Multimodal Analysis and Assessment of Therapist Empathy in Motivational Interviews

Trang Tran
Yufeng Yin
Leili Tavabi
ttran@ict.usc.edu
yin@ict.usc.edu
ltavabi@ict.usc.edu
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA

Joannalyn Delacruz
Brian Borsari
Joshua Woolley
joannalyn.delacruz@va.gov
brian.borsari@va.gov
josh.woolley@ucsf.edu
San Francisco VAHCS & UCSF
Psychiatry & Behavioral Sciences
San Francisco, CA, USA

Stefan Scherer
Mohammad Soleymani
scherer@ict.usc.edu
soleymani@ict.usc.edu
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA

ABSTRACT

The quality and effectiveness of psychotherapy sessions are highly influenced by the therapists' ability to meaningfully connect with clients. Automated assessment of therapist empathy provides cost-effective and systematic means of assessing the quality of therapy sessions. In this work, we propose to assess therapist empathy using multimodal behavioral data, *i.e.* spoken language (text) and audio in real-world motivational interviewing (MI) sessions for alcohol abuse intervention. We first study each modality (text vs. audio) individually and then evaluate a multimodal approach using different fusion strategies for automated recognition of empathy levels (high vs. low). Leveraging recent pre-trained models both for text (DistilRoBERTa) and speech (HuBERT) as strong unimodal baselines, we obtain consistent 2-3 point improvements in F1 scores with early and late fusion, and the highest absolute improvement of 6–12 points over unimodal baselines. Our models obtain F1 scores of 68% when only looking at an early segment of the sessions and up to 72% in a therapist-dependent setting. In addition, our results show that a relatively small portion of sessions, specifically the second quartile, is most important in empathy prediction, outperforming predictions on later segments and on the full sessions. Our analyses in late fusion results show that fusion models rely more on the audio modality in limited-data settings, such as in individual quartiles and when using only therapist turns. Further, we observe the highest misclassification rates for parts of the sessions with MI inconsistent utterances (20% misclassified by all models), likely due to the complex nature of these types of intents in relation to perceived empathy.

KEYWORDS

Motivational interview, empathy, speech, language, multimodal learning

ACM Reference Format:
Trang Tran, Yufeng Yin, Leili Tavabi, Joannalyn Delacruz, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2023. Multimodal Analysis and Assessment of Therapist Empathy in Motivational Interviews. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614105

1 INTRODUCTION

Empathy in psychotherapy, which is described as the therapist experiencing an accurate understanding of the client's awareness of their own state [49], is essential in reaching desired therapeutic outcomes [51, 57]. Due to its importance and effect on behavioral outcomes, therapist empathy is commonly used to assess the quality of therapy sessions. In this work, we focus on Motivational Interviewing (MI), which is an evidence-based therapeutic style of communication with particular attention to the language of change. It is designed to strengthen personal motivation for a specific goal by exploring and resolving ambivalence [39] and the therapist's ability to understand and elicit the client's own reasons for change is crucial. Therefore, empathy is a consistent evaluation metric in MI, commonly provided by standardized MI coding systems like the Motivational Interviewing Skill Code 2.5 (MISC) [40] and the Motivational Interviewing Treatment Integrity 3.1 (MITI) [42].

Following these coding systems, empathy along with other gestalt measures (*i.e.* MI spirit) are annotated by trained third-party coders who listen to the entire session's audio recordings to provide behavioral codes for both the client and therapist. This process is time-consuming, costly and therefore hard to scale. For example, obtaining the session-level quality ratings requires trained coders to listen to, review, and rate the session following the aforementioned standardized codings, MISC and MITI, on 7-point or 5-point Likert scales, respectively. Typical MI sessions are 45-60 minutes long, so the annotation process is labor-intensive.

The focus of our work is to build models that utilize therapist (and client) speech for estimating the session-level empathy ratings (standardized across datasets). For this goal, we use two clinical datasets consisting of real-world MI sessions on alcohol-related issues [4, 11]. Motivated by past work showing the potential relevance of certain temporal segments to outcomes [18, 24], we model empathy by taking the input content from certain segments, roughly associated with different stages of the session, and different topics

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614105

elicited by the therapist. To this end, we divide a therapy session into four equal-length sequential segments (quartiles) and focus on the segments hypothesized to have a higher salience for estimating empathy or behavioral outcome. Specifically, we focus on (1) the second quartile where the client discusses problematic behaviors, potentially with more opportunities for the therapist to provide an empathetic understanding; and (2) the last quartile based on prior evidence indicating the importance of the session's final segment in association with subsequent behavioral outcomes [1], in addition to the potential recency bias of annotators.

Our work leverages recently proposed language and audio processing methods for multimodal understanding and modeling of empathy. Our analyses provide insights into the utility of each modality, individually and in combination, *i.e.* how they are fused in empathy level classification. Our contributions are as follows.

- We present a comprehensive study of using speech transcript and audio in unimodal and multimodal empathy estimation, showing the benefits of each modality and fusion methods.
- We show that our multimodal fusion methods outperform both unimodal approaches, including the text modality that often dominates learning in prior multimodal work, and provide analyses on how each modality is used in each setting.
- Our findings show that the second quartile of MI sessions, *i.e.* an early phase in the conversation after introduction but before detailed strategy discussion, is most informative for empathy estimation in all models, including the larger-data setting of using full sessions.
- We perform error analysis and show that our models seem to learn and rely on important aspects of audio complementary to the text modality. In addition, we observe the highest misclassification rates for parts of the sessions with higher MI inconsistent utterances, likely due to the complex nature of these types of intents in relation to perceived empathy.

2 RELATED WORK

Quality assessment of therapy sessions can provide valuable insights into how a competent therapist operates and what kind of therapist-client interactions are productive. Researchers have explored approaches to building automatic systems for quality assessment in different types of therapy. For example, Xiao et al. [61] trained a model to predict empathy levels (high vs. low) using session-level lexical features, showing a significant correlation with expert-coded empathy scores. To model empathy in a temporally dynamic way, Chakravarthula et al. [8] integrated the language model features into a Hidden Markov Model (HMM) in order to allow for behavioral transitions throughout the session, showing performance improvements over static models. Researchers have also used the psycholinguistically-motivated Linguistic Inquiry and Word Count (LIWC) [45] features for the estimation of empathy levels. They showed that LIWC features provide performance gains for modeling empathy compared to the standard n-gram features. LIWC features were also effective for recognizing client intent (change vs. sustain talk) [56]. Lord et al. [30] used LIWC features to compute language style synchrony between clients and therapists, showing that client-therapist style synchrony is predictive of empathy ratings.

In addition to using session transcripts, Can et al. [5] used Conditional Random Fields (CRF) in a sequence tagging framework for utterance-level behavioral coding as an intermediate step to predict session-level quality measures like empathy. Leveraging the advances in neural network models, a series of more recent work have focused on using word embeddings, from non-contextualized representations such as GloVe [46] and word2vec [37], to newer large contextualized models such as BERT [13], for natural language understanding in therapy assessment. For example, Gibson et al. [20] modeled MI sessions using a recurrent neural network (RNN) applied to word2vec embeddings to obtain utterance-level representations, which are then used as features to predict empathy level. Additionally, Flemotomos et al. [17] used GLoVe embeddings as well as LIWC features to estimate Cognitive Behavioral Therapy (CBT) session quality as measured by the Cognitive Therapy Rating Scale (CTRS) scores. They found that the therapist's language has higher predictive power than client's language, which is intuitive given the ratings are provided based on the therapist's behaviors. In their follow-up study, Flemotomos et al. [16] incorporated highly contextualized representations, *i.e.* by using BERT-based embedding in their classifiers, achieving consistent improvements for session quality assessment.

Beyond transcripts of therapy sessions, speech is an important modality that has the potential to improve empathy assessment in ways complementary to text. Xiao et al. [62] studied empathy in connection to similarity or entrainment in interpersonal interactions. They used speech prosody features, including Mel Frequency Cepstral Coefficients (MFCCs) and pitch, to measure the similarity between speaker turns in terms of Kullback-Leibler (KL) divergence, showing a significant correlation with session-level empathy. In their follow-up work, Xiao et al. [60] used prosodic features like pitch, vocal energy, etc. for automatic prediction of empathy level. They showed that the therapist's use of high energy and pitch in their voice is associated with low empathy. Additionally, Xiao et al. [63] demonstrated that speech rate cues provide complementary information to speech prosody features in recognizing empathy. They also showed that the degree of turn-level entrainment of speech rate between the therapist and client correlates with the therapist's empathy rating. More recent work has also focused on modeling vocal entrainment in dyadic interactions using neural networks. Focusing on interactions involving individuals with Autism Spectrum Disorders, Lahiri et al. [27] utilizes conformers to capture both short- and long-term conversational context to model vocal entrainment patterns while using cross-subject attention to model intra- and inter-personal signals for modeling entrainment patterns throughout the interaction. In addition to modeling empathy in counseling sessions, other work developed models of empathy to facilitate empathetic human-machine interactions [35, 36]. For instance, Tavabi et al. [53] proposed multimodal models for identifying opportunities for expressing empathy in human-agent interactions, while Mathur et al. [34] modeled user empathy elicited by a story-telling robot using facial features.

Past work on computational analysis of therapy sessions for diagnosis [48] and quality assessment [54] demonstrated large differences in performance across different modalities, with language outperforming audio by a wide margin. This phenomenon has also been observed in other human-centered multimodal learning problems such as multimodal sentiment analysis [22, 64]. Superior performance from the text modality is due to its higher level of abstraction and the natural communicative affordances language provides over other modalities. Moreover, large language encoders are pre-trained with a vast amount of data, further enhancing language's advantage in multimodal analysis. This large performance gap between unimodal performances results in suboptimal learning in multimodal fusion where the unimodal encoder of the dominant, high-performing modality converges faster, resulting in underfitting in the other modalities [19, 28].

Most existing work on the recognition of therapist empathy focuses on unimodal language or audio models, leaving the multimodal analysis relatively less explored. In this work, we follow a multimodal approach, using both language and audio, while focusing on specific session stages for modeling therapist empathy.

3 DATASET

3.1 Dataset Overview

This work uses two clinical datasets of real-world Motivational Interviewing (MI) sessions. Our datasets come from MI sessions recorded from two populations: (1) College students mandated to take part in MI sessions due to alcohol-related problems [4] (denoted as 'Dataset1' in the rest of this paper) and (2) Community-based underage (ages 17-20) heavy drinkers transitioning out of high school who were not immediately planning to enroll in a 4-year college (denoted as 'Dataset2' in the rest of this paper). These participants, unlike in Dataset1, were volunteers who were recruited via advertisement held at local high schools, community colleges, etc. [11]. Both populations participated in single-session face-to-face MI meetings, which were approximately 50-60 minutes long.

Dataset1 contains 219 taped MI sessions with manual transcriptions. These sessions are coded following the MISC 2.5 guidelines [40] for local utterance-level behaviors, as well as global ratings of empathy and other MI-related measures like therapists' acceptance and MI spirit. The recordings were at 16 kHz mono, and we used a forced aligner, Speechmatics,[1] to obtain time stamps for each turn. Dataset2 comprises 82 MI sessions with digital audio recordings but no manual transcripts. The sessions were recorded at 44.1 kHz in stereo. We used the Google Cloud automatic speech recognition (ASR) service to automatically transcribe the sessions. The turn-level time stamps were extracted from the speech recognition output. We manually verified the ASR quality of a subset of sessions and found the transcriptions have only a few minor issues that do not affect this work, *e.g.* misrecognition of proper names associated with the sites, missing or inserting disfluencies (such as 'uh's and 'um's). Dataset2 sessions are also annotated with utterance-level codes and global session-level ratings of therapist skills like empathy and acceptance, following the MITI 3.1 coding system. During coding, an expert clinician listens to a session and marks where an utterance associated with an MI code starts and ends on the MITI software [42].

In each dataset, 20% of the sessions were randomly selected and double-coded to verify interrater reliability. Intraclass correlation coefficients (ICCs; two-way mixed, single measure) were calculated based on these double-coded sessions; ICCs range from 0.47 to 0.83 in both datasets, which is considered "fair" to "excellent" [10]. Dataset1's sessions were run by 13 therapists (min=1, max=44 sessions), Dataset2 has 3 therapists (min=5, max=53 sessions). Other data statistics are presented in Table 1.

Table 1: Dataset statistics: session length (in minutes) and average number of turns; numbers in parentheses are standard deviations.

	# Sess.	Avg. Duration	Avg. # Turns
Dataset1	219	49.9 (±13.7)	422 (±128.2)
Dataset2	82	54.3 (±12.1)	600 (±138.5)

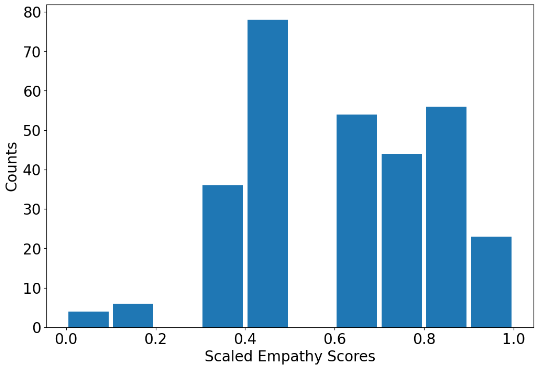

Figure 1: Histogram of scaled empathy ratings in our combined dataset.

Since our two datasets follow empathy rating systems on different Likert scales (5-point in Dataset1 and 7-point in Dataset2), we scale the empathy ratings between 0 and 1. In addition, following previous work in empathy prediction, we binarized these scaled empathy scores, mapping sessions with the scaled empathy score > 0.5 to "high" and the rest (≤ 0.5) to "low." This quantization resulted in a slightly skewed combined dataset, with 58% of positive (high) empathy sessions. Per-therapist scaled empathy scores' denoted as [mean range] ± [sdev range] are [0.36-0.83] ± [0.09-0.22] for Dataset1, [0.64-0.90] ± [0.14-0.17] for Dataset2. Figure 1 shows the score distribution of the overall dataset.

3.2 Session Quartiles

MI therapists are trained to follow a common (but flexible) structure, which can be studied in different temporal segments [1, 25]. Given the format of the intervention commonly following four stages with four high-level topics, we decided to divide the session into quartiles. Sessions typically start (Q1) with the therapist and client introducing themselves and the therapist providing details about the session's structure and asking questions about the client's life, for example, school. This part of the session is in accordance with the MI process of engagement, in which the context for a collaborative discussion is established [39] and therefore provides an early opportunity for rapport building. The second quartile (Q2) involves a discussion about the role of alcohol in the client's life and

[1] https://www.speechmatics.com/

their experiences around drinking. The third quartile (Q3) focuses on personalized feedback [38, 58, 59], during which the client is provided information about personal alcohol use and consequences, *e.g.* statistics and quantitative assessment of drinking behaviors of the client compared to their peers. In the final quartile (Q4), the therapist initiates a collaborative conversation about a personalized plan for change. This segment of the session is where the therapist can best express empathy, as one uses the knowledge from the earlier parts of the sessions to evoke and strengthen client's commitment to change [3]. Previous clinical studies also support the use of the final quartile for empathy analysis [7, 14, 33, 43].

Motivated by this session structure, we aim to explore how our models can estimate empathy levels by focusing on individual segments. We are specifically interested in Q2, since the topic of discussion may be an opportunity for the therapist to build an empathetic relationship, and Q4 based on evidence from the literature for its saliency regarding empathy and behavioral outcome. The importance of Q2 in estimating empathy has also been suggested in recent work [55]. We compare results from these two quartiles with predictions using the whole session. Since we do not have precise annotations of the start or end points of each stage in the sessions (in addition to the fact that the session might be more fluid without concrete stage boundaries), we divide the session into four quartiles by time. Quartile level analysis is consistent with previous work [6], *i.e.* a thin "slice" of contiguous segments in the session (of approximately 10 minutes) is sufficient to give MI fidelity assessment similar to human annotators. Following this finding, we use a window of 20 consecutive turns as the input to our model.

4 METHOD

In this section, we formulate the problem of multimodal empathy prediction (Section 4.1). We then describe the machine learning models (Section 4.2).

4.1 Problem Formulation

Let S be an MI session with a binary empathy label y; $y = 1$ for high empathy and 0 otherwise. S consists of N turns x_i, $i \in [1, N]$. Each turn x_i has its two modalities $x_{i,t}$ and $x_{i,a}$, denoting text and audio, respectively, and an optional learnable speaker representation $x_{i,k}$. Given a quartile (or the full session) consisting of turns $[x_i, x_{i+1}, \dots x_{i+M}]$, where M is the total number of turns in the segment ($M \leq N$), we extract overlapping context windows of fixed size W, hop (*i.e.* shift) factor P, obtaining $\lfloor M/P \rfloor$ (sub)samples of length W turns. Each subsample

$$\begin{aligned} X_i &= [x_i, x_{i+1}, \dots x_{i+W-1}], \\ X_{i+P} &= [x_{i+P}, x_{i+P+1}, \dots x_{i+P+W-1}], \\ &\dots \end{aligned}$$

then receives the same empathy label of the session S, *i.e.* $y_i = y_{i+P} = \ldots = y_S$. In our experiments, we set $W = 20$ and $P = 10$. We have also experimented with $W = 10, P = 5$ and $W = 40, P = 20$; the results were not significantly different. Hence, we fixed these hyperparameters as $W = 20$ and $P = 10$. Our model is trained to learn a function $\mathrm{F}(\cdot)$ to predict binary empathy labels y given an instance X_i, *i.e.* $\hat{y} = \mathrm{F}(X_i)$. The model $\mathrm{F}(X_i)$ consists of feature encoding modules for each modality and a fusion function to combine the modalities as described below.

4.2 Model

Figure 2 gives an overview of our model architectures. The model consists of separate encoders for each modality (text or audio) and an optional encoder for indicating the speaker of a turn (when both client and therapist talks are used). We considered both early and late multimodal fusion schemes.

Each data instance is a sequence of $W = 20$ turns in the Q2, Q4, or the whole session, using turns from only the therapist or both client and therapist. The encoders for each modality operate on the turn level, and a sequence of turn-level features are fed to a recurrent neural network, specifically Gated Recurrent Unit (GRU) [9] to obtain the sequence-level representation, *i.e.* the last hidden state of the GRU layer.

4.2.1 Feature Encoders.

Text. We use Distil-RoBERTa-Emotion [21] as the text encoder to extract the turn embeddings. Distil-RoBERTa-emotion is a distilled version of RoBERTa [29] which has been pretrained on multiple emotion datasets [12, 41, 47, 50, 52]. Our early experiments show that the distilled version has comparable performance with the original RoBERTa model for this task, likely due to our small dataset size. Additionally, the pre-trained language encoder fine-tuned on emotion recognition can improve empathy estimation due to its affect-related nature. For each turn, the input utterance is passed through the text encoder to extract the [CLS] token (the first hidden state) of the output as the utterance level feature $f_t \in \mathbb{R}^{W \times d_t}$, where $W = 20$ turns in the sequence, and $d_t = 768$ as standard BERT-* encodings.

Audio. We extract the acoustic features for each turn using HuBERT [26]. HuBERT is the state-of-the-art method for speech representation learning, pre-trained in a self-supervised manner. Specifically, HuBERT first generates pseudo-labels for each frame by performing K-means clustering on the pre-extracted features, *i.e.* MFCCs. The model then learns in a self-supervised manner through the task of predicting pseudo-labels for randomly masked frames. For Dataset1, the mono recording sampling rate 16 kHz is consistent with HuBERT feature extractors; for Dataset2 we converted the two channels into mono by averaging samples across channels. We then resample these recordings to 16 kHz using librosa.[2] In this work, to better capture the affective content in speech, we first fine-tune a HuBERT-base model on the MSP-Podcast corpus [32]. MSP-Podcast is the largest publicly available corpus for speech emotion recognition in English, containing emotionally rich podcast segments retrieved from audio-sharing websites. The HuBERT-base model is fine-tuned with the multitask objective for valence, arousal, and dominance estimation. For each turn, the input audio is passed to the audio encoder to obtain the temporal frame-level hidden states. To summarize the frame features for each turn, we concatenate the mean- and max-pooled temporal hidden states in the feature dimension and use these as the acoustic representations $f_a \in \mathbb{R}^{W \times d_a}$, where $W = 20$ turns in the sequence, and $d_a = 2 * 768$ for the

[2] https://librosa.org/

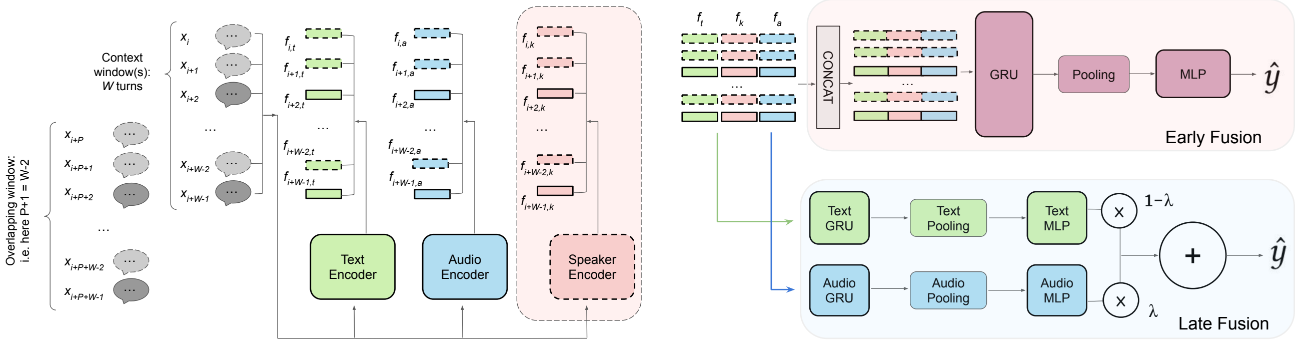

Figure 2: Model overview, including encoders and feature extraction process (left) and the fusion method (right). Light gray shade denotes client talk, and dark gray shade denotes therapist talk. Dash lines represent optional inputs, *i.e.* we studied models where both client and therapist turns are used vs. only therapist turns. Similarly, the speaker encoder block is optional.

concatenation of 768-dimensional HuBERT features.

Speaker Encoding. The optional speaker encoding module is used in settings where both therapist and client turns are included in the data, inspired by work in [23] showing improvements when the model has access to indication of turn change. Our speaker encoder is a learnable embeddings matrix that projects the therapist/client category to an embedding vector of dimension d_p. In our preliminary experiments, the models with speaker encodings slightly outperform those without, so we use speaker encodings in the rest of the experiments with both speakers; we set d_p to the same dimension as h_t and h_a.

Sequence Representation. For both text and audio modalities, we first apply a linear layer to project the features to a lower dimensional space. These features are then passed to a two-layer Bidirectional Gated Recurrent Unit (Bi-GRU) [9]. For each modality, the final representations h_t and h_a are obtained by their respective pooling modules. For the text modality, $h_t = [\text{MeanPool}(\text{Attn}(f_t)), \text{MaxPool}(\text{Attn}(f_t))]$, where Attn is a two-head self-attention module, and the outputs are mean- and max-pooled to obtain h_t. For the audio modality, $h_a = \text{FF}(f_a)$, where FF is a two-layer feedforward network operating on the last hidden state of the audio Bi-GRU.

4.2.2 *Unimodal Classifier.*

Unimodal representations from pre-trained models are used as features for our empathy classifiers. Specifically, we input the unimodal features (h_t or h_a) to a multilayer perceptron (MLP with two linear layers) and output the prediction.

$$\hat{y} = \text{MLP}(h_m), \quad m \in \{t, a\}. \tag{1}$$

4.2.3 *Multimodal Classifier.*

We experiment with early and late fusion schemes to combine the representations from both modalities.

Early Fusion. With the encoded hidden representations h_t and h_a, we concatenate them together and input the concatenated features into an MLP (two linear layers) for the final prediction.

$$h = \text{Concat}(h_t, h_a), \quad \hat{y} = \text{MLP}(h) \tag{2}$$

Late Fusion. Given the feature representations h_t and h_a, we input them into two MLPs (each has two linear layers) for separate unimodal prediction and calculate the weighted average of outputs as the final result.

$$z_m = \text{MLP}_m(h_m), \quad m \in \{t, a\}, \tag{3}$$

$$\hat{y} = (1 - \lambda) * z_t + \lambda * z_a, \tag{4}$$

λ is a learnable parameter; we look at λ in our experiments to see which modality contributes more to the prediction.

4.2.4 *Training.*

The training objective is the binary cross entropy (BCE) loss between the prediction $\hat{y}_i$ and the ground truth y_i.

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \text{BCELoss}(\hat{y}_i, y_i), \tag{5}$$

5 RESULTS AND DISCUSSION

5.1 Implementation Details

All methods are implemented in PyTorch [44]. We make all our code and model weights available.[3] At the feature encoding stage, to get better text representations, we fine-tune the last two layers of the text encoder while the audio encoder is fixed in all of our experiments since the time steps of audio are much longer than text, making it prohibitively expensive to fine-tune. The hidden dimensions for all the GRU and linear layers are 128. Each linear layer is followed by a batch normalization layer. A 10% dropout is applied to linear layers. We optimize the network weights using the AdamW optimizer [31] with a batch size of 8 on a single NVIDIA Quadro RTX8000 GPU. The weight decay is $1e^{-4}$. The gradient clipping is 1.0. We train the unimodal methods for 15 epochs with a learning rate of $1e^{-5}$. For multimodal methods, we first initialize them with the unimodal weights and then fine-tune them with 5 epochs and a learning rate of $1e^{-6}$.

[3] https://github.com/ihp-lab/mm_analysis_empathy

Table 2: Therapist prediction results in terms of Turn-level F1 score (% ↑). Results are computed from aggregated test fold predictions, per turn. "Both" means using both therapist and client turns while "Ther." denotes using therapist utterances only.

(a) Therapist-independent evaluations. Random baseline gets 48.83.

	Second quartile		Last quartile		Full session	
	Both	Ther.	Both	Ther.	Both	Ther.
Text only	57.61	61.33	56.21	57.13	56.89	58.76
Audio only	51.14	51.93	56.31	53.13	53.81	53.87
Early fusion	57.27	60.31	57.40	**58.25**	58.41	59.09
Late fusion	**60.31**	**62.30**	**58.73**	57.95	**58.94**	**60.15**

(b) Therapist-dependent evaluations. Random baseline gets 48.16.

	Second quartile		Last quartile		Full session	
	Both	Ther.	Both	Ther.	Both	Ther.
Text only	63.97	**66.91**	63.14	63.50	62.59	63.96
Audio only	59.14	59.33	60.67	58.12	62.66	61.42
Early fusion	62.29	66.49	61.57	63.51	62.95	**64.81**
Late fusion	**64.36**	66.79	**63.44**	**63.60**	**62.97**	64.53

Table 3: Therapist prediction results in terms of Voted F1 score (% ↑). Results are computed from aggregated test fold predictions. "Both" means using both therapist and client turns while "Ther." denotes using therapist utterances only.

(a) Therapist-independent evaluations. Random baseline gets 50.94.

	Second quartile		Last quartile		Full session	
	Both	Ther.	Both	Ther.	Both	Ther.
Text only	61.02	66.02	60.26	60.03	61.39	63.30
Audio only	53.15	54.32	61.04	55.53	56.73	57.18
Early fusion	62.46	65.03	64.36	**64.81**	61.36	65.45
Late fusion	**67.34**	**68.12**	**64.96**	62.46	**65.95**	**66.34**

(b) Therapist-dependent evaluations. Random baseline gets 48.52.

	Second quartile		Last quartile		Full session	
	Both	Ther.	Both	Ther.	Both	Ther.
Text only	68.06	70.12	68.38	65.71	69.42	67.89
Audio only	65.63	64.43	66.86	64.66	67.25	67.54
Early fusion	**68.23**	**70.15**	**69.29**	**72.61**	**71.09**	**69.90**
Late fusion	67.87	69.54	67.61	67.24	70.44	69.44

5.2 Experiment Setup

We use five-fold cross-validation for training and evaluation of the overall dataset, and obtain the test results by applying the model with the highest validation performance. We report the aggregated F1 score (↑) for the whole sessions, where the predictions are aggregated from the five test folds.

We perform both therapist-independent and dependent cross-validation. In therapist-independent cross-validation, sessions from the same therapist do not appear in both training and testing data. While in therapist-dependent cross-validation, the splits are not disjoint by therapists but by sessions. In each configuration, we conduct experiments using only therapist utterances and using both therapist and client utterances. Additionally, we report the results of our models when trained and tested on the second quartile, last quartile, and the whole sessions, as mentioned in 3.2.

Since the prediction is assigned to a sequence of W turns, we report two metrics: Turn F1 and Voted F1. Turn F1 is the macro F1 score computed on the turn level, *i.e.* each set of W turns in the session (or quartile) is considered a sample in computation. Voted F1 is the macro F1 score computed on the session level, where we first assign the most "voted" for in the samples to the session in consideration. The score is then computed from these voted predictions with respect to session-level empathy ground truth.

5.3 Experiment Results

5.3.1 *Classification Results.*

Tables 2a and 2b show the turn-level F1 scores of our models; Tables 3a and 3b show the Voted F1 prediction results of our models in the therapist-independent and therapist-dependent configurations, respectively. First, it is clear that the therapist-dependent setting shows better F1 scores on all settings and models. This is likely because of the variations in therapist ratings, as the 5-fold splits might have resulted in skewed label distributions. Specifically, the therapist-independent folds have a proportion of positive labels ranging from 37% to 73%, while the label distribution is more uniform in the therapist-dependent setting: all folds have 58%-60% positive labels.

Table 4: Fusion weights λ learned in our late fusion models. We report the averages of weights across five folds and their standard deviations. Top two rows show results from therapist-independent settings, bottom two for therapist-dependent settings.

		Second quartile	Last quartile	Full Session
Indep.	Both	0.21 ± 0.14	0.21 ± 0.12	0.08 ± 0.06
	Ther.	0.39 ± 0.10	0.41 ± 0.15	0.21 ± 0.36
Dep.	Both	0.08 ± 0.09	0.23 ± 0.39	$5e^{-3} \pm 5e^{-3}$
	Ther.	0.13 ± 0.13	0.39 ± 0.15	0.06 ± 0.10

Between turn-level F1 and voted F1, we observe higher scores in voted F1 across the board, which suggests that while individual data instances were classified incorrectly, in aggregate all the models are more often correct. The gap between turn-level F1 and voted F1 is larger when using both therapist and client utterances compared to when using only therapist turns, *i.e.* the differences in corresponding models for both speakers are as high as 7 points in Q2, and up to 6.5 for therapists in Q4. The largest difference is seen in Q4, *i.e.* the last quartile, suggesting that the last quartile contains information useful for empathy prediction in aggregate, especially when client talk is included. Since empathy is a dyadic construct, *i.e.* perceived empathy is observed in conversational contexts so monologues may not alone exhibit empathy.

Between unimodal-text and unimodal-audio, unimodal-text outperforms unimodal-audio in all settings except for the therapist-independent, using both speakers in the last quartile, although the

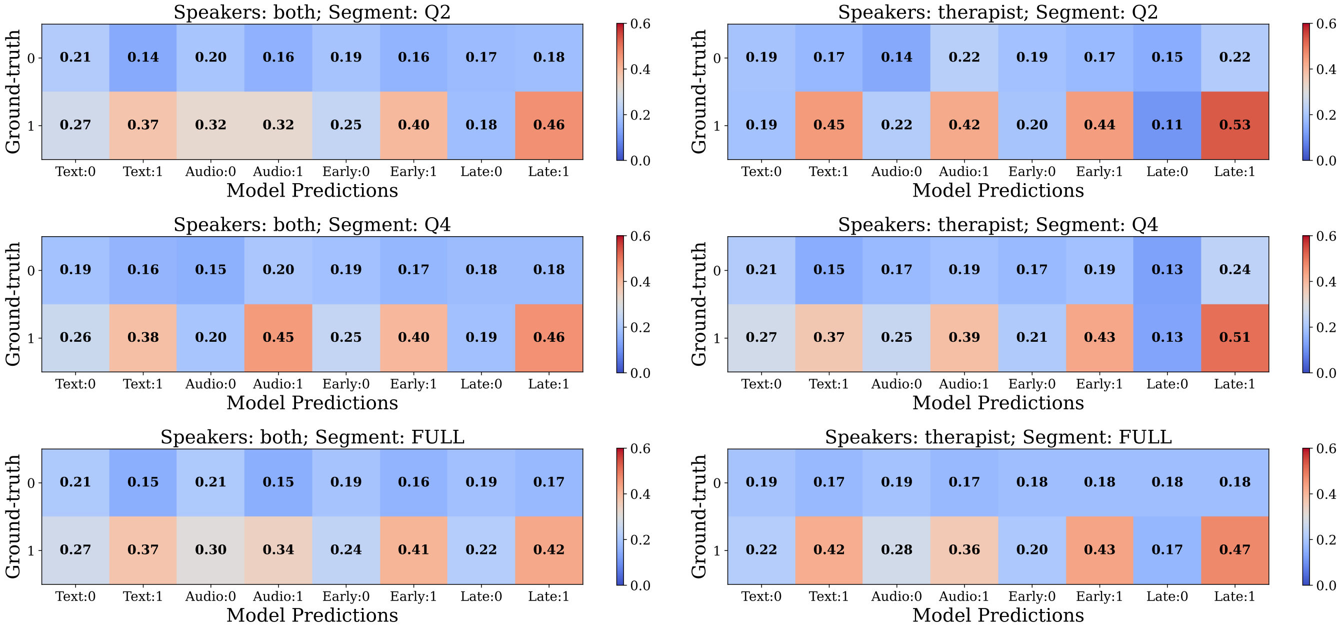

Figure 3: Confusion matrices of our models. The x-axes indicate model predictions, *e.g.* "Text:0" denotes the fraction of samples classified as 0 correctly (top row) and incorrectly (bottom row) by the unimodal-text classifier. Similarly, "Text:1" denotes the fraction of samples classified as 0 incorrectly (top row) and correctly (bottom row) by the unimodal-text classifier. "Early" and "Late" denote early and late fusion, respectively.

difference is small (56.31 for text and 56.21 for audio in terms of turn-level F1; 60.26 for text vs. 61.04 for audio in terms of Voted F1). One likely explanation for this is that the text encoder had a slight fine-tuning advantage (last two layers) while our audio encoder was used as a feature extractor only (due to computational limitations). However, while unimodal-audio underperforms unimodal-text by at least 5 points in most therapist-independent settings, this difference is narrowed in the therapist-dependent setting, where unimodal-audio is behind unimodal-text by at most 2 points; it even reaches almost parity in the full session, therapist-only setting in voted F1 and the full session, both speakers in turn-level F1.

All multimodal methods outperform unimodal ones, with a larger margin in the therapist-independent setting (by 2–6 points over unimodal-text, 4–14 points over unimodal-audio) compared to the dependent setting (by 0.1–6.9 points over unimodal-text, 2.4–7.9 points over unimodal-audio). We hypothesize this is due to the therapist effects, since it is likely that the therapist audio characteristics might be learned for common therapists in the non-disjoint splits. One exception is in the Q2, therapist-only setting (independent evaluation), where unimodal-text beats early fusion slightly (1 point for both turn-level and voted F1).

Among multimodal methods, late fusion achieves the best scores in most cases. Early fusion achieves the best voted F1 scores in the therapist-dependent evaluation. The results suggest that using the more powerful acoustic representations (HuBERT), we were able to alleviate the problem of having the language modality dominate as in previous work, *e.g.* [22, 54]. Overall, however, the discrepancies in fusion models are relatively small, with at most a 3 point difference (68.12 vs. 65.03) in therapist-independent evaluations. In therapist-dependent experiments, the difference is even smaller, with at most 1 point difference, except for the evaluations on Q4 (largest difference between 72.61 and 67.24 in the therapist-only setting). The differences between multimodal and unimodal results in the therapist-independent settings are statistically significant at $p < 0.01$, using the bootstrap test with 10,000 samples [15], as described in [2]. In the therapist-dependent settings, these differences are not statistically significant.

Regarding the predictive significance of different quartiles, using Q2 achieves the best F1 scores in all settings except for the therapist-dependent cross-validation in terms of voted F1, where using a full session seems to be more useful. As noted in Section 3.2, Q2 is where the first opportunity to show empathy occurs, so this result is consistent with the general MI structure clinicians follow. It is not clear why early fusion was the best model in this exception, though we note that the differences here are small, *i.e.* within a 1 point difference in most cases.

5.3.2 Modality Weights.

We examine how the learned modality fusion weight, λ, differs among the settings. Table 4 show these weights. First, we observe that in both therapist-independent and dependent settings, the audio modality is assigned a larger weight when using therapist-only turns than when using both speakers. More weight on the audio modality is also seen in quartiles vs. in the full session. A likely explanation for this is that audio (including the speakers' prosody and speaking styles) is more important in the limited data scenarios (individual quartiles, therapist-only) vs. when the model

T: ... and sounds to me like right now you're pretty happy but would you like to know [strategies] to avoid risks...
T: but these are just some things to keep in mind **(0 → 1)**
... ...
T: sound like you have a solid plan **(0 → 1)**
C: yeah ...
... ...
T: also they say don't take your card with you so if you want to get more you can't **(1 → 0)**
C: exactly
... ...
C: while driving that was a bad bad decision
T: the driving
T: yeah that was a bad decision **(1 → 0)**

Figure 4: Example dialog with incorrect classifications. T=therapist; C=client; blue (true → pred.) denotes misclassification by unimodal-text but correctly classified by unimodal-audio. brown (true → pred.) denotes misclassification by unimodal-audio but correctly classified by unimodal-text.

has access to larger contexts (full session, both speakers). Overall, however, this weight never exceeds 0.5, suggesting that the fusion models still rely more on the text modality.

We also look at the proportion of samples for which the early and late fusion models produce the same predictions with those by the text vs. audio modality. For early fusion, 89% of the samples receive the same classification as unimodal-text, and 59% same as unimodal-audio. Similarly, in late fusion, 89% of the samples also have the exact same prediction as unimodal-text, and 58% with unimodal audio. This result corroborates what we observe in the modality weights above.

5.3.3 Error Analysis.

Since the differences in multimodal vs. unimodal results were larger and statistically significant in the therapist-independent settings, all the following analyses are based on this setup. To understand how our models used each modality in the prediction, we looked at where a certain model fails while others do not. Figure 3 shows the confusion matrices for all models in all settings, therapist-independent evaluation. In almost all settings and models, the fraction of false negative is larger than the fraction of false positive, sometimes twice as much (*e.g.* 0.32 vs. 0.16 in Q1, both speakers for unimodal-audio and 0.30 vs. 0.15 for a full session). The difference between false positives and false negatives is similar and sometimes reversed in late fusion models, especially in the therapist-only settings.

Figure 4 shows some excerpts of the MI sessions in our data. We look at examples where one modality misclassified but another got it right. For example, in the misclassified instance by unimodal-text ("but these are just some things to keep in mind"), the utterance's language is an advice without permission (an MI inconstant code). However, the prosody of the therapist might have reflected an understanding and therefore perceived empathy, resulting in a correct classification by unimodal-audio. On the other hand, the instance misclassified by unimodal-audio as not empathetic ("sounds like you have a plan") might have gotten this prediction because the utterance is relatively short, and the audio has less information to rely on. Conversely, while the utterance "yeah that was a bad decision" by the therapist might have empathetic intonation reflected in audio, the MI code of this utterance is "confrontation," *i.e.* an MI inconsistent intent and discouraged by clinicians.

Finally, we also examine which types of MI codes are associated with correct vs. incorrect predictions. We observe the highest misclassification rates for parts of the sessions with MI inconsistent utterances (25% misclassified by all models), likely due to the complex nature of these types of intents in relation to perceived empathy. Conversely, utterances associated with reflections and facilitation have higher correct classification rates (30% correctly classified by all models). This is likely because reflections and facilitation utterances are representative of therapist empathy, both in terms of what they say (in text) and how these intents are conveyed (via speech). Looking further into the original empathy labels for these particularly difficult examples, *i.e.* where all models gave wrong predictions, we find that, surprisingly, most of these instances belong to the high-empathy label as opposed to having "ambiguous" empathy rating of 0.5. In fact, 59% of these samples have an original empathy rating higher than 0.6, and 32% have an empathy rating right at 0.5. These findings suggest that perceived high empathy remains challenging to define, and changing the task to empathy score regression is a promising future direction.

6 CONCLUSION

In this work, we present a comprehensive study of using speech transcript and audio in unimodal and multimodal binary empathy prediction. Our multimodal models outperform both unimodal models, including the text modality that often dominates learning in prior multimodal work. Our results demonstrate that the second quartile of MI sessions, *i.e.* an early phase in the conversation after introduction but before detailed strategy discussion, is the most informative segment for empathy prediction in all models, including the larger-data setting of using full sessions. In examining the learned weights in multimodal fusion, we find that a larger weight is put on the audio modality when only therapist turns are used and in shorter quartiles, suggesting that the model relies on audio more in limited-data settings.

Our error analyses suggest that our models do seem to learn and rely on important aspects of audio complementary to the text modality, *e.g.* detecting "empathetic" prosody when the transcript may suggest low empathy. In addition, we observe the highest misclassification rates for parts of the sessions with higher MI inconsistent utterances, likely due to the complex nature of these types of intents in relation to perceived empathy. Empathy is considered one of the key ingredients in establishing a good therapeutic relationship and, therefore, essential for facilitating successful behavioral outcomes. Our work contributes to a better understanding of empathy with potential applications in training therapists and assessing treatment fidelity.

ACKNOWLEDGMENTS

This work was supported by NIAAA grants R01 AA027225, R01 AA017427 and R01 AA12518. The content is solely the responsibility of the authors and does not represent the official views of the NIAAA, NIH, Dept. of Veterans Affairs, or the US Government.

REFERENCES

[1] Paul C Amrhein, William R Miller, Carolina E Yahne, Michael Palmer, and Laura Fulcher. 2003. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of consulting and clinical psychology* 71, 5 (2003), 862.

[2] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proc. Spoken Language Processing (ICSLP)*. ACL, Jeju Island, South Korea, 995–1005.

[3] Brian Borsari, Lindsey Hopkins, Jennifer Manuel, Timothy Apodaca, Nadine Mastroleo, Kristina Jackson, Molly Magill, Jerika Norona, and Kate Carey. 2019. Improvement in therapist skills over sessions in brief motivational interventions predicts client language and alcohol use outcomes. *Psychology of Addictive Behaviors* 33, 5 (2019), 484–494.

[4] Brian Borsari, John TP Hustad, Nadine R Mastroleo, Tracy O'Leary Tevyaw, Nancy P Barnett, Christopher W Kahler, Erica Eaton Short, and Peter M Monti. 2012. Addressing alcohol use and problems in mandated college students: a randomized clinical trial using stepped care. *Journal of consulting and clinical psychology* 80, 6 (2012), 1062.

[5] Doğan Can, David C Atkins, and Shrikanth S Narayanan. 2015. A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *Sixteenth Annual Conference of the International Speech Communication Association*. International Speech Communication Association, Dresden, Germany, 339–343.

[6] Derek Caperton, David Atkins, and Zac Imel. 2018. Rating motivational interviewing fidelity from thin slices. *Psychology of addictive behaviors* 32 (2018), 434–441. Issue 4.

[7] Kate Carey, Lori Scott-Sheldon, Lorra Garey, Jennifer Elliott, and Michael Carey. 2016. Alcohol interventions for mandated college students: A meta-analytic review. *Journal of Consulting & Clinical Psychology* 84, 7 (2016), 619–632.

[8] Sandeep Nallan Chakravarthula, Bo Xiao, Zac E. Imel, David C. Atkins, and Panayiotis G. Georgiou. 2015. Assessing empathy using static and dynamic behavior models based on therapist's language in addiction counseling. In *Interspeech*. International Speech Communication Association, Dresden, Germany, 668–672.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. , 9 pages.

[10] Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment* 6, 4 (1994), 284.

[11] Suzanne M Colby, Lindsay Orchowski, Molly Magill, James G Murphy, Linda A Brazil, Timothy R Apodaca, Christopher W Kahler, and Nancy P Barnett. 2018. Brief motivational intervention for underage young adult drinkers: Results from a randomized clinical trial. *Alcoholism: clinical and experimental research* 42, 7 (2018), 1342–1351.

[12] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Linda A. Dimeff, John S. Baer, Daniel R. Kivlahan, and G. Alan Marlatt. 2002. *Brief alcohol screening and intervention for college students (BASICS): A harm reduction approach.* Guilford Press, New York, NY.

[15] Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Philadelphia, PA.

[16] Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Torrey A Creed, David C Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *PloS one* 16, 10 (2021), e0258639.

[17] Nikolaos Flemotomos, Victor R Martinez, James Gibson, David C Atkins, Torrey A Creed, and Shrikanth S Narayanan. 2018. Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions.. In *Interspeech*. International Speech Communication Association, Hyderabad, India, 1908–1912.

[18] Kathryn Fokas, Jon Houck, and Barbara McCrady. 2020. Inside Alcohol Behavioral Couple Therapy (ABCT): In-session speech trajectories and drinking outcomes. *Journal of Substance Use & Addiction Treatment (JSAT)* 118 (2020), 7 pages.

[19] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. In *Advances in Neural Information Processing Systems* (Vancouver, BC, Canada), H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 3197–3208. https://proceedings.neurips.cc/paper_files/paper/2020/file/20d749bc05f47d2bd3026ce457dcfd8e-Paper.pdf

[20] James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment* 111, 2016 (2016), 21.

[21] Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base. online. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

[22] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 685–696. https://doi.org/10.18653/v1/2022.naacl-main.50

[23] Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker Turn Modeling for Dialogue Act Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2150–2157. https://doi.org/10.18653/v1/2021.findings-emnlp.185

[24] Jon Houck, Sarah Hunter, Jennifer Benson, Linda Cochrum, Lauren Rowell, and Elizabeth D'Amico. 2015. Temporal variation in facilitator and client behavior during group motivational interviewing sessions. *Psychology of Addictive Behaviors* 29, 4 (2015), 941–949.

[25] Jon M Houck, Sarah B Hunter, Jennifer G Benson, Linda L Cochrum, Lauren N Rowell, and Elizabeth J D'Amico. 2015. Temporal variation in facilitator and client behavior during group motivational interviewing sessions. *Psychology of Addictive Behaviors* 29, 4 (2015), 941.

[26] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[27] Rimita Lahiri, Md Nasir, Catherine Lord, So Hyun Kim, and Shrikanth Narayanan. 2023. A Context-Aware Computational Approach for Measuring Vocal Entrainment in Dyadic Conversations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Piscataway, NJ. https://doi.org/10.1109/ICASSP49357.2023.10095512

[28] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie (Yufan) Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. Curran, Red Hook, NY. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/37693cfc748049e45d87b8c7d8b9aacd-Paper-round1.pdf

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* 1 (2019), 13 pages.

[30] Sarah Lord, Elisa Sheng, Zac Imel, John Baer, and David Atkins. 2015. More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client. *Behavior Therapy* 46 (11 2015), 16 pages.

[31] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. OpenReview, New Orleans, LA, USA, 18 pages.

[32] Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* 10, 4 (2017), 471–483.

[33] Molly Magill, Tim Janssen, Nadine Mastroleo, Ariel Hoadley, Justin Walthers, Nancy Barnett, and Suzanne Colby. 2019. Motivational interviewing technical process and moderated relational process with underage young adult heavy drinkers. *Psychology of Addictive Behaviors* 33, 2 (2019), 128–138.

[34] Leena Mathur, Micol Spitale, Hao Xi, Jieyun Li, and Maja J Matarić. 2021. Modeling user empathy elicited by a robot storyteller. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, IEEE, Piscataway, NJ, 1–8.

[35] Scott W McQuiggan and James C Lester. 2006. Learning empathy: a data-driven framework for modeling empathetic companion agents. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. Association for Computing Machinery, New York, NY, USA, 961–968.

[36] Scott W McQuiggan and James C Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies* 65, 4 (2007), 348–360.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc., Lake Tahoe, Nevada, 9 pages.

[38] Mary Beth Miller, Thad Leffingwell, Kasey Claborn, Ellen Meier, Scott Walters, , and Clayton Neighbors. 2013. Personalized feedback interventions for college alcohol misuse: An update of Walters & Neighbors (2005). *Psychology of Addictive*

Behaviors 27, 4 (2013), 909–920.

[39] William Miller and Stephen Rollnick. 2013. *Motivational interviewing: Helping people change.* Guilford Press, New York, NY.

[40] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). , 50 pages.

[41] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, 1–17. https://doi.org/10.18653/v1/S18-1001

[42] TB Moyers, T Martin, JK Manuel, WR Miller, and D Ernst. 2010. Revised global scales: Motivational interviewing treatment integrity 3.1. 1 (MITI 3.1. 1). , 29 pages.

[43] James G. Murphy, Kathryn S. Gex, Ashley A. Dennhardt, Alex P. Miller, Susan E. O'Neill, and Brian Borsari. 2022. Beyond BASICS: A scoping review of novel intervention content to enhance the efficacy of brief alcohol interventions for emerging adults. *Psychology of Addictive Behaviors* 36, 6 (2022), 607–618.

[44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*. NeurIPS 2017, Long Beach, CA, USA, 4 pages.

[45] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[47] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 527–536. https://doi.org/10.18653/v1/P19-1050

[48] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice, France) *(AVEC '19)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/3347320.3357688

[49] Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology* 21, 2 (1957), 95.

[50] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3687–3697. https://doi.org/10.18653/v1/D18-1404

[51] Ruth Scheeffer. 1971. Toward effective counseling and psychotherapy. *Arquivos Brasileiros de Psicologia Aplicada* 23, 1 (1971), 151–152.

[52] Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology* 66, 2 (1994), 310.

[53] Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. 2019. Multimodal learning for identifying opportunities for empathetic responses. In *2019 International Conference on Multimodal Interaction*. Association for Computing Machinery, Suzhou, China, 95–104.

[54] Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal Automatic Coding of Client Behavior in Motivational Interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 406–413. https://doi.org/10.1145/3382507.3418853

[55] Leili Tavabi, Trang Tran, Brian Borsari, Joannalyn Delacruz, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2023. Therapist empathy assessment in motivational interviews. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, IEEE, Boston, MA, 1–8.

[56] Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of Behavior Classification in Motivational Interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics, Online, 110–115. https://doi.org/10.18653/v1/2021.clpsych-1.13

[57] Charles Truax. 1971. Research on certain therapist interpersonal skill in relation to process and outcome.

[58] Scott T. Walters. 2000. In Praise of Feedback: An Effective Intervention for College Students Who Are Heavy Drinkers. *Journal of American College Health* 48, 5 (2000), 235–238. https://doi.org/10.1080/07448480009599310

[59] Scott T. Walters and Clayton Neighbors. 2005. Feedback interventions for college alcohol misuse: what, why and for whom? *Journal of Addictive Behaviors* 30, 6 (2005), 1168–1182.

[60] Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *Proc. Interspeech 2014*. International Speech Communication Association, Singapore, 213–217.

[61] Bo Xiao, Dogan Can, Panayiotis Georgiou, David Atkins, and Shrikanth Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, IEEE, Hollywood, CA, USA, 1–4.

[62] Bo Xiao, Panayiotis G Georgiou, Zac E Imel, David C Atkins, and Shrikanth S Narayanan. 2013. Modeling therapist empathy and vocal entrainment in drug addiction counseling.. In *Interspeech*. International Speech Communication Association, Lyon, France, 2861–2865.

[63] Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2015. Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In *Proc. Interspeech 2015*. International Speech Communication Association, Dresden, Germany, 2489–2493. https://doi.org/10.21437/Interspeech.2015-537

[64] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/D17-1115