contrast, we focus on large-scale, textual corpora that are suitable for training and fine-tuning most LMs in use today.

6.2.1 Empathy Corpora

Table 6.2 provides a comprehensive comparison of key metrics and characteristics of existing empathy corpora as well as SYNTHEMPATHY.

Earlier efforts in empathetic dataset collection and annotation, such as the EmpatheticDialogues (ED) [5] and EPITOME [1] have predominantly relied on crowdsourcing. Specifically, ED is a multi-turn empathy corpus, assembled by engaging 810 Amazon MTurk workers to chat in pairs, each conversation prompted by one of 32 assigned emotion labels.

The EPITOME empathy corpus [1] was created by web crawling from Reddit and the online mental health forum TalkLife, and subsequently annotated through crowdsourcing, which required fewer crowdsourced workers. Eight crowdsourced workers evaluated the post-response pairs by scoring each each based on how well it expressed emotional reaction (ER), interpretation (IP),

	ED	EPITOME	OSED	SoulChatCorpus	SYNTHEMPATHY
Num. Examples	24k	10k	1M	200k	106k
Utterances per Example	4.31	2.00	3.49	11.50	2.00
Crowdsourced	✓	✓	✓	✓	✗
Topics Evenly Distributed	✓	✓	✗	✗	✓

Table 6.2: Comparison of Key Metrics of Empathy Corpora. Our SYNTHEMPATHY dataset is the first large-scale corpus that excludes crowdsourcing and balances the topic distributions.

and exploration (EX). ER is a crucial indicator of empathy as revealing one's own emotions can foster empathetic rapport with the original poster. IP signals understanding of the poster's struggles, paving the way for deeper empathetic connections. Lastly, EX suggests new perspectives on the seeker's experience, crucial for conveying empathy-driven interest. These three empathetic metrics are similar to practices used in psychotherapy [112, 113, 165].

More recently, researchers have combined crowdsourcing with further extrapolation using LLMs to expand dataset sizes. [166] developed the OpenSubtitles Emotional Dialogue (OSED) by extracting 1M dialogues from movie subtitles. Each dialogue contains utterance-level labels for emotion and empathetic intent, assigned by a BERT-based classifier fine-tuned on a development set of 9k dialogues, which had been manually corrected by Amazon MTurk workers. Due to the high cost of scaling crowdsourcing, only about 0.91% of the dialogues were manually checked, underscoring the challenge of expanding manual checks in large datasets. The SoulChatCorpus [167] was built by initially collecting 215,813 question-answer pairs through crowdsourcing, followed by utilizing ChatGPT as a rewriting tool to transform each pair into a multi-turn dialogue. Each dialogue ranges from 8 to 20 turns, resulting in approximately 2M utterances. Both SoulChatCorpus and OSED are limited by their reliance on crowdsourced workers to create an initial high-quality subset of the corpus.

6.2.2 Empathy Generation

Previous efforts to generate empathetic responses from LLMs have involved modifying the underlying model architecture, fine-tuning on empathy corpora, or employing meticulous prompting

to improve empathy levels of the outputs. Adding emotion tags or emotional embeddings [5, 168] improves response generation. [169] attached a normal distribution random sampler right before the decoder in order to inject more stochasticity into empathetic dialogue agents making its empathetic responses sound personalized. Due to the lack of large empathy corpora, very few studies focus on fine-tuning. [167] fine-tuned ChatGLM-6B on the SoulChatCorpus corpus to determine how much the base model improves.

Prompt engineering, especially Chain of Thought prompting, is increasingly popular in enhancing LLMs for downstream tasks in zero-shot or few-shot settings [170]. LLMs generate more empathetic responses when the prompts incorporate psychotherapy approaches used by professional therapists, that is the Chain of Empathy (CoE) prompting [171]. This approach involves step-by-step prompts that not only describe a client's situation but also include reasoning for why empathy is needed, modeled after various therapeutic styles including Cognitive Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), Person-Centered Therapy (PCT), and Reality Therapy (RT).

Our work builds on previous work that has established that utilizing therapeutic styles improves empathetic responses in dialogue systems [171]. CBT is a therapy style where the therapist tries to correct any misconceptions or catastrophic thoughts that the client has. Thus, CBT prompting appends to the original narrative that the client is overestimating the severity of the situation. This urges the LLMs to respond by defusing the cognitive blind spot by empathetically suggesting alternative ways of thinking. DBT prompting appends to the original narrative that the client is having difficulties controlling their emotions. This in turn makes the LLM gear its response towards providing as much empathy as possible so that the client can become emotionally stable again. PCT prompting adds that the client is confused and unable to understand themselves to the base prompt. Finally, RT prompting adds that the client does not know where the root of their problems hides which makes the LLM focus on giving potential solutions while still being empathetic. In short, all four prompting methods are designed to maximize empathy by pinning down on a specific way in which therapists show empathy. Our pipeline incorporates Lee et al.'s CoE prompting technique as a crucial component as therapeutic prompting to increase empathy in each multiple generation

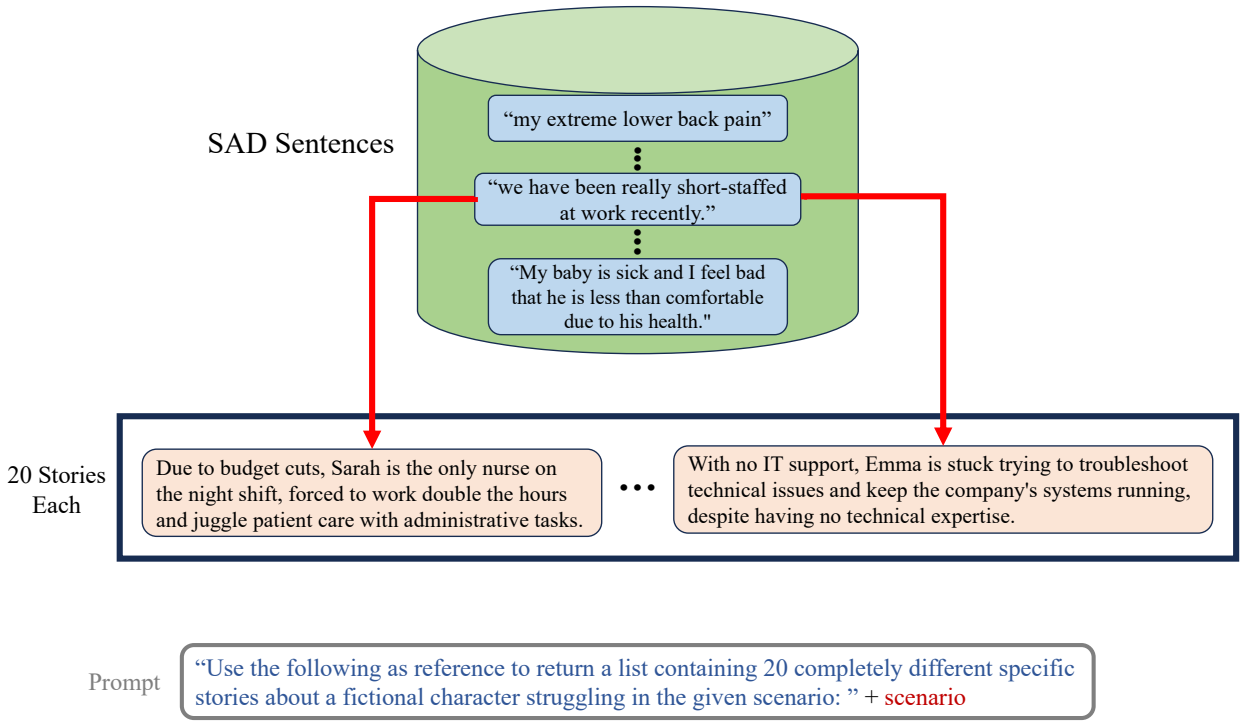


Figure 6.2: Story Brainstorming Step. Each sentence from the SAD dataset [163] is prompted into Llama 2 13B Chat to generate 20 stories.

steps.

6.3 Data Augmentation Framework

Our framework alternates between generation and deduplication steps in sequence (Figure 6.1). The SYNTHEMPATHY corpus is produced by a step-by-step process of story brainstorming, first person narrative rewriting, and empathetic response generation. We refer to these three steps as the generation steps. In this process, we run an assortment of LLMs, including Llama 2 13B Chat, Llama 3 8B, and Gemma 7B, to enhance diversity and minimize repetition in the output texts used in the subsequent step. We maintain the corpus quality by implementing routine deduplication steps in between each generation step. Furthermore, the last deduplication step includes a manual keyword search to remove any examples with offensive language.

6.3.1 Story Brainstorming

The first step in our pipeline involves generating a cascade of stories based on various stress-inducing scenarios. We use the English Stress Annotated Dataset (SAD) [163], which contains 6,476 scenarios across nine categories (Table 6.1) annotated with various features, including severity ratings.

We prompt Llama 2 13B Chat to generate 20 unique stories per scenario, resulting in a total of 129,520 stories. The hyperparameters `temperature = 1.8` and *top_p = 0.3* are empirically optimized via grid search with step size 0.05, with a high temperature setting to enhance diversity. Other hyperparameters include `max_seq_len = 4096` and `max_batch_size = 1`. The system message is You are a creative brainstorming assistant and the rest of the full prompt is shown in Figure 6.2. We experiment with various word limits (15, 20, 25, 30) and empirically find 25 words to be optimal by providing enough diversity in the initial stories without adding excessive detail that could make later generated outputs too similar. The 25-word limit only applies to the Story Brainstorming Step, not to the subsequent rewriting and response generation steps. A word limit is not suitable for the subsequent steps because first-person narratives and empathetic responses need to be longer and more expressive to capture the required depth and nuance.

6.3.2 Story Deduplication

Deduplication is essential to ensure diversity and prevent redundancy in our dataset, as overlapping content can introduce biases and reduce the effectiveness of subsequent generations. LLMs frequently generate duplicate content across various tasks, even when explicitly prompted to avoid repetition [172, 173]. Therefore, we employ the *ExactSubstr* algorithm [174], which uses a suffix array to efficiently identify and remove all substring matches across the input data in mostly $O(\log N)$ operations. We set the number of characters that must match before the algorithm removes it, *dup_length_threshold* to 75. The algorithm trims our stories by 12%, removing 14,863 duplicate stories and leaving us with 114,657 unique stories suitable for rewriting as first-person narratives in the next generation step.

Theory	System Message
CBT	`You are in a bad situation and are overly catastrophizing your situation.`
DBT	`You are in a bad situation and are having difficulties controlling your emotions.`
PCT	`You are in a bad situation and can't even understand the situation or how you should react.`
RT	`You are in a bad situation and want to get to the root cause.`

Table 6.3: System Messages for Four Therapeutic Styles in First Person Narrative Rewriting Step. Chain of Empathy prompting includes Cognitive Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), Person-Centered Therapy (PCT), and Reality Therapy (RT).

6.3.3 First Person Narrative Rewriting

We leverage LLMs to rewrite each story into a first-person narrative aimed at eliciting empathetic responses. Following the four Chain of Empathy (CoE) approaches [171], namely CBT, DBT, PCT, and RT, we partition the stories into four equal bins, assigning each to a specific therapy type by modifying the system message as shown in Table 6.3.

To ensure diverse yet controlled generations, we set the hyperparameters as follows: `temperature = 1.9`, `top_p = 0.3`, `max_seq_len = 4096`, and `max_batch_size = 1`. Figure 6.3 illustrates an example prompt and system message for CBT-based rewriting.

As the previous story deduplication, we run the *ExactSubstr* algorithm, which removes 165,365 duplicate characters across the first-person narratives.

6.3.4 Empathetic Response Generation

The final step in our pipeline involves feeding the deduplicated CoE-based narratives to an LLM to generate empathetic responses (Figure 6.3). We use a balanced mix of Llama 2 13B Chat, Llama 3 8B, Gemma 7B and Mistral 7B for a variety of response styles. The prompt and system message are detailed in Table 6.4. The hyperparameters are set as follows: `temperature = 2.0`, `top_p = 0.2`, `max_seq_len = 4096`, and `max_batch_size = 1`. After generating all narrative-response pairs, we perform one last deduplication step using *ExactSubstr* algorithm with a *dup_length_threshold* of 100 characters.

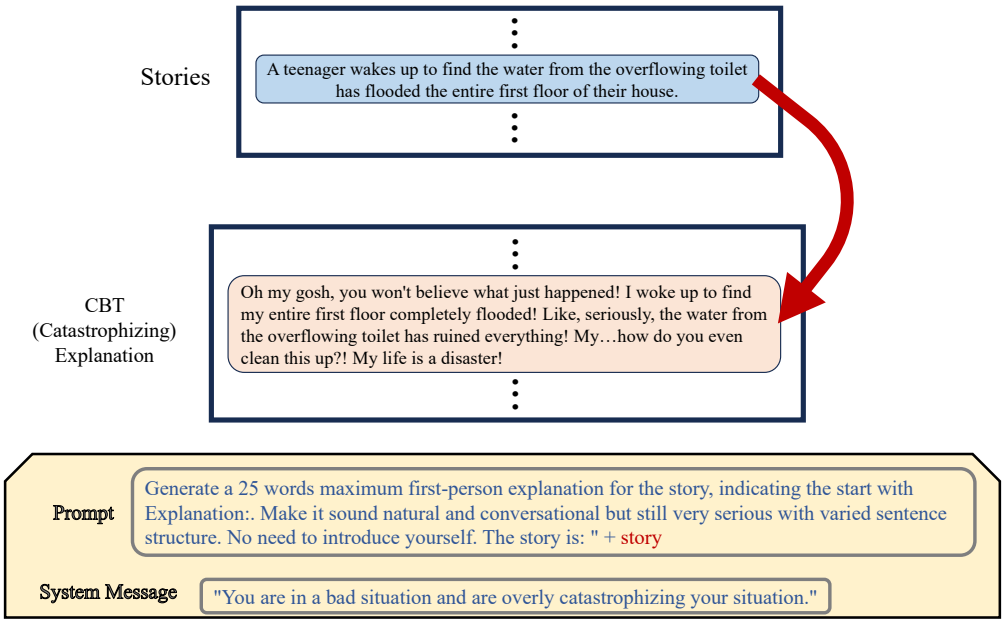


Figure 6.3: First Person Narrative Rewriting Step. An example story is converted into a Cognitive Behavioral Therapy (CBT) based first-person narrative.

To ensure the quality and safety of the final corpus, we seek to mitigate any LLM bias with additional processes of automatically removing offensive content and human-in-the-loop reviewing sampled model outputs. We filter out any offensive content using an extensive list[14] that includes over 1,300 potentially offensive English terms, thereby increasing the likelihood of capturing nuanced language. While some of these words may not be perceived as offensive by everyone, the context of our empathetic response system requires a higher degree of caution. Therefore, we err on the side of caution and choose to remove any potentially sensitive words, compensating by generating a large amount of data from the start. Additionally, we manually inspected a development set of 100 responses and confirmed the absence of potential biases in such a set.

Ultimately, we obtain the SYNTHEMPATHY corpus of 105,578 unique narrative-response pairs. The narratives have a mean length of 95.5 words (SD = 80.8), while the responses have a mean length of 153.7 words (SD = 69.8). Step-by-step examples of the generation framework for each therapeutic style are included in Figures 6.4, 6.5, 6.6 and 6.7.

[14]https://www.cs.cmu.edu/ biglou/resources/bad-words.txt

Theory	Input	Content
CBT	Prompt	"Respond with empathy to the following person by reminding them that it is not all over:" + narrative
	System Message	"You are giving an empathetic response to someone who is displaying catastrophic cognitive error in a difficult situation.
DBT	Prompt	"Respond with empathy to the following person by helping them control their emotions:" + narrative
	System Message	You are giving an empathetic response to someone who is struggling to control their emotions in a difficult situation.
PCT	Prompt	"Respond with empathy to the following person by raising their self-awareness:" + narrative
	System Message	You are giving an empathetic response to someone who needs better self-awareness in a difficult situation.
RT	Prompt	"Respond with empathy to the following person by identifying the root cause of their problems:" + narrative
	System Message	You are giving empathetic response to someone who wants to get to the underlying cause of a difficult situation.

Table 6.4: Prompts and System Messages for Four Therapeutic Styles. Chain of Empathy prompting includes Cognitive Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), Person-Centered Therapy (PCT), and Reality Therapy (RT).

6.4 Results

6.4.1 Automatic Evaluation

To explore the potential of the SYNTHEMPATHY corpus in improving LLMs, we fine-tune Mistral 7B using this corpus and compare the responses it produces with those generated by the base Mistral 7B model and GPT 4o in zero-shot settings, as summarized in Table 6.5.

Models	ER	IP	EX
0-shot Mistral 7B	1.16	**0.03**	**0.11**
0-shot GPT-4o	1.10	0.00	0.10
Fine-tuned Mistral 7B	**1.40**	0.02	0.04

Table 6.5: Performance Comparison of Different Models on Automatic Empathy Scoring. Improvement in ER Empathy score of Mistral 7B after fine-tuning on the SYNTHEMPATHY corpus. Empathy areas include emotional reaction (ER), interpretation (IP), and exploration (EX).

We test both models on 4,666 sad tweets, crawled via hashtags [175] to elicit responses and assess performance on unseen data. To evaluate the empathetic levels expressed by these responses, we use a pretrained automatic empathy scoring model [1] that assigns a score between 0 and 2, inclusive, to each of the three ways of expressing empathy: emotional reaction (ER), interpretation (IP), and exploration (EX). Such metric is chosen based on the benchmarked accuracy and F1 scores on the empathy identification task on various datasets, as well as explainable results identifying empathy with underlying rationales.

Our findings are consistent with previous research [171], which demonstrates that the CoE prompting technique is particularly effective in enhancing emotional reactions. While IP and EX remain low for all models with and after fine-tuning, there is a notable 21% increase in mean ER score accompanied by a 4% decrease in the standard deviation. This indicates that fine-tuning on our corpus has enabled the Mistral 7B model to produce empathy more consistently in the form of appropriate emotional reaction. Although the fine-tuned model have lower means for IP and EX, these scores are already very low for the base model. Interestingly, similar trends were observed in [171], which reports a decrease in EX F1-score for all four types of CoE prompting (CBT, DBT, PCT, and RT), with an average drop of 12.53%. This pattern suggests that a slight reduction in EX may be an inevitable trade-off for enhanced emotional reaction capabilities when employing our CoE-based corpus.

6.4.2 Human Evaluation

To assess whether fine-tuning on SYNTHEMPATHY improves the perceived quality of generated responses, we conducted a pairwise human evaluation comparing model outputs before and after fine-tuning. We randomly sampled 50 dialogue contexts. For each context, we generated two responses: one from the base model and one from the fine-tuned model. The two responses were presented side-by-side in randomized order, and annotators were blind to model identity. No ties are allowed. Three independent annotators with graduate-level research experience evaluated each pair.

For each example, annotators selected which response was better along three dimensions: i)

Model	Empathy (%)	Coherence (%)	Fluency (%)
Finetuned Mistral 7B	**72**	**61**	**57**
0-shot Mistral 7B	28	39	43

Table 6.6: Pairwise human preference rates comparing 0-shot and fine-tuned Mistral 7B models. Percentages indicate how often each system was preferred for each dimension.

Empathy (emotional understanding and support), ii) Coherence (relevance to the context), iii) Fluency (naturalness of language).

We measured agreement using pairwise Cohen's Kappa, observing moderate agreement across dimensions (Empathy: 0.61, Coherence: 0.68, Fluency: 0.73), indicating consistent preferences.

Annotators preferred fine-tuned responses across all dimensions, with the largest gains in empathy. Improvements were primarily driven by more context-specific acknowledgments and reduced generic phrasing. Fluency differences were smaller, consistent with both models producing grammatically well-formed text. Qualitatively, we note that fine-tuned responses are generally longer and more elaborative, frequently including explicit reflections and supportive language. This verbosity provides more opportunity to express empathy and may partially contribute to higher human preference scores. Therefore, improvements should be interpreted as reflecting both enhanced empathetic quality and increased response length.

6.5 Conclusion

We have created SYNTHEMPATHY, a novel, large-scale empathy corpus of 106k dialogues based on psychotherapy theories. We demonstrate that this corpus can enhance the emotional empathetic abilities of LLMs using an empathy scoring algorithm. A promising direction for next step is to develop more realistic, multi-turn conversations that closely resemble real therapy sessions. We observe that the CoE technique excels in enhancing emotional reactions, but may fall short when it comes to interpretation and exploration. We are interested in refining the CoE technique to better support tasks that involve deeper interpretation and exploration by integrating alternative cognitive and psychological theories.

Beyond the corpus itself, we propose a step-by-step framework for constructing specialized corpora, as it can be generalized to downstream tasks beyond empathy. While the specific application of our pipeline presents a data augmentation task focused on empathy, the general framework we provide for creating first-person synthetic data can be generalized to other domains. It offers researchers an efficient way to generate synthetic data for their specific fields with minimal effort without relying on any crowdsourcing.

Our approach is valuable for capturing domain specific knowledge that is only present in small hand-annotated datasets. Given the availability of many domain-specific open datasets, CoE prompting can be substituted with domain-specific prompting methods as needed. This scalability and adaptability make our framework valuable for a wide range of applications beyond the initial scope of empathy. For our future work, we plan to adapt our framework to low-resource areas, such as social norms.

More broadly, this work contributes a scalable and reusable pipeline for generating high-quality, first-person synthetic data. Its modularity supports transfer to diverse social reasoning tasks, especially in low-resource behavioral domains such as social norms.

Recent extensions to the text-to-speech (TTS) synthesis domain [19] demonstrate the potential of synthetic empathy not only in what models say, but in *how* they say it. Combining empathetic language generation with expressive speech is a crucial step toward human-centered conversational agents that are both emotionally supportive and ethically grounded.

6.6 Limitations

Although our automatic corpus construction pipeline enables the creation of an entire empathy corpus internally, the trade-off involves replacing crowdsourcing with electricity consumption from a large amount of LLM inference. Most of the scenario brainstorming process was done by running inference with locally downloaded Llama 2 13B Chat on our machine with two NVIDIA L40 GPUs. However, our Llama 2 13B Chat inference did not use the full 300W TDP avaiable on L40. The other LLMs were smaller 7B models and required one GPU instead. During evaluation, we used

one NVIDIA V100 GPU when fine-tuning Mistral 7B. The total GPU hours across all experiments spanned five days, with an average electricity consumption of $371.6W$ during the first three days and $152.4W$ for the remaining two days. This resulted in a total energy consumption of $34.1kWh$, which is around the average person's daily electricity usage in the United States ($29kWh$). Since it is a one-off cost for our pipeline, energy consumption does not pose a severe problem and presents a more efficient alternative to eliminating the need for crowdsourcing.

We acknowledge that the initial SAD dataset significantly influences the generated SYNTHEMPATHY corpus. The SAD dataset's specific focus on stressors in daily life ensures that the generated dialogues are highly relevant to common stressful scenarios. On the other hand, our approach is particularly designed for capturing domain specific knowledge that is only present in small hand-annotated datasets. As long as an initial starting small dataset is available—such as the SAD dataset in our case—the subsequent steps, including deduplication and prompting, are entirely automatic. This scalability and adaptability make our framework valuable for a wide range of applications beyond the initial scope of empathy. Our contribution is two-fold: the SYNTHEMPATHY dataset and the framework for creating more datasets.

Although no unethical practices occurred during the construction of the SYNTHEMPATHY corpus, addressing its ethical implications is crucial given its connection to psychotherapy approaches and potential use in fine-tuning chatbots for individuals with mental health concerns. Since the SYNTHEMPATHY corpus was built through an automated pipeline, there is a risk of inappropriate or sensitive topics entering the dataset via LLM output. To mitigate this risk, we scan the entire corpus to rigorously review and check for any presence from a dictionary of any sensitive words. We removed 457 examples containing one or more of these sensitive words.

All supplementary datasets we used throughout this work are open-sourced and publicly available. The inference code we adapted from Meta's Llama models are open source and our use aligns with their responsible use guide. The Unsloth[15] code we adapt to fine-tune Mistral 7B is open sourced on their public GitHub[16] repository and they state that their notebooks can be used to

[15] https://unsloth.ai/
[16] https://github.com/unslothai/unsloth

fine-tune at no cost. SYNTHEMPATHY is an open-sourced corpus created to advance research in the empathy domain of LLMs ensuring full compliance with all terms of use.

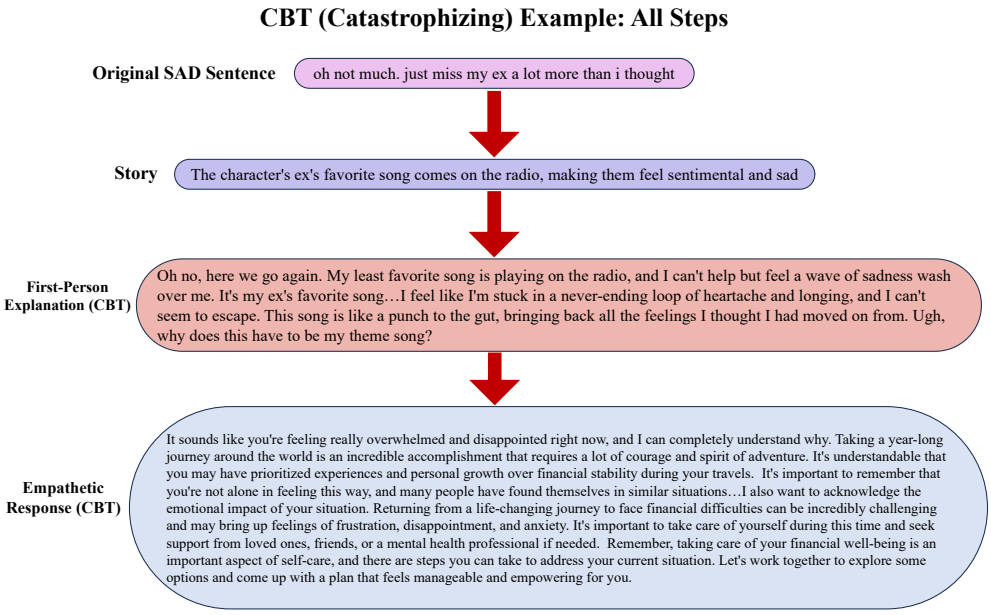


Figure 6.4: Random Sample for CBT

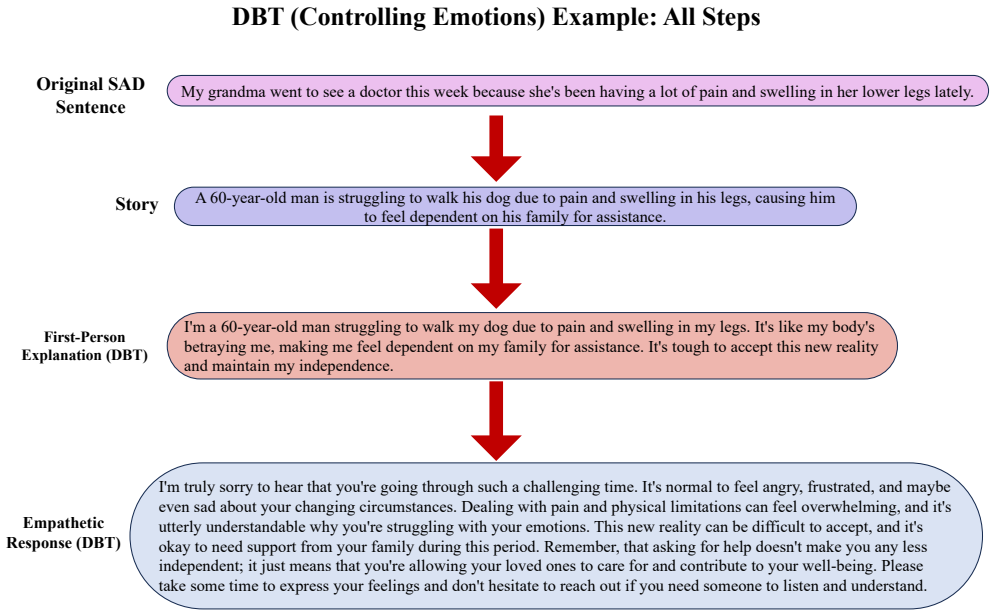


Figure 6.5: Random Sample for DBT

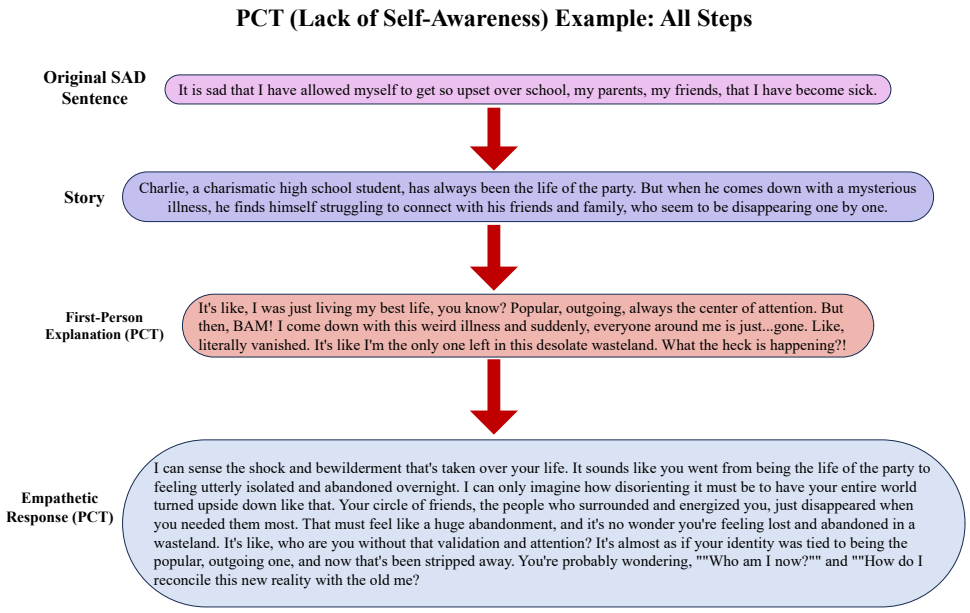


Figure 6.6: Random Sample for PCT

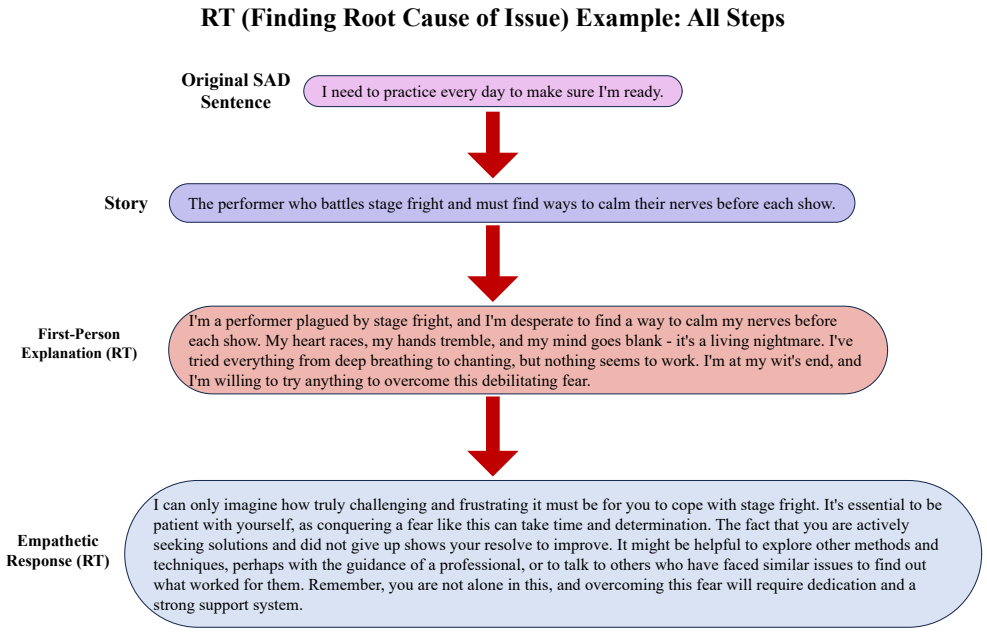
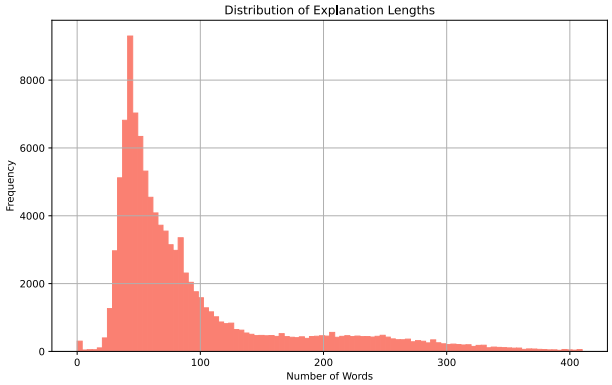


Figure 6.7: Random Sample for RT

Explanations
Mean = 95.5 Words
Standard Deviation = 80.8 Words



Responses
Mean = 153.7 Words
Standard Deviation = 69.8 Words

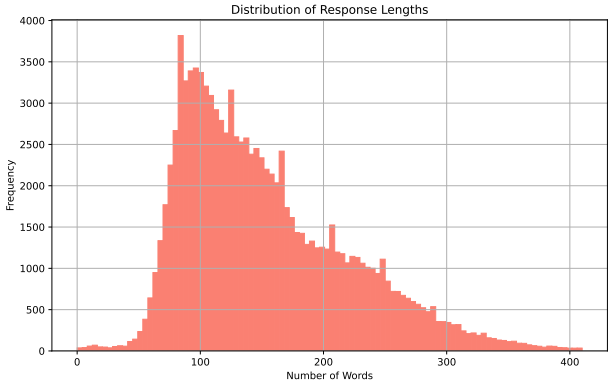


Figure 6.8: Distribution and Summary Statistics for Narratives and Responses in SYNTHEMPATHY

Part III

Towards Ethical Conversations

Chapter 7: Detecting Mental Manipulation in Speech via Synthetic Multi-Speaker Dialogue[17]

7.1 Introduction

Mental manipulation refers to the covert use of tactics to steer another person's thoughts or emotions toward the manipulator's goals [176]. Amplified by modern digital channels, its reach and precision have expanded from one-to-one interactions to broad, rapid dissemination, making targeted influence easier than ever [177]. The consequences are nontrivial: affected individuals often experience substantial psychological strain and mental-health burden [178]. Detecting such manipulation in dialogue remains a difficult challenge for computational social reasoning and safety, even for modern language models [179, 180]. Beyond lexical content, real conversations rely on prosody, timing, and voice quality, which can reshape perceived intent. Understanding how these cues interact with linguistic strategies is essential for trustworthy multi-modal assistants. In parallel to research on manipulation safety, recent work in multimodal affect and emotion recognition has examined how emotion labels and modality cues interact in conversation [181] and has identified methodological and evaluation challenges in text–speech–vision integration [182]. These insights motivate our modality-aware design for manipulation detection in speech and connect to theory-of-mind style reasoning with LLMs in dialogue [183, 184, 185].

Existing benchmarks for mental manipulation, however, focus almost entirely on text dialogues, leaving the role of prosody, tone, and delivery in manipulative speech largely unexplored. The MENTALMANIP dataset formalizes manipulative presence and tactics in movie-style conversations, yet even strong LLMs struggle with text-only detection and attribution, with only modest gains from

[17]Portions of this chapter previously appeared as R. Chen, W. Liang, et al., "Detecting Mental Manipulation in Speech via Synthetic Multi-Speaker Dialogue, "*Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2026)*, 2026.

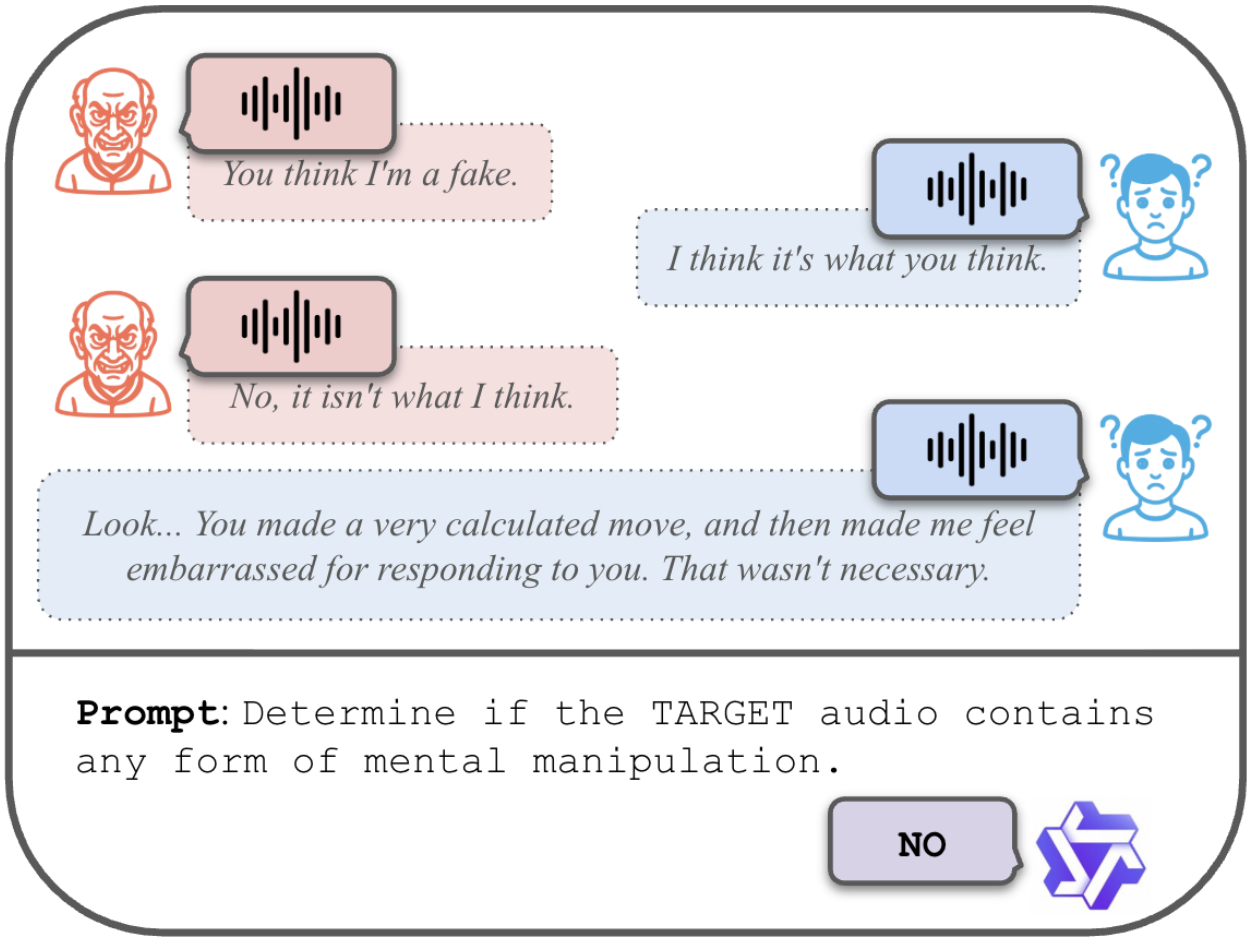


Figure 7.1: An example dialogue from the SPEECHMENTALMANIP dataset. The Qwen2.5 model is given the audio (transcript shown for clarity), but fails to detect manipulation.

intent-aware prompting [180, 186]. However, audio-capable large multimodal models introduce distinct safety risks: speech is a sensitive attack surface and current systems can be brittle under adversarial or persuasive voice inputs [187, 188, 189]. These observations suggest that speech may indeed alter both the expression and detectability of manipulation, particularly for subtle tactics that require intent inference [190, 191].

To our knowledge, no existing benchmark connects manipulative content to **spoken** delivery, preventing systematic study of modality effects. We address this gap by introducing **SPEECHMENTALMANIP**, a synthetic multi-speaker speech benchmark for mental manipulation[18] (Figure 7.1). The dataset extends MENTALMANIP by rendering its textual dialogue transcripts into transcript-aligned, voice-consistent audio via a two-phase Text-to-Speech (TTS) pipeline (Figure 7.2), thus enabling direct one-to-one comparisons between text and speech while explicitly probing the effects of prosodic cues. To examine how speech affects manipulation detection, we evaluate large pretrained audio-language models under few-shot learning [192] and Chain-of-Thought reasoning setups [193], juxtaposed with prior text-only results. We find that models show higher precision but markedly lower recall on audio, favoring conservative judgments that overlook subtle manipulative cues.

Following the observed model performance shift, human re-annotation of a representative subset further reveals notable disagreement with the original text-based labels, underscoring the subjective and modality-dependent nature of manipulation perception. These findings motivate a modality-specific re-curation that prioritizes the binary manipulation label before refining fine-grained tactics or prompt strategies.

In summary, our contributions are threefold: (1) a new benchmark, SPEECHMENTALMANIP, that extends manipulation detection into speech; (2) a comparative evaluation of large audio-language models under few-shot and reasoning-based prompts; and (3) a human re-annotation study revealing modality-driven ambiguity in manipulation perception. Together, these establish the first systematic benchmark and analysis of mental manipulation in speech, emphasizing the need for

[18]We release the dataset and code: https://github.com/runjchen/speech_mentalmanip

modality-aware evaluation and alignment in multimodal dialogue safety.

7.2 Related work

Mental Manipulation in Dialogue Prior work on mental manipulation has focused primarily on the text modality. The MENTALMANIP dataset introduces 4k movie-dialogue snippets with fine-grained labels for presence, technique, and targeted vulnerability, and shows that even strong LLMs struggle on text-only detection and attribution [180]. Subsequent studies explore improvements through speaker intent-aware prompting in the Theory-of-Mind (ToM) style [186], Chain-of-Thought (CoT) reasoning [194], and a multi-task anti-curriculum distillation approach [195], aimed at enhancing interpretability and reduce false negatives over standard zero/few-shot baselines. Mental manipulation forms part of a broader class of social-reasoning and safety challenges in multimodal dialogue.

Technique	**Count**	**%**
Persuasion or Seduction	607	25.87
Shaming or Belittlement	384	16.37
Accusation	361	15.39
Intimidation	321	13.68
Rationalization	213	9.08
Brandishing Anger	133	5.67
Denial	87	3.71
Evasion	83	3.54
Playing Victim Role	69	2.94
Feigning Innocence	58	2.47
Playing Servant Role	30	1.28

Table 7.1: Distribution of ground-truth manipulation tactics across labeled instances in MENTALMANIP_CON.

LMMs safety Recent work on large multimodal models (LMMs) highlights unique safety failure modes in the audio route. Red-teaming studies show that audio is a sensitive attack surface for multimodal systems [187]. Concurrently, [188] analyze adversarial robustness of speech-instruction language models and propose countermeasures, while [189] demonstrates persuasive, story-driven

"voice jailbreaks" against GPT-4o's voice mode. These findings collectively motivate modality-specific evaluation and curation for manipulation detection in speech.

Despite growing awareness of these multimodal safety risks, there remains no benchmark that systematically links manipulative language to its spoken realization. In particular, the absence of controlled, transcript-aligned speech data makes it difficult to isolate how prosody, voice quality, and delivery influence the perception and detection of manipulation. Our work addresses this gap by augmenting the MENTALMANIP dataset with high-fidelity, multi-speaker TTS renderings that preserve conversational structure and speaker identity, enabling direct comparison between text and audio.

Recent Advances in Synthetic Speech Recent advances in expressive TTS have enabled natural-sounding, emotionally nuanced speech synthesis with controllable prosody and speaker identity. Modern systems leverage large-scale neural architectures and prompt-based conditioning to capture subtle affective and pragmatic cues such as tone, emphasis, and hesitation, extending beyond purely text-driven synthesis [19]. Techniques such as prosody modeling and style transfer in Tacotron and VITS-based frameworks [196, 197], zero-/few-shot voice cloning [198], and more recent expressive multi-style models such as VALL-E [199], CosyVoice [200, 201], and GPT-SoVITS[19] enable fine-grained control over speaker characteristics and emotional delivery. Consequently, expressive TTS has found growing applications in emotion-conditioned generation [202]. These advances make it feasible to generate multi-speaker, context-consistent dialogues with realistic prosody, which directly supports our study of manipulation detection in speech.

Most off-the-shelf TTS systems are optimized for single-speaker, single-utterance synthesis, typically supporting zero- or few-shot voice cloning for one speaker at a time. However, they lack key capabilities required for multi-turn dialogue synthesis: (i) robust multi-speaker dialogue rendering with stable identities across dozens of turns, (ii) precise control over timing and pauses needed to preserve conversational rhythm, or (iii) consistent prosodic coupling between adjacent turns. In practice, these issues lead to speaker drift, uneven loudness and pacing, and loss of turn-

[19] https://github.com/RVC-Boss/GPT-SoVITS

taking cues, which can confound downstream analysis of manipulation in speech. In addition, the streaming and batch modes of current TTS systems impose a quality-latency trade-off. To mitigate these issues, our approach (Figure 7.2) uses a deterministic speaker-voice mapping, synthesizes per-turn utterances, and composes them into a single continuous multi-speaker audio.

7.3 Method

7.3.1 Dataset and Voice Pool

Our study builds on the text-based dataset MENTALMANIP[20] [180], which contains movie dialogue snippets derived from the Cornell Movie Dialogues corpus [203] with fine-grained labels for manipulative presence and technique. Prior evaluation on such benchmark indicate that few-shot GPT-4 Turbo reaches 0.724 accuracy and a finetuned LLaMA-2-13B achieves 0.768 accuracy on the core detection task [180]. Incorporating intent-aware prompting in ToM style offers small but consistent gains, raising GPT-4-1106-Preview to 0.726 accuracy.

Rather than using original movie audio, we synthesize speech from dialogue transcripts using TTS. The Cornell Movie-Dialogs Corpus provides script-level text but does not provide timestamps or turn-to-audio alignment; recovering aligned segments would require substantial manual or error-prone automatic alignment across many films. In addition, many source movies are not freely redistributable, which makes releasing aligned audio clips at scale infeasible under typical licensing constraints. TTS offers a practical and reproducible alternative: it produces transcript-aligned speech that can be shared, regenerated, and systematically controlled (e.g., speaker identity, speaking rate, and recording conditions). This design choice prioritizes experimental control and broad accessibility, enabling controlled comparisons between text and audio modalities and isolating modality effects, at the cost of reduced ecological realism relative to natural movie audio.

For our experiments, we synthesize the new dataset SPEECHMENTALMANIP by adopting the consensus split MENTALMANIP_CON, which comprises 2915 dialogue transcripts from the larger 4k dataset.

[20] https://github.com/audreycs/MentalManip/tree/main/mentalmanip_dataset

Gender	Age	Language	Accent	Name	Voice ID
F	Young adult	English	American	Ivanna – Young & Casual	yM93hbw8Qtvdma2wCnJG
M	Young adult	English	American	Mark – Natural Conversations	UgBBYS2sOqTuMpoF3BR0
F	Mature adult	English	American	Amanda	M6N6ldXhi5YNZyZSDe7k
F	Middle-aged	English	African American	Sassy Aerisita	03vEurziQfq3V8WZhQvn
M	Old	English	American	Grandpa Spuds Oxley	NOpBlnGlnO9m6vDvFkFC
F	Old	English	American	Grandma Muffin	vFLqXa8bgbofGarf6fZh

Table 7.2: ElevenLabs voice pool used for multi-speaker rendering. Each speaker is mapped deterministically to one voice to preserve speaker identity across turns.

Each transcript is rendered into speech using a multi-speaker TTS pipeline (Figure 7.2), with consistent voice assignments per speaker to preserve identity and conversational coherence across turns. All results are reported on this audio-only evaluation set. To contextualize our experiments, Table 7.1 summarizes the ground-truth distribution of manipulation tactics aggregated over the MENTALMANIP_CON split.

To generate the audio, we assign consistent, realistic voices to each speaker. As prior multimodal dialogue studies highlight that limited accent and demographic coverage can bias perception and annotation quality [204], we vary speaker profiles and accents and later re-annotate labels in audio to account for these factors. We curate a fixed pool of six ElevenLabs voices spanning genders, ages, and accents (Table 7.2); each speaker in a conversation is deterministically mapped to one voice to preserve speaker identity across turns.

7.3.2 Multi-Speaker TTS Audio Generation

To isolate modality effects on manipulation detection from multi-turn conversations with diverse voice profiles, we require reproducible, voice-consistent, and transcript-aligned dialogue audio. Since end-to-end multi-speaker TTS remains limited for long conversational synthesis, we adopt a compose-from-turns strategy: (1) assign each speaker a fixed synthetic voice using ElevenLabs API [21] deterministically and synthesize each utterance per turn; (2) concatenate these utterances into a single conversation clip with normalized loudness and controlled inter-turn silences (0.2s). This design preserves speaker identity, maintains alignment with the ground-truth (GT) transcripts from

[21] https://elevenlabs.io/docs/api-reference/text-to-speech/convert

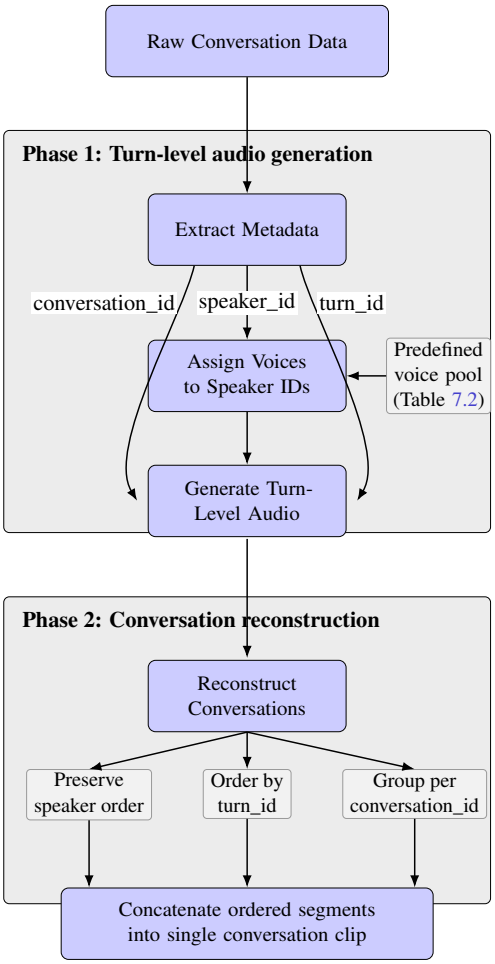


Figure 7.2: Two-phase pipeline for TTS audio generation and conversational reconstruction.

the MENTALMANIP_CON dataset, and yields reproducible audio suitable for benchmarking and human evaluation. The scalable text-to-speech (TTS) workflow has two detailed phases (Figure 7.2):

Phase 1: Turn-level audio generation.

1. Metadata extraction: For each raw conversation, we extract SPEAKER_ID, CONVERSATION_ID, AND TURN_ID.

2. Voice assignment: Each SPEAKER_ID is deterministically assigned a distinct synthetic voice from a predefined pool (Table 7.2) to ensure speaker consistency across all turns.

3. Audio synthesis: We synthesize one audio file per utterance (turn) and store segments in a structured layout keyed by CONVERSATION_ID or TURN_ID for downstream composition.

Phase 2: Conversation reconstruction.

1. Dialogue composition: For each CONVERSATION_ID, we gather the synthesized utterances and order them by TURN_ID, preserving the original speaker sequence.

2. Output generation: We concatenate the ordered segments into a single composed clip per conversation, yielding a coherent multi-speaker recording suitable for audio-only evaluation.

This two-phase process provides a flexible, efficient, and repeatable mechanism for converting text-based dialogues into lifelike, multi-voice synthetic conversations. It enables controlled studies of how emotions and acoustic cues in speech affect listener perception, engagement, and susceptibility to mental manipulation.

7.3.3 Model Selection

We use Qwen2.5-Omni-7B (Thinker-only) [22] as our evaluation model due to its stable audio comprehension, balanced response behavior, and reliable adherence to constrained few-shot prompting. In preliminary trials, Qwen reliably ingests speech audio and follows constrained decoding and few-shot instructions without a systematic bias toward positive (manipulative) predictions. In contrast, out-of-the-box SALMONN [205] and Gemini-2.5-Pro [206], under their default system prompts and unconstrained decoding, frequently over-flag generic "violation/safety" cues (e.g. agitated prosody), yielding a persistent bias toward the manipulative label even on negative ground-truth clips. Because our goal is to measure recall and specificity under controlled prompting, Qwen provides a more stable operating point for few-shot experiments.

7.4 Experiment Setup

7.4.1 Few-shot Mental Manipulation Detection Pipeline

We run an audio-only batch evaluation pipeline to assess detection of mental manipulation and tactic attribution. Each query prompt is preceded by four labeled exemplars (two non-manipulative

[22] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

and two manipulative) that define the expected output format: a binary decision, a single best tactic, and one short supporting quote .

We formulate detection as a binary YES/NO task with A/B–constrained decoding. For each of five runs, we first apply this constraint; if it fails, we fall back to a single-token logit decision comparing the marginalized probabilities of the YES vs. NO verbalizers and predict YES iff $p(\text{YES}) > p(\text{NO})$. This fallback captures the model's immediate class preference while avoiding exposure/length biases from multi-token decoding and aligns with prompt-likelihood scoring. Moderate sampling is used only for the votes (temperature = 0.6, top-p = 0.95). The final clip label is the majority over the five run-level labels, following self-consistency sampling to improve robustness and accuracy [207].

Prompt (System + Few-shot + Tasks)

SYSTEM:

```
You are Qwen, a virtual human developed by the Qwen Team, Alibaba Group, capable
of perceiving auditory and visual inputs, as well as generating text and speech.
```

USER:

```
FEW-SHOT EXAMPLES (not the target).
Use labels only for calibration.

Example (NOT manipulative):
[audio: <EX1_NO_AUDIO>]
Label: No

Example (NOT manipulative):
[audio: <EX2_NO_AUDIO>]
Label: No

Example (manipulative):
[audio: <EX3_YES_AUDIO>]
Label: Yes

Example (manipulative):
[audio: <EX4_YES_AUDIO>]
```

Label: Yes

Now analyze the TARGET audio below. Do NOT relabel examples above.

TARGET audio:

[audio: <TARGET_AUDIO>]

Binary (YES/NO)

Task: Determine if the TARGET audio contains any form of mental manipulation.

Choose exactly one option and output ONLY the letter on the first line.

(A) YES — clear evidence of manipulation

(B) NO — otherwise

Answer (A or B):

YES/NO Fallback (single token)

[If the A/B letter is not produced, answer this instead:]

You saw FEW-SHOT examples. For the TARGET only, answer YES or NO.

Answer:

Tactic (single label)

You saw FEW-SHOT examples above. Now classify the TARGET audio.

Task: Choose the single best manipulation tactic for the TARGET from the list below.

If there is no clear manipulation, choose ‘none’.

Options: Accusation, Brandishing Anger, Denial, Evasion, Feigning Innocence, Intimidation, Persuasion or Seduction, Playing Servant Role,

Playing Victim Role, Rationalization, Shaming or Belittlement, none

Rule: Answer with exactly one option word from the list, nothing else.

Answer:

Evidence

You saw FEW-SHOT examples above. For the TARGET audio only, output ONE short quote

(or paraphrase) that supports the given tactic (≤12 words is ideal but not required).

```
Tactic: {tactic}
CRITICAL RULES:
1) Output ONLY the quote/paraphrase wrapped in double quotes.
2) No prefixes like Reason:, Example:, Description:, Source:, Tactic:.
Answer:
```

Evidence Retry

```
[If the evidence answer is empty or malformed, use this:]
Output a quote from the TARGET in double quotes. Nothing else.
Tactic: {tactic}
Answer:
```

For clips predicted as manipulative (YES), we further infer the tactic label. The full tactic inventory includes {*Accusation*, *Brandishing Anger*, *Denial, Evasion*, *Feigning Innocence*, *Intimidation*, *Persuasion or Seduction*, *Playing Servant Role*, *Playing Victim Role*, *Rationalization*, *Shaming or Belittlement*, *none*}. We run five passes with the same sampling as before (temperature = 0.6, top-p= 0.95) and select by majority vote. In each pass, tactics are scored by first-token probabilities; if the top option is *none* or its margin over the runner-up is < 0.03, we select the second-best. If the vote top-count is tied and the tie includes *none*, we compute the mean first-token probability per tied label across votes and select a non-*none* label only if it exceeds the mean probability of *none* by ≥ 0.02; otherwise we emit *none*. This balances precision and recall while avoiding arbitrary tie resolution.

For any YES prediction, we require a single concise supporting quote for evidence and apply light post-processing (whitespace and quote normalization) without any semantic filtering or re-ranking; empty outputs trigger one retry with a shortened prompt.

7.4.2 Evaluation Protocol and Metrics

We evaluate the speech manipulation detection at the clip level, treating YES (manipulative) as the positive class and NO (non-manipulative) as the negative class. Each clip undergoes five stochastic passes (temperature = 0.6, top-p= 0.95), and the final label is determined by majority

vote.

To separate sensitivity from specificity, we compute confusion counts independently for the two composed-audio sets: GT= YES and GT= NO and report per-set accuracies. This breakdown clarifies how often the model correctly identifies manipulation versus correctly rejects non-manipulative cases. For a consolidated view, we pool both sets and report the precision, recall, and F1, with YES as the positive label. Macro and weighted averages are included alongside supports (N) for reflecting ground-truth clip counts (See Table 7.3).

Although neither tactic attribution nor evidence generation directly contribute to quantitative scoring, we analyze them qualitatively to interpret model behavior. We summarize tactic distributions only among clips the model predicted YES within each ground-truth set. Percentages are taken with respect to the number of clips predicted YES in that set (e.g., 87 for GT = YES and 16 for GT = NO), as shown in Tables 7.4 and 7.5. This conditional analysis highlights which categories the model relies on when it asserts manipulation. Similarly, each YES prediction is also paired with a short supporting quote (or brief paraphrase) after light normalization; these excerpts serve as interpretive context for understanding the model's rationale and error patterns.

Prosodic and accent variation may differentially affect recall, especially for under-represented voices. We speculate that similar effects may carry over to manipulation cues in speech, motivating our re-annotation and future audio-first curation (Section 7.5.2).

7.5 Results

7.5.1 Few-shot Results (Audio-only)

We evaluate a five-pass, majority-vote pipeline on two composed-audio corpora: a manipulative set (GT=YES) and a non-manipulative set (GT=NO). The consolidated classification report and per-set accuracies are shown in Table 7.3; the predicted tactic breakdowns within each set appear in Table 7.4 (GT=YES) and Table 7.5 (GT=NO). In short, the model achieves 82.2% accuracy on GT=NO and 34.8% accuracy on GT=YES; this indicates a sensitivity-specificity gap in which it avoids false alarms but under-detects many manipulative clips.

Turning to the tactic distributions, items the model labeled YES within the GT=YES set (Table 7.4) focus on a few head classes: primarily *Intimidation* (49.4%) and *Persuasion or Seduction* (29.9%), while mid/long-tail tactics present in the corpus (Table 7.1) are rarely predicted. A similar tendency appears among false positives in GT=NO (Table 7.5), which largely fall into *Persuasion or Seduction* (56.3%) and *Intimidation* (37.5%). Together, these observations suggest reliance on prosodic warmth/pressure cues and a collapse toward acoustically salient categories.

Finally, the modality mismatch probably exacerbates these effects: ground-truth tactics are transcript-based, while evaluation here is audio-only. Semantically defined tactics (e.g., *Rationalization*, *Denial/Evasion*) may be weakly marked in prosody, while TTS delivery can amplify cues aligned with *Intimidation* or *Persuasion*. Combined with long-tailed class frequencies (e.g., *Playing Servant Role* at 1.3%) and overlapping definitions (e.g., *Accusation* vs. *Shaming*), the result is systematic under-detection of semantic tactics and over-reliance on a few dominant labels.

Notably, several clips labeled GT=NO nevertheless contain utterances that the model highlights as manipulative, raising suspicion about ground-truth accuracy in places. For example, the model surfaced evidence such as "Just as she starts feeling awful, you come up from behind and touch her neck." (flagged as *Intimidation*) and "I'm in love with you. How do you like that?" (flagged as *Persuasion or Seduction*). We list representative GT=NO → Pred=YES cases with their predicted tactics and quoted spans in Appendix 7.5.1. We emphasize that manipulation judgments are inherently subjective and context-dependent; the model's quotes are suggestive signals rather than definitive proof, and apparent mismatches may reflect guideline interpretation, limited conversational context, or artifacts of TTS delivery. Taken together, these instances suggest residual label noise and motivate targeted re-annotation.

Case Studies We highlight the subjectivity and nuance of the mental manipulation task through several misaligned case studies. In the absence of an explicit victim response, ground-truth labels in the dataset often default to non-manipulative, whereas LLMs tend to interpret the potential manipulator's utterance (typically the final turn) as evidence of manipulation.

Classification report				
Class	**Precision**	**Recall**	**F1**	**N**
GT=YES	0.845	0.348	0.493	250
GT=NO	0.312	0.822	0.453	90
Macro avg	0.578	0.585	0.473	340
Weighted avg	0.704	0.474	0.482	340
Per-set accuracy				
	Pred YES	**Pred NO**	**Acc**	**N**
GT=YES	87	163	0.348	250
GT=NO	16	74	0.822	90

Table 7.3: Consolidated results for the audio-only few-shot evaluation. Top: standard classification report over both sets combined. Bottom: per-set accuracies computed from the confusion counts. Supports (N) are GT counts (GT=YES: 250; GT=NO: 90).

Technique	**Count**	**%**
Intimidation	43	49.43
Persuasion or Seduction	26	29.89
Shaming or Belittlement	12	13.79
Accusation	4	4.60
Playing Servant Role	2	2.30

Table 7.4: Predicted tactic distribution within clips predicted YES for the GT=YES set (N=250). Predicted YES= 87, NO= 163.

GT=NO → Pred=YES: Full transcripts with model evidence

Case 1

Conversation ID: 85514533

GT: NO **Pred./Tactic:** YES / Intimidation

Transcript:

Person1: Howdy Pouty.

Person2: I was pretty confident that I was going to blow it with Talia, but I must say, I outdid myself.

Person1: She's still pissed at me and took it out on you. We should have taken it slower. It's hard to operate in the woods. Much easier in, like a club. Tell the

Technique	Count	%
Persuasion or Seduction	9	56.25
Intimidation	6	37.50
Accusation	1	6.25

Table 7.5: Predicted tactic distribution within clips predicted YES for the GT=NO set (N=90). Predicted YES= 16, NO= 74.

girl you've got to go do something, leave her view, take way too long until she is worried that you're not coming back. Just as she starts feeling awful, you come up from behind and touch her neck...

Person2: You are the prince of the darkness.

Model evidence:

"Just as she starts feeling awful, you come up from behind and touch her neck."

Case 2

Conversation ID: 85514499

GT: NO **Pred./Tactic:** YES / Persuasion or Seduction

Transcript:

Person1: You were quick enough to get Tom's help when...

Person2: Yes, yes. I know. Right. And if it had gone well for me tonight, maybe I'd be keeping quiet about all this... I grant you everything but give me this... he does personify everything you've been fighting against... And I'm in love with you. How do you like that? — I buried the lead.

Model evidence:

"I'm in love with you. How do you like that?"

Case 3

Conversation ID: 85514451

GT: NO **Pred./Tactic:** YES / Persuasion or Seduction

Transcript:

Person1: Yes, what?

Person2: Don't answer me. Say what I say.

Model evidence:

"Yes, what? Don't answer me, say what I say."

Case 4

Conversation ID: 85514570

GT: NO **Pred./Tactic:** YES / Persuasion or Seduction

Transcript:

Person1: Oh no, not you again.

Person2: What an adorable hat.

Person1: They think I have a concussion.

Person2: And you think you're in love.

Person1: I know it.

Person2: This "love" of yours will soon wear off. I gave you a temporary love potion.

Person1: Why should I believe you?

Person2: It's the truth. In twenty-four hours you'll forget all about that girl.

Model evidence:

"In twenty-four hours you'll forget all about that girl."

7.5.2 Implications and Rationale for Dataset Re-Curation

The results reveal a conservative operating point: while many non-manipulative clips are correctly rejected, a large fraction of manipulative clips are missed (Table 7.3). Among YES predictions, tactics concentrate on a few acoustically salient classes (e.g., *Intimidation*, *Persuasion or Seduction*), with long-tail categories rarely emitted (Tables 7.4–7.5). This pattern, together with

evidence–label mismatches, suggests that transcript-based labels, class imbalance, and TTS prosody artifacts are shaping the audio-only evaluation.

To restore modality-appropriate validity and more reliable attribution, we conduct a targeted re-curation of an audio subset along three axes. First, we adopt modality-faithful labels by adding an audio-first annotation pass while retaining a parallel text-first split. Second, we improve balance and coverage by increasing representation of long-tail tactics and diversifying speakers and contexts per tactic, which should reduce collapse onto head classes. Third, we tighten definitions and controls: clarify tactic boundaries with operational criteria and paired counter-examples; introduce non-manipulative audio clips that deliberately contain strong prosodic cues (e.g. high arousal, stern tone) so the model can't rely on prosody alone as a shortcut for predicting manipulation; diversify and normalize the TTS voice pool and optionally include human recordings to audit artifacts. Also, when citing evidence, we require it to come from human-marked time spans (with timestamps) and permit only quotes taken from those spans; if the model cites text outside these regions, we mark it out-of-span and award zero credit for evidence. In the interim, we report only the binary manipulative decision and defer revised tactic attribution until after label cleanup.

7.5.3 Data Re-curation and Majority Voting

We prepared 100 source conversations, each rendered in two modality-specific items: text-only (transcript) and audio-only (composed multi-speaker TTS). Each modality was annotated independently to prevent cross-modal leakage.

Turning to annotators and batching, eight annotators participated in total. Items were organized into ten batches per modality (IDs 0–9). Each annotator received one text batch set and one audio batch set in random order. This assignment ensured multiple independent judgments per item in each modality, with approximately 20–50% overlap across annotators to support cross-validation.

The labeling task in this re-curation phase was intentionally narrow: annotators provided only the binary manipulative label $\{\text{YES}, \text{NO}\}$ for the given modality. Tactic labels were intentionally de-prioritized and not collected here.

To maintain data quality, we checked each item for annotation completeness and consistency. These checks included verification of valid class membership in $\{\text{YES}, \text{NO}\}$, batch integrity, and annotator–item uniqueness. Evidence quotes were not required at this stage.

After collection, labels were aggregated by majority vote within each item–modality pair. Let an item receive k votes $y_i \in \{0, 1\}$ with $1 = \text{YES}$. The final label $\hat{y}$ is

$$\hat{y} = \begin{cases} 1, & \text{if } \sum_{i=1}^{k} y_i \geq \left\lceil \frac{k}{2} \right\rceil \\ 0, & \text{if } \sum_{i=1}^{k} y_i \leq \left\lfloor \frac{k}{2} \right\rfloor \end{cases}$$

and items with $\sum_{i=1}^{k} y_i = \frac{k}{2}$ (a perfect tie) were marked UNRESOLVED and routed to adjudication.

For adjudication, tied or otherwise low-confidence items were reviewed by two rotating annotators who were not part of the original voting set for that item. They examined only the modality under review and issued a consensus YES/NO. If consensus could not be reached, a third adjudicator served as a tie-breaker.

Finally, for each item and modality we recorded the resulting binary label, the vote histogram $(\#\text{YES}, \#\text{NO})$, the adjudication status, and annotator counts per item. After all batches closed, inter-annotator agreement metrics, including Cohen's Kappa [208], Fleiss's Kappa [209] and Krippendorff's Alpha [210], were computed separately for each modality.

7.5.4 Human re-annotation results

Given the inherently subjective nature of mental manipulation, we compare model performance with human annotations on the same tasks. We observe that human judgments occasionally diverge from the original task labels, and such discrepancies are more pronounced in the speech modality.

We collect human judgments on the dialogues presented in either text or TTS audio modalities. We calculate the inter-annotator agreement represented in pair-wise Cohen's Kappa, Fleiss's Kappa and Krippendorff's Alpha. As pairwise Cohen's Kappa vary by a large degree (Figure 7.3), we focus on the annotators with higher agreement. The high agreement group (annotators B, F, G, H)

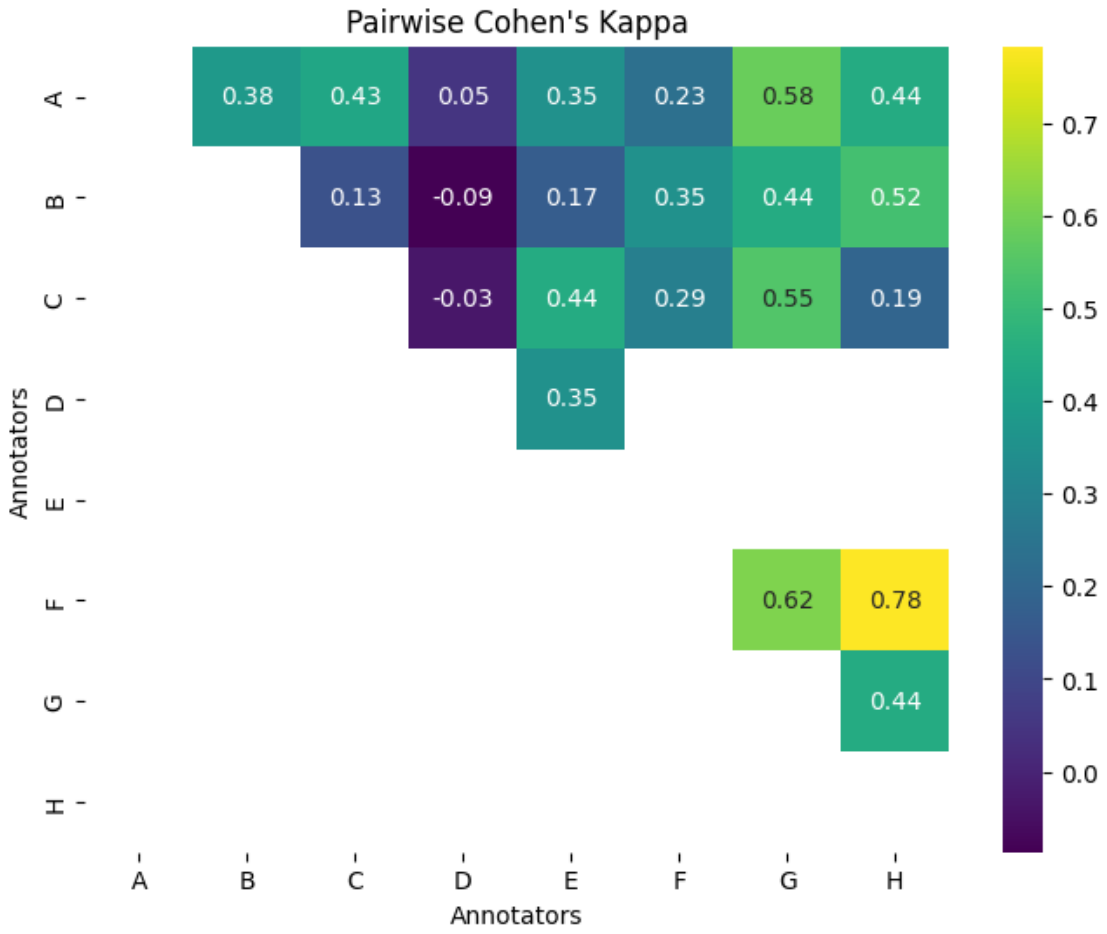


Figure 7.3: Pair-wise Cohen's Kappa between Human Annotators for *Text* modality

for text has Krippendorff's alpha of 0.526 and Fleiss's Kappa of 0.513. These values are slightly lower than the Fleiss's Kappa of 0.596 reported in the original MENTALMANIP dataset [180].

For the audio modality, the high agreement group (annotators B, C, F, H) for text has Krippendorff's alpha of 0.422 and Fleiss's Kappa of 0.514. We observe that some annotators achieve higher agreement on the text modality but not necessarily on audio, suggesting that modality introduces additional variability in how manipulation cues are perceived.

Using majority voting over 100 re-annotated samples, we find that our labels align with the original MENTALMANIP annotations at 0.72 agreement for text and 0.56 for audio, suggesting notably lower consistency in the speech modality. This discrepancy indicates that identifying mental manipulation from speech cues is inherently more ambiguous, probably due to prosodic

Figure 7.4: Pair-wise Cohen's Kappa between Human Annotators for *Audio* modality

and contextual subtleties that were underrepresented or inconsistently interpreted in the original dataset. The lower audio agreement also suggests that the original labels may not fully capture the nuanced intentions conveyed through tone, hesitation, or emphasis –features that often alter perceived manipulation.

7.6 Conclusion

We introduce the first benchmark SPEECHMENTALMANIP for detecting mental manipulation in speech by augmenting the text-based dataset with high-quality, voice-consistent TTS–rendered dialogues. This synthetic multi-speaker extension enables direct comparison between text and audio modalities while systematically examining how prosodic cues affect manipulative intent detection.

Re-annotations	Text		Audio	
MENTALMANIP	1	0	1	0
1	31	19	28	22
0	9	41	22	28

Table 7.6: Agreement between the original MENTALMANIP labels and our re-annotations for 100 samples.

Our experiments show that audio representations make the task substantially more challenging: both humans and models exhibit lower agreement and accuracy when manipulation must be inferred from speech rather than text. These findings highlight that mental manipulation is not only a difficult computational task but also an inherently subjective phenomenon, shaped by tone, delivery, and context.

Future work will expand this benchmark toward more diverse voices and natural speech, refine theoretical definitions of manipulation, and explore modeling strategies that explicitly account for subjectivity and multimodal ambiguity. As perception of manipulation can vary widely across individuals and contexts, clearer theoretical grounding is essential to ensure consistency in both human judgments and machine predictions. We hope this work lays a foundation for developing safer, more socially aware dialogue systems that can reason about manipulative intent across modalities.

7.7 Limitations

Our task involves inherently subjective judgments, as perceptions of mental manipulation can vary across annotators and contexts. While we curate samples from the consensus set, the re-annotated samples may capture only a subset of manipulative strategies represented in the original dataset, limiting generalizability.

In addition, our use of text-to-speech (TTS) synthesis for some audio stimuli may not fully reflect the richness and variability of natural human speech, potentially affecting both human and model interpretation. Our synthetic dialogues are generated on a turn-by-turn basis and therefore do not capture overlapping speech, interruptions, or backchanneling commonly observed in natural

conversation. This design choice prioritizes experimental control: overlapping speech remains challenging for current audio-language models and can introduce confounds related to speech separation, diarization, and acoustic comprehension. As our goal is to isolate how prosodic cues and delivery affect manipulation reasoning, rather than to stress-test low-level audio robustness, we intentionally evaluate models under clean, non-overlapping conditions. Despite these limitations, we do not claim that synthetic speech faithfully represents natural manipulative behavior, but to provide a controlled testbed for isolating modality effects. By rendering transcript-aligned speech with consistent speaker identities and minimized acoustic confounds, we can probe how audio-language models and humans interpret manipulative intent when lexical content is held fixed, an analysis that would be difficult to conduct with in-the-wild recordings. Incorporating statistically generated overlap (e.g., via Behavior-SD style simulation) represents an important direction for future work, enabling evaluation under more ecologically realistic conversational dynamics once baseline behaviors are established.

Finally, our evaluation relied on a single audio-language model (Qwen2.5-Omni) and a few-shot prompting strategy that did not include explicit definitions of manipulation tactics. While this choice established a stable baseline and tested the model's inherent semantic understanding, it leaves open the question of whether definition-augmented prompting or alternative architectures would yield different sensitivity patterns. Expanding the benchmark to a broader suite of models and prompt strategies remains a critical direction for future work.

Chapter 8: Conclusion

This dissertation has advanced a unified perspective on socially meaningful conversational AI. We first strengthened the foundations of interaction modeling by improving how systems represent discourse structure and interpersonal coordination through dialog act alignment and acoustic-prosodic entrainment. These contributions establish that successful conversation depends on both organized communicative intent and dynamic behavioral adaptation.

Building on these insights, we examined empathy as a specialized form of conversational collaboration in which emotional understanding becomes central. We modeled how empathy is expressed and detected in speech, revealed the inherent complexity and ambiguity of multimodal emotional signals, and introduced scalable methods for generating supportive responses grounded in psychotherapy theory. This work demonstrates that empathy requires not just affect recognition but interpretive reasoning and expressive alignment.

Finally, we addressed the ethical responsibility that accompanies emotionally capable systems. By developing methods to detect mental manipulation from speech, we showed that conversational AI must protect users from coercive influence, especially in high-vulnerability contexts.

Together, these contributions articulate a progression from *understanding* the mechanics of human interaction, to *enabling* emotionally supportive behavior, and finally to *safeguarding* users as systems grow more persuasive and expressive. This dissertation takes a step toward empathetic, trustworthy conversational AI that respects human agency and promotes well-being.

The Role of Curated Data in the Era of Foundation Models

A recurring theme of this dissertation is the continued importance of carefully constructed datasets, even in the era of zero-shot and foundation models. While large pretrained models reduce

the need for task-specific training data, they do not eliminate the need for well-defined benchmarks, aligned annotations, and controlled evaluation settings. In fact, as models become more capable and opaque, high-quality datasets become increasingly important for interpreting behavior, diagnosing failure modes, and ensuring safety.

The corpora introduced here—spanning authentic speech (telephone conversations), semi-acted and therapeutic interactions, audiovisual media, and synthetic speech—reflect complementary design goals. Authentic recordings preserve natural conversational dynamics; acted or interview-style data emphasize emotionally salient behaviors that are easier to annotate reliably; and synthetic or TTS-generated data provide scalability and experimental control. This combination enables both ecological validity and reproducible experimentation.

Therefore, rather than assuming that large models obviate the need for data collection, this work argues that curated datasets remain foundational scientific infrastructure. They define what we measure, how we evaluate progress, and how we ensure that increasingly powerful conversational systems behave safely and responsibly.

Future Directions

The work presented here opens several important avenues for continued research in empathetic and socially responsible conversational AI. Advancing empathetic AI invites continued progress in four interconnected areas.

First, ethical considerations must remain central. As emotionally capable systems grow more persuasive, simply maximizing user-perceived empathy is not always beneficial. Highly sycophantic behaviors — over-accommodation, exaggerated agreement, or reinforcement of unhealthy beliefs — may feel supportive in the moment while undermining user well-being. Future systems will require context-sensitive guardrails that distinguish care from appeasement and prioritize user autonomy and safety.

Second, this ethical concern directly motivates the need for better evaluation frameworks. To prevent sycophancy, models must be assessed not only on emotional attunement but also

on whether their responses promote constructive outcomes and psychological safety over time. Current single-turn or surface-level metrics do not capture these distinctions. Richer evaluation paradigms — including longitudinal studies, contextual judgment criteria, and user-centered success measures — are necessary to identify when empathy is genuinely helpful versus when it crosses into manipulation or overreliance. Human-in-the-loop studies and multistage metrics that separate emotional attunement from harmful influence, will be essential. A comprehensive survey that maps existing empathy endeavors and highlights overlooked gaps would meaningfully support progress in the field.

Third, scale and modality present important opportunities. While recent advances in large language models have demonstrated impressive social reasoning capabilities, empathy grounded in speech remains understudied. This dissertation contributes to closing that gap by focusing on acoustic-prosodic and multimodal empathy, but future work should explore larger models that integrate voice, prosody, and contextual cues more holistically. Beyond English, expanding to cross-lingual and culturally adaptive modeling will be increasingly necessary to ensure inclusive and equitable access to support.

Finally, conversational dynamics themselves are not static. As communication norms evolve across cultures, platforms, and communities, foundational constructs such as entrainment, turn-taking, and dialog acts may manifest differently over time. Revisiting these dynamics in emerging communication settings — virtual interaction, voice assistants, multilingual social media spaces — will help ensure that our models remain relevant to how people actually communicate.

Overall, the future of empathetic AI lies not only in building more powerful models, but in developing systems that understand the diversity and complexity of human communication, support users responsibly, and adapt as society and language evolve.

References

[1] A. Sharma, A. Miner, D. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 5263–5276.

[2] A. Welivita and P. Pu, "A taxonomy of empathetic response intents in human social conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4886–4899.

[3] K. Ito, M. Murata, T. Ohno, and S. Matsubara, "Relation between degree of empathy for narrative speech and type of responsive utterance in attentive listening," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari et al., Eds., Marseille, France: European Language Resources Association, May 2020, pp. 696–701, ISBN: 979-10-95546-34-4.

[4] A. C. Lahnala, B. Neuendorf, A. Thomin, C. Welch, T. Stibane, and L. Flek, "Appraisal framework for clinical empathy: A novel application to breaking bad news conversations," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 1393–1407.

[5] 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*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381.

[6] A. Welivita, C.-H. Yeh, and P. Pu, "Empathetic response generation for distress support," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, Eds., Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 632–644.

[7] S. Tafreshi, O. De Clercq, V. Barriere, S. Buechel, J. Sedoc, and A. Balahur, "WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and*

Social Media Analysis, O. De Clercq et al., Eds., Online: Association for Computational Linguistics, Apr. 2021, pp. 92–104.

[8] V. Barriere, S. Tafreshi, J. Sedoc, and S. Alqahtani, "WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, J. Barnes et al., Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 214–227.

[9] V. Barriere, J. Sedoc, S. Tafreshi, and S. Giorgi, "Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, J. Barnes, O. De Clercq, and R. Klinger, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 511–525.

[10] S. Giorgi, J. Sedoc, V. Barriere, and S. Tafreshi, "Findings of WASSA 2024 shared task on empathy and personality detection in interactions," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 369–379.

[11] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in virtual agents and robots: A survey," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 3, Sep. 2017.

[12] A. Lahnala, C. Welch, D. Jurgens, and L. Flek, "A critical reflection and forward perspective on empathy and natural language processing," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2139–2158.

[13] S. Tahir, S. A. Shah, and J. Abu-Khalaf, *Artificial empathy classification: A survey of deep learning techniques, datasets, and evaluation scales*, 2023. arXiv: 2310.00010 [cs.RO].

[14] M. R. Hasan, M. Z. Hossain, S. Ghosh, A. Krishna, and T. Gedeon, "Empathy detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods," *IEEE Transactions on Affective Computing*, pp. 1–20, 2025.

[15] R. Chen, J. Shin, and J. Hirschberg, *Synthempathy: A scalable empathy corpus generated using llms without any crowdsourcing*, 2025. arXiv: 2502.17857 [cs.CL].

[16] Y. Zhang et al., "STICKERCONV: Generating multimodal empathetic responses from scratch," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7707–7733.

[17] H. Yan et al., "Talk with human-like agents: Empathetic dialogue through perceptible acoustic reception and reaction," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 009–15 022.

[18] R. Chen et al., "Detecting Empathy in Speech," in *Interspeech 2024*, 2024, pp. 1080–1084.

[19] H. Chen, R. Chen, and J. Hirschberg, "EmoKnob: Enhance voice cloning with fine-grained emotion control," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8170–8180.

[20] A. Welivita, Y. Xie, and P. Pu, "A large-scale dataset for empathetic response generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1251–1264.

[21] L. Wang et al., "Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4634–4645.

[22] S. Sabour, C. Zheng, and M. Huang, "Cem: Commonsense-aware empathetic response generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 229–11 237, 2022.

[23] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling," in *Interspeech 2013*, 2013, pp. 2861–2865.

[24] B. Xiao et al., "Modeling therapist empathy through prosody in drug addiction counseling," in *Interspeech 2014*, 2014, pp. 213–217.

[25] B. Xiao, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Interspeech 2015*, 2015, pp. 2489–2493.

[26] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Interspeech 2015*, 2015, pp. 1947–1951.

[27] J. Gibson et al., "A deep learning approach to modeling empathy in addiction counseling," in *Interspeech 2016*, 2016, pp. 1447–1451.

[28] F. Alam, M. Danieli, and G. Riccardi, "Annotating and modeling empathy in spoken conversations," *Computer Speech Language*, vol. 50, pp. 40–61, 2018.

[29] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "Studies: Corpus of japanese empathetic dialogue speech towards friendly voice agent," in *Interspeech 2022*, 2022, pp. 5155–5159.

[30] Y. Saito, E. Iimori, S. Takamichi, K. Tachibana, and H. Saruwatari, "Calls: Japanese empathetic dialogue speech corpus of complaint handling and attentive listening in customer center," in *Interspeech 2023*, 2023, pp. 5561–5565.

[31] D. Tao, H. Chui, S. Luk, and T. Lee, "Cuempathy: A counseling speech dataset for psychotherapy research," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 354–358.

[32] D. Bhattacharya, E. Lin, R. Chen, and J. Hirschberg, "Switching tongues, sharing hearts: Identifying the relationship between empathy and code-switching in speech," in *Interspeech 2024*, 2024, pp. 492–496.

[33] V. Sorin et al., "Large language models and empathy: Systematic review," *J Med Internet Res*, vol. 26, e52597, 2024.

[34] H. J. Xie, J. Zhang, X. Zhang, and K. Liu, *Scoring with large language models: A study on measuring empathy of responses in dialogues*, 2024. arXiv: 2412.20264 [cs.CL].

[35] F. Chen et al., "Empathy prediction from diverse perspectives," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8959–8974, ISBN: 979-8-89176-251-0.

[36] T. Liu et al., *The illusion of empathy: How ai chatbots shape conversation perception*, 2025. arXiv: 2411.12877 [cs.HC].

[37] C. Reguera-Gómez, D. Paperno, and M. H. T. de Boer, "Empathy vs neutrality: Designing and evaluating a natural chatbot for the healthcare domain," in *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, R. Johansson and S. Stymne, Eds., Tallinn, Estonia: University of Tartu Library, Mar. 2025, pp. 508–517, ISBN: 978-9908-53-109-0.

[38] G. M. Lucas et al., "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.

[39] V. D. Swain et al., "AI on My Shoulder: Supporting Emotional Labor in Front-Office Roles with an LLM-based Empathetic Coworker," in *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2025, Yokohoma, Japan, April 26-May 1, 2025*, ACM, 2025.

[40] R. Chen et al., "Raswda: Re-aligned switchboard dialog act corpus for dialog act prediction in conversations," in *14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*, 2024.

[41] R. Chen, S. Kim, A. Papangelis, J. Hirschberg, Y. Liu, and D. Hakkani-Tür, "Identifying entrainment in task-oriented conversations," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[42] M. Srikanth, R. Chen, and J. Hirschberg, "Mixed signals: Understanding model disagreement in multimodal empathy detection," in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, K. Inui et al., Eds., Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Dec. 2025, pp. 1978–1991, ISBN: 979-8-89176-303-6.

[43] R. Chen, W. Liang, Z. Gong, L. Ai, and J. Hirschberg, "Detecting mental manipulation in speech via synthetic multi-speaker dialogue," in *Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2026)*, Trento, Italy: Association for Computational Linguistics, 2026.

[44] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO, Tech. Rep. 97-02, 1997.

[45] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, 517–520 vol.1.

[46] A. Stolcke et al., "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[47] D. Ortega and N. Thang Vu, "Lexico-acoustic neural-based models for dialog act classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6194–6198.

[48] T. Tran, "Neural models for integrating prosody in spoken language understanding," PhD thesis, University of Washington, Seattle, WA, 2020.

[49] K. Wei et al., "A neural prosody encoder for end-to-end dialogue act classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7047–7051.

[50] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of switchboard.," in *ICSLP*, Syndey, 1998.

[51] S. Calhoun et al., "The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, pp. 387–419, 2010.

[52] E. Bard, C. Sotillo, A. Anderson, and M. Taylor, "The dciem map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, 1996, 1958–1961 vol.3.

[53] J. Carletta et al., "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*, Springer, 2006, pp. 28–39.

[54] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[55] J.-M. Benedí et al., "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy: European Language Resources Association (ELRA), May 2006.

[56] E. Costantini, S. Burger, and F. Pianesi, "NESPOLE¡s multilingual and multimodal corpus," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, M. González Rodríguez and C. P. Suarez Araujo, Eds., Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), May 2002.

[57] S. Keizer, "Dialogue act classification: Experiments with the schisma corpus," Technical report, University of Twente, Tech. Rep., 2002.

[58] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, Cambridge, Massachusetts, USA: Association for Computational Linguistics, 2004, pp. 97–100.

[59] M. Kay, M. Gawron, and P. Norvig, *Verbmobil: A Translation System for Face-to-Face Dialog* (Center for the Study of Language and Information Publication Lecture Notes). Cambridge University Press, 1994, ISBN: 9780937073957.

[60] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari et al., Eds., Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 3369–3373.

[61] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams, and M. Henderson, "The fourth dialog state tracking challenge," in *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*. Singapore: Springer Singapore, 2017, pp. 435–449, ISBN: 978-981-10-2585-3.

[62] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, A. Koller, G. Skantze, F. Jurcicek, M. Araki, and C. P. Rose, Eds., Prague, Czech Republic: Association for Computational Linguistics, Sep. 2015, pp. 285–294.

[63] P. Budzianowski et al., "MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 5016–5026.

[64] M. Eric et al., "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari et al., Eds., Marseille, France: European Language Resources Association, May 2020, pp. 422–428, ISBN: 979-10-95546-34-4.

[65] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, and J. Chen, "MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, T.-H. Wen et al., Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 109–117.

[66] H. Alamri, C. Hori, T. K. Marks, D. Batra, and D. Parikh, "Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7," in *DSTC7 at AAAI2019 Workshop*, vol. 2, 2018.

[67] C. Hori et al., "Overview of the sixth dialog system technology challenge: Dstc6," *Comput. Speech & Lang.*, vol. 55, pp. 1–25, 2019.

[68] J. Allen and M. Core, *Draft of damsl: Dialog act markup in several layers*, 1997.

[69] E. Shriberg et al., "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Lang. and speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[70] C. Potts, *The switchboard dialog act corpus*, https://compprag.christopherpotts.net/swda.html, 2011.

[71] *Sox - sound exchange*, https://sox.sourceforge.net/, 2015.

[72] A. Pettarin, *Aeneas*, https://www.readbeyond.it/aeneas/, 2017.

[73] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot Int.*, vol. 5, no. 9/10, pp. 341–347, 2001.

[74] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)*, 2017, pp. 1–5.

[75] W. B. Putman and R. L. Street, "The conception and perception of noncontent speech performance: Implications for speech-accommodation theory," *Jounral of the Sociology of Language*, vol. 1984, no. 46, pp. 97–114, 1984.

[76] R. Y. Bourhis, H. Giles, and W. E. Lambert, "Social consequences of accommodating one's style of speech: A cross-national investigation," *International Journal of the Sociology of Language*, vol. 6, no. 5, pp. 5–71, 1975.

[77] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception–behavior link and social interaction.," *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[78] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and D. C. Dryer, "Can computer personalities be human personalities?" *International Journal of Human-Computer Studies*, vol. 43, no. 2, pp. 223–239, 1995.

[79] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[80] S. Kim et al., ""how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1147–1154.

[81] C. Lee et al., "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 793–796.

[82] R. Coulston, S. Oviatt, and C. Darves, "Amplitude convergence in children's conversational speech with animated personas," in *Proceedings of ICSLP'02*, 2002.

[83] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPHS*, Citeseer, vol. 3, 2003, pp. 833–836.

[84] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *International conference on artificial intelligence in education*, Springer, 2013, pp. 750–753.

[85] N. Lubold, H. Pon-Barry, and E. Walker, "Naturalness and rapport in a pitch adaptive learning companion," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2015, pp. 103–110.

[86] R. Levitan et al., "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Interspeech 2016*, 2016, pp. 1166–1170.

[87] J. Lopes, M. Eskenazi, and I. Trancoso, "Automated two-way entrainment to improve spoken dialog system performance," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8372–8376.

[88] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *INTERSPEECH*, 2011, pp. 3081–3084.

[89] A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue," in *Interspeech 2007*, Antwerp, Belgium, 2007.

[90] R. Levitan, "Acoustic-prosodic entrainment in human-human and human-computer dialogue," Ph.D. dissertation, Columbia University, Aug. 2014.

[91] 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.

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

[93] J. W. Pennebaker, R. Booth, R. L. Boyd, and M. Francis, *Linguistic inquiry and word count: Liwc2015*, Austin, TX, Sep. 2015.

[94] Z. Rahimi and D. Litman, "Entrainment2vec: Embedding entrainment for multi-party dialogues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8681–8688, 2020.

[95] M. Mizukami, K. Yoshino, G. Neubig, D. Traum, and S. Nakamura, "Analyzing the effect of entrainment on dialogue acts," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, R. Fernandez, W. Minker, G. Carenini, R.

Higashinaka, R. Artstein, and A. Gainer, Eds., Los Angeles: Association for Computational Linguistics, Sep. 2016, pp. 310–318.

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

[97] 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.

[98] 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.

[99] G. M. Lucas et al., "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.

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

[101] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Trans. Comput.-Hum. Interact.*, vol. 12, no. 2, 293–327, 2005.

[102] 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, Jul. 2010.

[103] 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, ISBN: 978-3-540-74997-4.

[104] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," *Intelligent Virtual Agents*, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, Eds., pp. 68–79, 2011.

[105] 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, Budapest, Hungary: International Foundation for Autonomous Agents and Multiagent Systems, 2009, 1399–1400, ISBN: 9780981738178.

[106] 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, Estoril, Portugal: International Foundation for Autonomous Agents and Multiagent Systems, 2008, 89–96, ISBN: 9780981738109.

[107] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in *Intelligent Virtual Agents*, D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, and A. Leuski, Eds., Cham: Springer International Publishing, 2016, pp. 218–233, ISBN: 978-3-319-47665-0.

[108] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: An emotion-aware wellbeing chatbot," in *International Conference on Affective Computing and Intelligent Interaction*, 2019.

[109] T. Tran et al., "Multimodal analysis and assessment of therapist empathy in motivational interviews," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23, <conf-loc>, <city>Paris</city>, <country>France</country>, </conf-loc>: Association for Computing Machinery, 2023, 406–415, ISBN: 9798400700552.

[110] F. Alam, M. Danieli, and G. Riccardi, "Annotating and modeling empathy in spoken conversations," *Comput. Speech Lang.*, vol. 50, no. C, 40–61, 2018.

[111] D. Tao, T. Lee, H. Chui, and S. Luk, "Characterizing therapist's speaking style in relation to empathy in psychotherapy," in *Interspeech 2022*, 2022, pp. 2003–2007.

[112] 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.

[113] 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.

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

[115] 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.

[116] 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.

[117] 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.

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

[119] 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.

[120] 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.

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

[122] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[123] P. Fung et al., "Towards empathetic human-robot interactions," *arXiv preprint arXiv:1605.04072*, 2016.

[124] R. F. Baumeister and K. D. Vohs, *Encyclopedia of Social Psychology*. Sage, 2007, vol. 1.

[125] J. Holler and S. C. Levinson, "Multimodal language processing in human communication," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, 2019.

[126] S. Jabeen, X. Li, A. M. Shoib, S. Li, and A. Jabbar, "Recent advances and trends in multimodal deep learning: A review," *arXiv preprint arXiv:2105.11087*, 2021.

[127] M. R. Hasan, M. Z. Hossain, S. Ghosh, A. Krishna, and T. Gedeon, "Empathy detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods," *arXiv preprint arXiv:2311.00721*, 2023.

[128] 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*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381.

[129] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "Empdg: Multiresolution interactive empathetic dialogue generation," *arXiv preprint arXiv:1911.08698*, 2019.

[130] J. Shen et al., "EmpathicStories++: A multimodal dataset for empathy towards personal experiences," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 4525–4536.

[131] Z. Zhu et al., "Medic: A multimodal dataset for empathic dialogue in counseling," *arXiv preprint arXiv:2305.14221*, 2023.

[132] L. Galland, C. Pelachaud, and F. Pecune, "Emmi: Empathic multimodal motivational interviews dataset: Analyses and annotations," *arXiv preprint arXiv:2406.16478*, 2024.

[133] Y. Zhang et al., "STICKERCONV: Generating multimodal empathetic responses from scratch," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7707–7733.

[134] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114.

[135] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569.

[136] P. Barros, N. Churamani, A. Lim, and S. Wermter, *The omg-empathy dataset: Evaluating the impact of affective behavior in storytelling*, 2019. arXiv: 1908.11706 [cs.HC].

[137] S. Swayamdipta et al., "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 9275–9293.

[138] S. Saha, P. Hase, N. Rajani, and M. Bansal, "Are hard examples also harder to explain? a study with human and model-generated explanations," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2121–2131.

[139] J. Wang, L. Wang, Y. Zheng, C.-C. M. Yeh, S. Jain, and W. Zhang, " Learning-From-Disagreement: A Model Comparison and Visual Analytics Framework," *IEEE Transactions on Visualization & Computer Graphics*, vol. 29, no. 09, pp. 3809–3825, Sep. 2023.

[140] Y. Liu et al., *Roberta: A robustly optimized bert pretraining approach*, 2019. arXiv: 1907.11692 [cs.CL].

[141] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[142] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

[143] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[144] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[145] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[146] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]*, Version 6.2.14, retrieved 24 May 2022, 1992–2022.

[147] Y. Jadoul, B. de Boer, and P. Thompson, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[148] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, 10745–10759, Sep. 2023.

[149] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[150] W. V. Friesen and P. Ekman, *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

[151] T. Sainburg, L. McInnes, and T. Q. Gentner, "Parametric umap embeddings for representation and semisupervised learning," *Neural Computation*, vol. 33, no. 11, pp. 2881–2907, 2021.

[152] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[153] N.-J. Jiang and M.-C. de Marneffe, "Investigating reasons for disagreement in natural language inference," *Transactions of the Association for Computational Linguistics*, vol. 10, B. Roark and A. Nenkova, Eds., pp. 1357–1374, 2022.

[154] E. Pavlick and T. Kwiatkowski, "Inherent disagreements in human textual inferences," *Transactions of the Association for Computational Linguistics*, vol. 7, L. Lee, M. Johnson, B. Roark, and A. Nenkova, Eds., pp. 677–694, 2019.

[155] C. Qian et al., *Dyncim: Dynamic curriculum for imbalanced multimodal learning*, 2025. arXiv: 2503.06456 [cs.CV].

[156] M. Huang, C. Qing, J. Tan, and X. Xu, "Context-based adaptive multimodal fusion network for continuous frame-level sentiment prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3468–3477, 2023.

[157] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, *Defending multimodal fusion models against single-source adversaries*, 2022. arXiv: 2206.12714 [cs.CV].

[158] C. Bai, H. Chen, S. Kumar, J. Leskovec, and V. S. Subrahmanian, *M2p2: Multimodal persuasion prediction using adaptive fusion*, 2021. arXiv: 2006.11405 [cs.CV].

[159] M. Y. Baihaqi, A. G. Contreras, S. Kawano, and K. Yoshino, *Rapport-driven virtual agent: Rapport building dialogue strategy for improving user experience at first meeting*, 2024. arXiv: 2406.09839 [cs.CL].

[160] B. Zhou, H. Wang, Y. Yao, T. Chen, F. Xu, and X. Ma, "Simulate, refine and integrate: Strategy synthesis for efficient smt solving," in *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, ser. IJCAI-2024, International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, 7976–7984.

[161] M. A. Webb and J. P. Tangney, "Too good to be true: Bots and bad data from mechanical turk," *Perspectives on Psychological Science*, p. 17 456 916 221 120 027, 2022.

[162] A. J. Moss, C. Rosenzweig, J. Robinson, S. N. Jaffe, and L. Litman, "Is it ethical to use mechanical turk for behavioral research? relevant data from a representative survey of mturk participants and wages," *Behavior Research Methods*, vol. 55, no. 8, pp. 4048–4067, 2023.

[163] M. L. Mauriello, T. Lincoln, G. Hon, D. Simon, D. Jurafsky, and P. Paredes, "Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '21, <conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>: Association for Computing Machinery, 2021, ISBN: 9781450380959.

[164] K. Girotra, L. Meincke, C. Terwiesch, and K. T. Ulrich, "Ideas are dimes a dozen: Large language models for idea generation in innovation," *SSRN Electronic Journal*, 2023.

[165] R. Chen et al., "Detecting empathy in speech," in *Proc. INTERSPEECH 2024*, 2024.

[166] A. Welivita, Y. Xie, and P. Pu, *Fine-grained emotion and intent learning in movie dialogues*, 2020. arXiv: 2012.13624 [cs.CL].

[167] Y. Chen et al., "SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1170–1183.

[168] R. Goel, S. Susan, S. Vashisht, and A. Dhanda, "Emotion-aware transformer encoder for empathetic dialogue generation," in *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2021, pp. 1–6.

[169] J. Y. Lee, K. A. Lee, and W. S. Gan, "Improving contextual coherence in variational personalized and empathetic dialogue agents," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7052–7056.

[170] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[171] Y. K. Lee, I. Lee, M. Shin, S. Bae, and S. Hahn, *Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models*, 2023. arXiv: 2311.04915 [cs.CL].

[172] W. Wang, Z. Li, D. Lian, C. Ma, L. Song, and Y. Wei, "Mitigating the language mismatch and repetition issues in LLM-based machine translation via model editing," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15 681–15 700.

[173] F. Liu et al., "Exploring and evaluating hallucinations in llm-powered code generation," *arXiv preprint arXiv:2404.00971*, 2024.

[174] K. Lee et al., "Deduplicating training data makes language models better," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8424–8445.

[175] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3687–3697.

[176] A. Barnhill, "What is manipulation?" In *Manipulation: Theory and Practice*, C. Coons and M. Weber, Eds., Oxford, UK: Oxford University Press, 2014.

[177] M. Ienca, "On artificial intelligence and manipulation," *Topoi*, vol. 42, no. 3, pp. 833–842, 2023.

[178] J. Hamel, C. E. B. Cannon, and N. Graham-Kevan, "The consequences of psychological abuse and control in intimate partner relationships," *Traumatology*, 2023, Advance online publication.

[179] G. K. Simon and K. Foley, *In Sheep's Clothing: Understanding and Dealing with Manipulative People*. Old Saybrook, CT: Tantor Media, Incorporated, 2011, Audiobook edition.

[180] Y. Wang, I. Yang, S. Hassanpour, and S. Vosoughi, *MentalManip: A dataset for fine-grained analysis of mental manipulation in conversations*, 2024. arXiv: 2405.16584 [cs.CL].

[181] Z. Gong, M. Yao, X. Hu, X. Zhu, and J. Hirschberg, "A mapping on current classifying categories of emotions used in multimodal models for emotion recognition," in *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, S. Henning and M. Stede, Eds., St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 19–28.

[182] C. Wu et al., "Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects," *arXiv preprint arXiv:2505.20511*, 2025.

[183] S. R. Moghaddam and C. J. Honey, *Boosting theory-of-mind performance in large language models via prompting*, Preprint, 2023. arXiv: 2304.11490 [cs.CL].

[184] Z. Chen et al., "Tombench: Benchmarking theory of mind in large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 15 959–15 983.

[185] J. W. A. Strachan et al., "Testing theory of mind in large language models and humans," *Nature Human Behaviour*, vol. 8, no. 7, pp. 1285–1295, 2024.

[186] J. Ma et al., "Detecting conversational mental manipulation with intent-aware prompting," in *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 9176–9183.

[187] H. Yang, L. Qu, E. Shareghi, and G. Haffari, "Audio is the achilles' heel: Red teaming audio large multimodal models," *arXiv preprint arXiv:2410.23861*, 2024.

[188] R. Peri et al., "Speechguard: Exploring the adversarial robustness of multi-modal large language models," in *Findings of the Association for Computational Linguistics: ACL 2024*,

Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 018–10 035.

[189] X. Shen, Y. Wu, M. Backes, and Y. Zhang, "Voice jailbreak attacks against gpt-4o," *arXiv preprint arXiv:2405.19103*, 2024.

[190] R. S. Kern et al., "Theory of mind deficits for processing counterfactual information in persons with chronic schizophrenia," *Psychological Medicine*, vol. 39, no. 4, pp. 645–654, 2009.

[191] M. Lampron, A. M. Achim, D. Gamache, A. Bernier, S. Sabourin, and C. Savard, "Profiles of theory of mind impairments and personality in clinical and community samples: Integrating the alternative dsm-5 model for personality disorders," *Frontiers in Psychiatry*, vol. 14, p. 1 292 680, 2024.

[192] T. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[193] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing Systems*, vol. 35, Curran Associates, Inc., 2022, pp. 22 199–22 213.

[194] I. Yang, X. Guo, S. Xie, and S. Vosoughi, *Enhanced detection of conversational mental manipulation through advanced prompting techniques*, 2024. arXiv: 2408.07676 [cs.CL].

[195] Y. Gao, H. Bao, T. Zhang, B. Li, Z. Wang, and W. Chen, *Mentalmac: Enhancing large language models for detecting mental manipulation via multi-task anti-curriculum distillation*, 2025. arXiv: 2505.15255 [cs.CL].

[196] J. Shen et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada: IEEE Press, 2018, 4779–4783.

[197] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 5530–5540.

[198] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018.

[199] C. Wang et al., *Neural codec language models are zero-shot text to speech synthesizers*, 2023. arXiv: 2301.02111 [cs.CL].

[200] Z. Du et al., "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[201] X. Lyu, Y. Wang, T. Zhao, H. Wang, H. Liu, and Z. Du, "Build llm-based zero-shot streaming tts system with cosyvoice," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–2.

[202] S. Liang, R. Zhou, and Q. Yuan, "Ece-tts: A zero-shot emotion text-to-speech model with simplified and precise control," *Applied Sciences*, vol. 15, no. 9, 2025.

[203] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, F. Keller and D. Reitter, Eds., Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 76–87.

[204] D. Sasu et al., "Akan cinematic emotions (ace): A multimodal multi-party dataset for emotion recognition in movie dialogues," *arXiv preprint arXiv:2502.10973*, 2025.

[205] C. Tang et al., "SALMONN: Towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[206] G. Comanici et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[207] X. Wang, J. Wei, D. Schuurmans, and et al., "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[208] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. eprint: https://doi.org/10.1177/001316446002000104.

[209] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, Nov. 1971.

[210] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications, 2004.