USC-SIPI REPORT #428

Modeling Expert Assessment of Empathy through Multimodal Signal Cues

by

Bo Xiao

May 2016

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

MODELING EXPERT ASSESSMENT OF EMPATHY THROUGH MULTIMODAL SIGNAL CUES

by

Bo Xiao

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

May 2016

Copyright 2016 Bo Xiao

Dedicated to my parents Yanmei Quan and Pingxin Xiao.

Contents

List of Tables

List of Figures

Acknowledgements

This dissertation is made possible with the encouragement and help from my family, friends, and colleagues.

Firstly, I would like to thank my adviser Dr. Shrikanth Narayanan for his advice and support. He has been an outstanding mentor and inspirer in the past six years. I would like to thank Dr. Panayiotis Georgiou for the guide on my research and countless times of valuable discussion. I also thank my other dissertation committee, Dr. C.-C. Kuo, Dr. Antonio Ortega and Dr. Gayla Margolin for their insightful comments and suggestions.

I would like to thank Dr. Brian Baucom, Dr. David Atkins, and Dr. Zac Imel for sharing their knowledge in the Psychology field, providing data for the study, and working together on research. It's my pleasure to work with these talented collaborators, and I have learned a lot from them.

It's my fortune to have amazing colleagues in the Signal Analysis and Interpretation Lab. I would like to recognize their contribution to my research work. I have greatly enjoyed interactions with Jimmy Gibson, Dogan Can, Daniel Bone, CheWei Huang, Rahul Gupta, Prasanta Kumar Ghosh, Victor Rozgic, Maarten Van Segbroeck, Nassos Katsamanis, Chi-Chun Lee, Tanaya Guha, Theodora Chaspari, Naveen Kumar, Ming Li, Kartik Audhkhasi, Zhaojun Yang, and many others.

Finally, I thank all my family and friends for their support. I thank my wife Shan Shi who made me a better person with empathy, encouragement and love. I thank my parents Yanmei Quan and Pingxin Xiao who have devoted their lifetime love and care to me. Last but not least, I want to thank everyone that I forgot to mention.

Abstract

Empathy is an important psychological process facilitating human interaction through emotional simulation, perspective taking, and emotion regulation mechanisms. Higher empathy level of the care-provider relates to better outcome of interactions in scenarios such as psychotherapy and medical care. However, traditional manual assessment of empathy is not scalable in practice, leaving the quality of services largely unknown. Computational modeling of empathy is a novel approach providing useful information to aid human decision making.

Empathy is a latent process that is difficult to measure directly. Human expert assesses empathy level through the observation of human interactive behaviors. Taking addiction counseling as an example scenario, this dissertation analyzes therapist empathy computationally based on the observed behavioral signals. Specifically, this dissertation proposes a fully automatic system to predict expert assessment of empathy based on modeling of therapist language cues. This system integrates Voice Activity Detection, Diarization, Automatic Speech Recognition, and speaker role matching modules to obtain machine generated transcripts of therapist language. It then employs Natural Language Processing methods including Maximum Entropy model, Maximum Likelihood model, and decoding lattice rescoring to estimate empathy. It finally predicts expert assessment by integrating the output of these methods.

This dissertation also proposes modeling of empathy through prosodic, speech rate entrainment, and turn-taking cues. These cues are correlated with expert assessment of empathy, including interaction session level joint distribution of a group of prosodic features; behavioral entrainment cues based on averaged turn-by-turn similarity of speech rates; and turn taking cues based on therapist and client speech ratio.

Experiments of empathy assessment prediction are conducted on audio recordings of real addiction counseling sessions in a particular treatment type named Motivational Interviewing. Results of the experiments demonstrate that the proposed automatic system and the multimodal cues can predict expert assessments of empathy in a machine-learning framework. Fusion of these cues improves the prediction accuracy. These findings suggest the feasibility of quantifying empathy via automated behavioral analysis, and may offer new insights in understanding empathy in human interactions.

Chapter 1

Introduction

1.1 Background

In this section we review the background of empathy modeling [1].

1.1.1 Definition of Empathy

Usage of the word "empathy" in the psychology literature started in 1909 with Titchener's translation of the German term "*Einfühlung*" [2] in his 1909 lecture notes on experimental psychology.

The term of empathy takes multiple interpretations. Hoffman defined it as "an affective response more appropriate to another's situation than one's own" [3], while Batson listed eight distinct phenomena that are all named empathy [4]. The discussion of empathy's definition continues in a recent summary by Cuff *et al.* [5]. Despite conceptual variations, consensus on the understanding of empathy consists of three major subprocesses [4, 6, 7], including:

(a) emotional simulation — an affective response which often entails sharing the emotional state;

(b) perspective taking — a cognitive capacity of knowing another's internal states including thoughts and feelings;

(c) emotion regulation — regulating personal distress from the other's pain to allow compassion and helping behavior.

Interdisciplinary research on empathy modeling has broadened and deepened the understanding of empathy. Preston suggested that a Perception-Action Model has the explanatory power to integrate different views of empathy into a common mechanism framework. The model states that "attended perception of the object's state automatically activates the subject's representations of the state, situation, and object, and the activation of these representations automatically primes or generates the associated autonomic and somatic responses, unless inhibited" [8]. Decety and Jackson modeled empathy as "parallel and distributed processing in a number of dissociable computational mechanisms", including shared neural representations, self-awareness, mental flexibility, and emotion regulation, which are supported by specific neural systems [6]. De Vignemont and Singer argued that empathic brain response may be contextual rather than automatic, modulated by the appraisal processes, taking into account factors such as information about the emotional stimuli, their situative context, characteristics of the empathizer and his/her relationship with the target [9].

1.1.2 Importance of Empathy

Acquired in evolution [8, 10], empathy likely serves to motivate sympathetic, helping, cooperative, and prosocial behaviors, and facilitates social communication [7, 9]. In the context of psychotherapy, Elliott *et al.* have conducted a meta-analysis that revealed an overall positive correlation of 0.31 between therapist empathy and client outcome. Thus empathy is among the most consistent predictors of psychotherapy outcome [7].

In clinical fields of oncology and general medical practice, positive correlations between empathy measures and patient outcomes have also been found in meta-analyses [11,12]. Moyers and Miller also summarized the importance of empathy in

psychotherapy, and proposed that empathic listening skills should be emphasized in hiring and training therapists [13]. Concerning whether empathy may be taught, a recent review concluded that empathy training tends to be effective in general [14].

1.1.3 Challenges

There are still important challenges in promoting empathy in clinical settings. Empathy is in part an internal mental process, which is difficult to gauge directly by observation. For example, there are four steps in each "empathy cycle" [15]: (1) client expression of experience; (2) therapist empathic resonation; (3) therapist expressing empathy; (4) client perceiving empathy, and continue to (1).

Measurement of empathy relies on human perception and subjective assessment, either by the client, the therapist, or an outside reviewer [7]. These measures vary from the true psychological process, thus being fundamentally a probabilistic estimate with associated statistical inaccuracy. They may also be biased, exacerbating the problem of coder-reliability. Human ratings also tend to be time consuming, and hence is prohibitive for large scale measurement of therapist empathy [16]. The gain of empathy from training may decay over time, while day-to-day monitoring and reinforcement of empathy by human experts is generally out of reach. In addition to being relatively slow, human ratings may not be sufficiently sensitive to capture particular nuanced and latent facets of the empathic process (*e.g.,* synchrony). As a result, research on how to decode human behaviors with respect to empathy expression, perception and action is still in its early stage, partly due to physical constrains on acquiring large amounts of data of therapist behaviors against empathy evaluations.

1.1.4 Empathy and Computation

Computational methods provide potential solutions to the aforementioned problems with scale and specificity. Recent technological advances have enabled low-cost, large scale, and widely deployable audio, visual, and physiological sensing abilities; concurrent advances in signal processing and machine learning techniques have made possible for computers to analyze complex human behaviors from vast amounts of diverse multimodal data. If automated computational methods are able to discern empathy, the advantages are clear: machines provide objective assessments and enable unconstrained sensing and computational bandwidth to support scalability.

The method of Behavioral Signal Processing (BSP) [17] provides a holistic view for the behavior modeling problem in a computational framework. It studies "measurement, analysis, and modeling of human behavior signal that are manifested in both overt and covert multimodal cues (expressions), and that are processed and used by humans explicitly or implicitly (judgments and experiences)". Following such a framework, this dissertation focuses on studying the multimodal behavioral cues that convey therapist empathy.

Figure 1.1 illustrates the general idea of the framework. Latent mental process such as empathy modulates the multimodal expressions, which are perceived and interpreted by the interlocutor. The perceived cues then influence the latent mental process of the interlocutor. The behavioral cues in the expressions are also perceived by the human expert being an observer. Computational modeling of these cues underpins automatic assessment of empathy; and human expert assessments are employed to train the computational model as well as to examine the outcome of the automatic processing. Finally, automatic processing provides feedback to human expert, and produces behavioral informatics about the interaction.

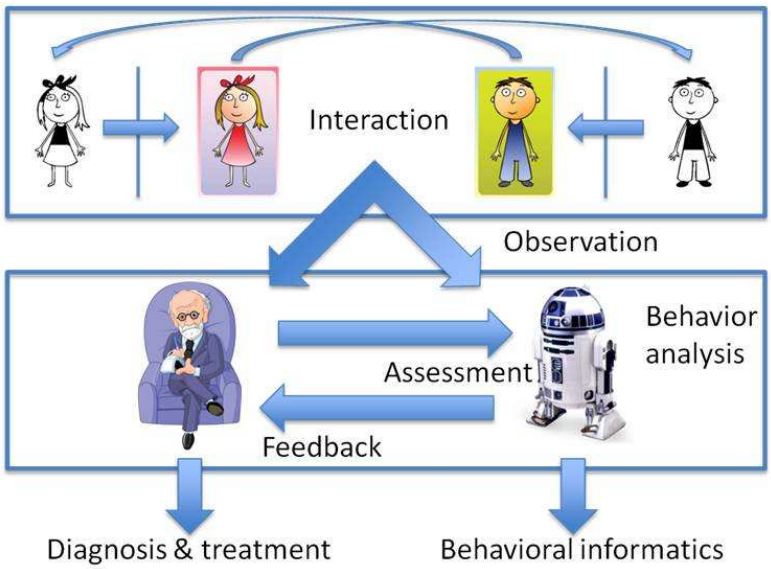

Figure 1.1: Illustration of the general framework of Behavioral Signal Processing

1.2 Dissertation Overview

This dissertation proposes modeling multimodal behavioral cues to predict expert assessment of therapist empathy. It models mainly three types of behavioral cues including prosodic, lexical, and speech rate entrainment cues.

1.2.1 Prosodic Cues

Prosody refers to the intonation of speech rather than the verbal content. It describes "how one says it" instead of "what one says", conveying rich emotional and communicative cues. Several types of prosodic features are extracted from the speech signal including energy, pitch, duration, jitter, and shimmer. These features depict the property of prosody in short intervals. The research question is to map time streams of prosodic features to session level assessment.

This dissertation proposes quantizing prosodic cues into three levels based on the averages in speech segments [18]. The quantization transforms real valued

features into discrete levels that are easier to model and interpret. The speech segments serve as cognitively coherent units of expression. Such constrains on time and feature range enable an analysis of session level joint distributions of prosodic cues. The joint distributions, representing an overall property of prosodic cues, are examined for their relation to empathy. Such a modeling approach allows backward interpretation of the findings, which may be pointing to certain meaningful categories of prosodic patterns. For example, experiment shows medium duration, high pitch, and high energy segments by the therapist are linked to lower empathy.

1.2.2 Lexical Cues in the "Sound to Code" System

Lexical cues are modeled through the property of language use by the therapist. Compared to prosody, language is more structured and encodes more abstract semantic information. This dissertation proposes employing competing language models of high *vs.* low empathy in three different methods [19].

The first method uses Maximum Entropy model to formulate the posterior of empathy for each speech utterance. N-grams in high *vs.* low empathy utterances are used as feature functions. Model parameters are optimized based on the training data. The second method uses Maximum Likelihood language model of high *vs.* low empathy. Posterior of empathy is derived based on the likelihoods following the Bayesian theorem. In addition, when the language is derived from Automatic Speech Recognition, high *vs.* low empathy language model rescoring are applied to the decoding lattice, in order to raise empathy relevant words in the lattice, which may not appear in the original best path due to the lower likelihood based on a generic language model. Scores of the averaged N-best paths from the rescored high *vs.* low empathy lattices are used as features indicating empathy.

These methods evaluate empathy on the utterance level. In the experiment only session level high *vs.* low empathy labels are available. One solution is to use all utterances of therapist in high empathy sessions as positive samples, and *vice versa*. In the testing case, utterance level scores of empathy are averaged to derive session level prediction.

In practice, obtaining therapist language by human transcription is a costly process. An automatic system directly taking audio recording as the only input would enable large scale processing of psychotherapy. The system outputs an empathy assessment in lieu of an empathy code given by a human coder. A prototype system is proposed connecting several speech processing front-end modules including Voice Activity Detection (speech *vs.* non-speech), Speaker Diarization (group speech segments by the same speaker), Automatic Speech Recognition (transcription), and speaker role matching(therapist or patient). The decoded therapist language is then used to predict empathy assessment.

1.2.3 Speech Rate Entrainment

Entrainment refers to the phenomena that behaviors of interlocutors become more similar or coordinated as the interaction proceeds. It is a psychological process closely tied to empathy, following the theories of Perception-Action-Link and mirror neurons. Clinical evidences show that stronger empathy relates to more prominent entrainment.

Entrainment is manifested through multimodal behaviors. This dissertation proposes quantifying entrainment from one aspect — speech rates of the interlocutors [20]. ASR-derived forced speech-text alignment provides word level time marks. Speech rates are then computed as the count of words, syllables, or phonemes in a speech segment divided by its time duration. Experiment results

lend support to the hypotheses that empathy correlates to the averaged turn-by-turn absolute difference of speech rates between therapist and client. Larger speech rate difference is associated with lower empathy.

As another aspect of timing in interaction, turn taking is also modulated by the mental processes. Turn taking cues such as the time ratio of therapist and client speech correlate to empathy. Pause (*i.e.,* intra-speaker silence) and gap (*i.e.,* inter-speaker silence) time ratios also reflect therapist empathy. For example, experiments show that therapists tend to speak less when they show more empathy to the client, which may be the case that they are able to invoke more client talk through expressing empathy.

1.2.4 Multimodal Empathy Modeling

The above multimodal cues and their fusion are tested in experiments to classify high *vs.* low empathy. Given limited data, the experiments are conducted in a leave-one-therapist-out cross validation. The results demonstrate the feasibility of quantifying therapist empathy through signal processing of the multimodal cues. They also show that the integration of multiple features improves the classification accuracy.

The rest of the dissertation is organized as follows. Chapter 2 summarizes related work on empathy modeling. Chapter 3 explains prosody modeling in more detail. Chapter 4 introduces the "sound to code" system, its various sub-modules, and the empathy detection algorithms using language modeling approaches. Chapter 5 examines the relation of speech rate entrainment and empathy through hypotheses testing. Chapter 6 concludes the dissertation with remarks on future directions.

Chapter 2

Related Work

In behavioral studies of empathy, human raters (who are often external to the interaction and data generation setting) typically use behavioral cues of the target to infer and annotate whether a particular empathic process has occurred (*e.g.,* a group of behavioral cues proposed by Riess [21]). Regenbogen *et al.* have examined the utility of three behavioral channels (facial expressions, prosody and speech content) towards emotional recognition and response via "neutralizing" one channel and testing the differential effect on empathic responses. The study showed that all three channels contributed to empathic responses [22]. This suggests that an observer may have employed information from the above channels to draw an conclusion of the therapist's empathy. Still, this process of empathy evaluation is challenging and non-scalable; computational methods may provide a useful alternative. Like manual evaluation, computational empathy analysis studies how to capture and model multimodal behavioral cues for detecting empathy.

Two kinds of research methodologies are commonly applied:

- **Feature analysis** — finding behavioral cues that correlate with human annotator-derived empathy ratings through statistical analyses, a common method in behavioral sciences.

- **Prediction** — data driven computational learning of models using machine learning techniques that serve as functions mapping automatically measured behavioral cues to empathy ratings. The performance of the automated prediction is typically evaluated by comparing machine assessments against

human expert ratings on new or held-out interactions not seen in model construction [23].

The standard in clinical psychology and psychiatry is to build and evaluate models in a complete dataset (*e.g.,* to fit a regression model with various correlates of empathy). In engineering approaches, *prediction* is a much stronger test than *correlation*. It partitions data into mutually exclusive training and evaluation sets to establish validity and generalizability of results.

As an emerging field, computational empathy analysis has been pursued most notably in two domains. Firstly, in addiction counseling using Motivational Interviewing (MI) [24], empathy is a key index for treatment fidelity [25]. Human experts use the *Motivational Interviewing Treatment Integrity* (MITI) manual [26] to code the degree of therapist empathy in an interaction on a Likert scale. MITI defines empathy as "the extent to which the clinician understands or makes an effort to grasp the client's perspective and feelings", emphasizing the cognitive component of empathy.

Secondly, in four-person casual conversations the researchers operationally defined empathy as emotion contagion [27], emphasizing the affective component of empathy. Human coders marked the empathy states of each pair of interlocutors on the time line.

Though in its early stage, computational empathy analysis has examined a number of multimodal behavioral cues. In addition, *entrainment* (synchrony) — an interaction process wherein behaviors of interlocutors becoming more similar or coordinated — is a phenomenon that is tied closely to empathy, based on the theory of Perception-Action Link and the function of mirror neurons [8, 10, 28]. Modeling entrainment across various modalities serves as an indirect but useful mechanism for quantifying empathy.

Other related studies focused on empathy synthesis, *i.e.,* designing Embodied Computer Agent (ECA) that can simulate human empathic behavior [29–32].

2.1 Lexical Cues

Spoken language encodes a multitude of information including a speaker's intent, emotions, desires as well as other physical, cognitive and mental state and traits (*e.g.,* speaker age and gender). By analyzing the language transcripts of interactions we may infer the empathy processes that are driving, and reflected in, the language expressions (*e.g.,* qualitative findings on empathic word use by Coulehan *et al.* [33]).

Xiao *et al.* have used N-gram Language Models [34] of empathic *vs.* other (background) utterances of the therapists in MI type counseling [35]. They showed that a Maximum Likelihood classifier based on these language models were useful to automatically identify empathic utterances. Further, utterance level evidences of empathy can be summed to derive measures that can better correlate with interaction session level empathy ratings (*i.e.,* MITI codes).

Extending this work, Chakravarthula *et al.* proposed a model that considers the therapist's likelihood to transition among high *vs.* low empathy states over time using a Hidden Markov Model [36], instead of assuming a static state of empathy throughout the interaction [37]. They showed that the dynamic model provided improved predictions of the session level assessments offered by human experts compared to the static model while providing short-term empathy information.

The above N-gram language model based methods do not exploit the semantic meaning of words. Linguistic features such as those generated by the *Linguistic Inquiry and Word Count* (LIWC) software [38] associate words with categories of

various psychological processes, personal concerns, spoken categories, *etc.* Moreover, novel computational methods afford affective text analyses to be applied broadly beyond words specified in the lexica [39]. Computational Psycholinguistic Norms (PN) [39] further expand the ability to include both affect states and word's relation to additional cognitive processes (*e.g.,* age of acquisition, imageability, gender ladenness). Gibson *et al.* compared LIWC and PN features to N-gram features in predicting therapist empathy ratings, showing that though N-gram features performed the best, LIWC and PN features provided complementary information resulting in boosted prediction performance by feature fusion [40].

The above methods investigate language cues that directly correlate with and can predict empathy. Although these cues appear to be effective, their ties to psychological theories about empathy largely remain implicit. On the other hand, analysis of language style synchrony investigates one possible realization of the perception-action link. Lord *et al.* extracted LIWC features on each speaking turn of the therapist/client, and quantified if the same category of words appeared both in the therapist's turn and the client's turn [41]. As a result, they found 11 word categories that associated with stronger synchrony in high empathy sessions. Language style synchrony has even stronger correlation to empathy than the well accepted traditional indicator — count of *reflections* by the therapist.

2.2 Vocal Cues

Human vocal expression is highly dependent on internal state, and as such it is linked to empathy. This has been supported by diverse work: *e.g.,* brain areas important for prosodic mechanisms are linked to empathic ability [42], and empirically prosodic continuity (*e.g.,* therapist continued the intonation/rhythm of the

client's preceding turn) by the therapist has been associated with higher empathy [43].

Interlocutor vocal entrainment serves as an indirect feature for empathy. Imel *et al.* investigated vocal entrainment through the correlation of mean fundamental frequencies (pitch) [44] between interacting therapist and standardized patient (SP) [45]. They found strong correlation (0.71) that did not exist in fake interactions with random pairings of therapists and SPs. Moreover, this correlation was higher in high empathy sessions compared to low empathy ones, demonstrating the link between entrainment and empathy.

Xiao *et al.* modeled entrainment with a more detailed measure of acoustic similarity [46]. They extracted MFCC, *i.e.,* Mel-Frequency Cepstrum Coefficients [44], and pitch features from the speech of interacting therapists and SPs. These features defined the Principal Component Analysis (PCA [47]) spaces of the therapist/SP. Kullback-Leiber divergence (KLD [48]) was employed to compute the similarity of PCA components — one's own PCA space and the other's that is mapped to the former. They found significant correlation between statistics of turn-level KLDs and human specified empathy ratings.

2.3 Facial Expression and Reaction Timing Cues

Facial expressions also carry rich emotional information [49]. Kumano *et al.* investigated if the co-occurrence of facial expression patterns amongst the interlocutors could predict the empathy labels [50]. They discretized facial expressions into six types, and modeled empathy state in three classes as *empathy*, *unconcern*, and *antipathy*. A Dynamic Bayesian Network model [51] was constructed to associate empathy states with facial expressions and gaze directions along time. Automatic

recognition of facial expressions was compared with manual labeling. Experiment results showed that facial expressions were effective predictors of empathy labels.

Kumano *et al.* extended this framework by investigating reaction timing and facial expression congruence information [52]. They demonstrated that these two aspects were related to the annotated empathy labels. For example, a congruent but delayed reaction in facial expression is less likely to have an empathy label. By further incorporating annotations of head gesture types, they improved the accuracy of empathy state prediction.

Moreover, Kumano *et al.* studied the inference of empathy labels by multiple human annotators [53]. Instead of assigning one class label for empathy, they estimated the distribution of empathy labels by a group of evaluators. They found that training the model with multiple annotations outperformed training with only the majority-voted empathy labels.

Chapter 3

Modeling Empathy through Prosodic Cues

3.1 Introduction

In this Chapter, we build computational models to analyze the relation of *prosodic* cues and therapist empathy (as perceived by human experts) in drug addiction counseling. Prosody refers to the non-verbal part of speech, such as intonation, volume, and other voice quality factors, which account for "how one says" rather than "what one says".

Neurology studies have showed not only that the production and perception of prosody share the same brain area, but also that this area is related to affective empathy [42]. Psychology study found empirically that prosodic continuity (defined as continued intonation/rhythm of the client's preceding turn, and produced with a lower and/or quieter voice and with narrower pitch span) by the therapist points to higher empathy; whereas prosodic disjuncture (therapist evaluated or challenged the client's emotional descriptions and voice was higher and/or louder and the pitch span wider than in the client's previous turn) points to the opposite [43]. Correlation between the therapist's and the client's mean pitch values is higher in high empathy sessions [45].

Thus, past works have proved prosodic cues as indicators of empathy, but have yet to include a robust analysis of prosodic feature toward automatic prediction of

empathy. Toward this end, in this Chapter we consider five dimensions of prosodic features: pitch, vocal energy, jitter, shimmer, and utterance duration (a result of conversational factors and speaking rate). Pitch and vocal energy are integral to intonation. Jitter and shimmer — measures of short-term variation in pitch period duration and amplitude, respectively — are acoustic correlates of atypical voice quality attributes including breathiness, hoarseness, and roughness [54]. In addition to empathy, these prosodic features can capture important behavioral cues in various domains [55, 56].

We describe the addiction counseling dataset and the annotation of therapist empathy in Sec. 3.2. We explain the prosodic features as well as the extraction and normalization in Sec. 3.3. For robustness and generalization, we quantize each prosody feature into three levels, and analyze the values on the unit of speaking utterances. This allows us to characterize the pattern of an utterance with a single or multiple prosodic features, and compute the distribution of various types of utterances in a session, as described in Sec. 3.4. We examine the relation between these distributions and therapist empathy, and attempt to capture salient prosodic patterns; we then carry out the prediction of "high" or "low" empathy of the therapist using the captured patterns in experiments in Sec. 3.5. We discuss the results in Sec. 3.6 and conclude this Chapter in Sec. 3.7. An overview of the modeling approach is illustrated in Figure 3.1.

3.2 Dataset

For the experiments in this Chapter, we use the data from a counselor training study that follows the Motivational Interviewing (MI) counseling approach [57]. MI is a style of counseling focused on helping people to resolve ambivalence and

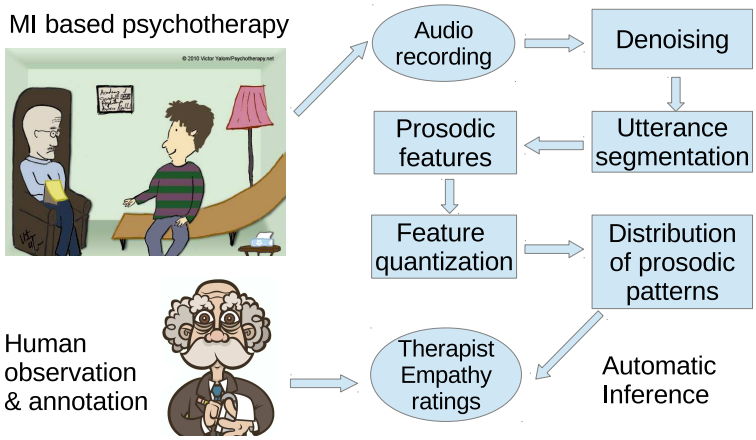

Figure 3.1: Overview of prosodic modeling of therapist empathy.

emphasizing the intrinsic motivation of changing addictive behaviors. Therapist empathy is hypothesized to be one of the key drivers of change in patients receiving MI [58]. In the above study, 144 therapists serving in the community participated at the beginning, and 123 of them completed the entire process. Three researchers acted as *Standardized Patients* (SP), *i.e.,* taking the role of clients, in about half of all the counseling sessions recorded. The rest of the sessions involved real clients. Each interaction session is roughly 20 minutes long, recorded with a single channel far field microphone. At collection time the intended consumers were human annotators, and as such the audio quality is challenging for machine processing.

Three human coders reviewed the recordings and assessed the performance of the therapist using a specially designed coding system, the *Motivational Interviewing Treatment Integrity* (MITI) [58]. The therapist in each session received an overall rating of empathy on a Likert scale (discrete) from 1 to 7. Inter-coder reliability assessed via *Intra-Class Correlation* (ICC) had a mean of 0.67±0.16, while ICC for the same coder over time had a mean of 0.79±0.13. Correlation of the empathy scores given at the first and second time is 0.87, based on all 182 sessions that were coded twice. No session was triple-coded.

In this Chapter we employ 117 sessions that involve a SP and from 91 different therapists, with empathy ratings on the two extremes (mean value if double coded). From the 117 sessions, 71 have high-empathy scores with range 5∼7 and mean 6.05±0.65, while 46 sessions have low-empathy scores with range 1∼3.5 and mean 2.17±0.57. Since only overall ratings of empathy are available rather than localized labels for empathic events, we choose sessions on the extremes where empathic/non-empathic behaviors are more frequent and prominent, and thus binarize our data. The above sessions are manually diarized into therapist's speech and client's speech separately.

3.3 Prosodic Feature Extraction

3.3.1 Audio Preprocessing

We first apply speech enhancement to reduce noise in the audio recordings due to the challenging audio quality. We adopt the approach of minimum Mean-Square-Error estimation of spectral amplitude [59] for denoising, implemented in the *Voicebox* speech processing toolbox [60]. The effectiveness of noise reduction was empirically confirmed on a few sessions.

The sessions were manually annotated for speakers; however, the segmentation boundaries were not precisely aligned with speech onsets or offsets, and pauses within the same speaker were not marked out. Therefore, we exploit our previously designed Voice Activity Detection (VAD) system to finely segment the audio into speech utterances [61]. The VAD system is based on a number of robust speech features with Neural Network learning. In this Chapter, we train the model on 10 sessions of Motivational Interviewing which were manually segmented and are disjoint to the data we use for prosody analysis. During decoding the VAD outputs

a probability measure for the presence of speech over time with a value that varies between 0 (non-speech) and 1 (speech). We empirically set a high threshold equal to 0.8.

We break a speech segment belonging to a single speaker if a pause inside the segment is longer than 0.2 seconds, otherwise we consider it as a single continuous segment. We also set a threshold for minimum duration of speech segment as 0.5 seconds; therefore detected speech of less than that was assigned as non-speech. No lower bound is set for the gap between speakers due to probable interruptions. However, we ignore speech regions that are labeled as overlapped speech, since they cannot represent the prosodic properties of a single speaker.

We denote the resultant sequence of speech utterances in a session as U_n, $n = 1, 2, \cdots, N$, where N is the total number of utterances. Let $r_n \in \{\texttt{Therapist(T)}, \texttt{Patient(P)}\}$ be the speaker of U_n. Let d_n in seconds be the time duration of U_n.

3.3.2 Pitch and Jitter

We compute pitch using the method in [62] that is inspired by the subharmonic summation proposed in [63]. We suppress doubling and halving errors through dynamic programming. Pitch values are confined to the frequency range 50-800 Hz and are computed on a 30 ms window with a 10 ms shift. In order to reduce interference, we compute pitch values separately for the two interlocutors. We further prune the pitch against doubling/halving errors and other noises, respectively for the therapist and patient by the following two steps: (1) Find the central pitch p_0 for the speaker as the mode of the pitch values $p(t)$. (2) Discard the pitch value if $p(t) > 1.5p_0$ or $p(t) < p_0/1.5$ (symmetric in log domain). We observed that in average the pruning removed 6% pitch values in time.

Let $\overline{p_T}$ be the mean pitch after pruning for the therapist in a session. For each utterance U_n, $r_n = \mathtt{T}$ we obtain the mean-normalized log pitch feature as in (3.1):

$$p_n = \frac{1}{K}\sum_{t_n=1}^{K} \mathtt{log}\frac{p(t_n)}{\overline{p_\mathtt{T}}}, \tag{3.1}$$

where t_n is the acoustic frame index within the time span of U_n.

We denote $g(t_n)$ the reciprocal of $p(t_n)$, *i.e.,* the fundamental period of the glottal pulse. Based on extracted pitch values, we approximate relative jitter values $\tilde{j}_n$, *i.e.,* normalized by the average fundamental period, for U_n as in (3.2)$\sim$(3.3):

$$\tilde{j}_n = \frac{1}{K-1}\sum_{t_n=2}^{K}\left|\frac{g(t_n)-g(t_n-1)}{\overline{g_\mathtt{T}}}\right| \tag{3.2}$$

$$= \frac{\overline{p_\mathtt{T}}}{K-1}\sum_{t_n=2}^{K}\left|\frac{1}{p(t_n)}-\frac{1}{p(t_n-1)}\right| \tag{3.3}$$

Moreover, we compute the averaged relative jitter $\overline{j_\mathtt{T}}$ for the therapist in the entire session (accumulating all therapist utterances) by applying (3.3), as the individual baseline for jitter. Finally, we define the normalized jitter feature $j_n = \tilde{j}_n - \overline{j_\mathtt{T}}$ for U_n. We obtain the pitch and jitter features for patient utterances in the same way.

3.3.3 Vocal Energy and Shimmer

We compute short time vocal energy over a 300 ms window with 10 ms shift as the mean-squared value of speech signal. We denote the log scale of the energy as $e(t)$. Due to the variations of microphone gain and speaker-to-microphone distance, it is necessary to normalize the energy for each interlocutor. Let the mean and variance

of the therapist's energy be μ_T and σ_T^2. We define the vocal energy feature e_n for U_n, $r_n = \texttt{T}$ as in (3.4):

$$e_n = \frac{1}{K} \sum_{t_n=1}^{K} \frac{e(t_n) - \mu_T}{\sigma_T}, \tag{3.4}$$

where t_n is the acoustic frame index within the time span of U_n.

We compute the averaged difference of $e(t_n)$ as shimmer value $\tilde{s}_n$ for U_n, as in (3.5):

$$\tilde{s}_n = \frac{1}{K-1} \sum_{t_n=2}^{K} \left| \frac{e(t_n) - e(t_n - 1)}{\sigma_T} \right| \tag{3.5}$$

Moreover, we compute the averaged shimmer $\overline{s_{\texttt{T}}}$ as an individual baseline for the therapist by applying (3.5) over the accumulated speech signal of the therapist in the entire session. We finally define the normalized shimmer feature as $s_n = \tilde{s}_n - \overline{s_{\texttt{T}}}$ for U_n.

We obtain the vocal energy and shimmer features for the patient in a similar way. In summary, $(d_n, p_n, j_n, e_n, s_n)$ is the five-dimensional prosodic feature for U_n.

3.4 Modeling Prosodic Features

3.4.1 Feature Quantization

We quantize each prosodic feature into Q equally populated intervals, for the therapist and the patient separately. We find boundaries of the intervals on aggregated training samples of utterances from multiple sessions involving different therapists and patients. Such aggregate quantization is applicable due to the normalization and subtraction of individual baselines. Note that the disparities of feature distributions still exist in different sessions, hence the equally populated quantization

does not imply that the quantized features are uniformly distributed in each session. Unseen utterances (test set) can be quantized with the same boundaries obtained on the training set.

Taking $Q = 3$ for the therapist utterances for example, we quantize each feature by its 33 and 67 percentile into discrete values. These discrete bins conceptually represent low, medium and high values for each feature dimension. Similarly we carry out the quantization for patient utterances.

3.4.2 Distribution of Prosodic Patterns

We denote the quantized feature values as $(\hat{d}_n, \hat{p}_n, \hat{j}_n, \hat{e}_n, \hat{s}_n)$ for utterance U_n. We compute the joint distributions of $P_U(r_n, F_n)$ and $P_U(r_n, F_n, r_{n+1}, F_{n+1})$, where r_n is binary in Therapist or Patient, *i.e.,* $r_n \in \{\mathtt{T}, \mathtt{P}\}$, and F_n can be any combination drawn from the five quantized prosodic features. Because of speech segmentation and quantization of the feature set, there are integer counts of utterances in each pattern and finite types of prosodic patterns. We count the occurrences of each discrete pattern of (r_n, F_n) and $(r_n, F_n, r_{n+1}, F_{n+1})$, and divide by the total number of segments. The above probabilistic model is akin to a maximum likelihood "bag-of-words" model.

Specifically, we consider the following feature combinations in $P_U(r_n, F_n)$: (1) $F_n = f_n^1$ where f_n^1 is one of the five prosodic features. (2) $F_n = (f_n^1, f_n^2)$ where (f_n^1, f_n^2) is any combination of two features. (3) $F_n = (f_n^1, f_n^2, f_n^3)$ where (f_n^1, f_n^2, f_n^3) is any combination of three features. For $P_U(r_n, F_n, r_{n+1}, F_{n+1})$, we set $F_n = f_n^1$, $F_{n+1} = f_{n+1}^1$, *i.e.,* a single feature out of the five prosodic features. For the robustness of probability estimation, we do not incorporate more complex prosodic patterns (*e.g.,* combination of more features) due to the limit of samples (speech segments) in each session.

We consider the joint rather than conditional probability with respect to the speaker, according to the previous finding that therapist empathy is correlated with the ratio of therapist's speech [46]. The total dimension of different probability entries is given in (3.6) ($\mathtt{C}_m^n$ represents combinatorial function), which equals 930 in case of $Q = 3$. Note that these probability entries can also be viewed as the frequencies of occurrence for different prosodic patterns; we examine the relation of therapist empathy and these probabilities in the experiments.

$$2(Q\mathtt{C}_5^1 + Q^2\mathtt{C}_5^2 + Q^3\mathtt{C}_5^3) + (2Q)^2\mathtt{C}_5^1 \tag{3.6}$$

3.5 Experiment and Results

3.5.1 Correlation of Therapist Empathy and Prosody

For the analysis of correlation between therapist empathy and prosody, we extract prosodic features in each session and derive the quantization of $Q = 3$ as well as sessions-wise distribution P_U over the entire dataset. We will discuss the choice of Q in Sec. 3.6.

The coded therapist empathy rating E, as introduced in Sec. 3.2, is in the range of 1 to 7. We compute the Pearson's correlation ρ between E and elements of P_U, and test the significance using Student's t-distribution. In Table 3.1 we report some of the most prominent prosodic patterns associated positively and negatively with E. We can see that high pitch and energy are negatively associated with therapist empathy; this is consistent with the empirical findings from psychology literature *e.g.,* [43]. We discuss the results further in Sec. 3.6.

Table 3.1: *Prominent prosodic patterns for correlations ρ between E and P_U: T — Therapist, P — Patient, L — Low, M — Medium, H — High*

r_n	f_n^1	f_n^2	f_n^3	ρ	p-value
T	$\hat{d}_n$ =M	$\hat{p}_n$ =H	$\hat{e}_n$ =H	-0.47	8×10^{-8}
T	$\hat{d}_n$ =M	$\hat{p}_n$ =H	—	-0.42	2×10^{-6}
T	$\hat{d}_n$ =M	$\hat{e}_n$ =H	$\hat{s}_n$ =M	-0.41	4×10^{-6}
T	$\hat{d}_n$ =M	$\hat{p}_n$ =H	$\hat{j}_n$ =M	-0.41	5×10^{-6}
...				...	

r_n	f_n^1	r_{n+1}	f_{n+1}^1	ρ	p-value
T	$\hat{e}_n$=M	T	$\hat{e}_{n+1}$=M	-0.40	7×10^{-6}
T	$\hat{j}_n$=M	T	$\hat{j}_{n+1}$=H	-0.34	2×10^{-4}
P	$\hat{d}_n$=H	T	$\hat{d}_{n+1}$=L	0.34	2×10^{-4}
P	$\hat{p}_n$=M	P	$\hat{p}_{n+1}$=L	0.34	2×10^{-4}
...				...	

In total 51 features	$\lvert\rho\rvert > 0.3$	$p < 10^{-3}$

3.5.2 Classification of Therapist Empathy Level

We carry out leave-one-therapist-out cross-validation in prediction of the binary levels of therapist empathy $\hat{E}$ ($\hat{E} = 1$ if $E \geq 4.5$, otherwise $\hat{E} = 0$) using P_U. This means we do the following operations in each round. For training (1) determine the quantization boundaries of the prosodic features; (2) quantize using these thresholds; (3) compute P_U separately for each session; (4) train the classifier of $\hat{E}$. For testing employ the test data and (1) quantize using thresholds derived at training and compute P_U; (2) predict $\hat{E}$. We use linear Support Vector Machine (SVM) as the classifier.

For comparison, we design a baseline method for classification using functionals of prosodic features $(d_n, p_n, j_n, e_n, s_n)$ in each session, separately for the therapist and the patient utterances. This is hypothesizing that the overall empathy is reflected in the ensemble statistics of individual prosodic features. Specifically, we employ the following functionals: (1) 1, 25, 50, 75, 99 percentile; (2) range of

1∼25, 25∼50, 50∼75, 75∼99 percentile; (3) mean, variance, skewness and kurtosis of the prosodic feature. This in total derives 14 (functional) × 5 (prosody) × 2 (speaker) = 140 dimensional functional features for the SVM classifier. Note that the mean value of the prosodic features are not necessarily zero, since the normalization is applied to acoustic frames while the functional is computed over utterances. Numerically, it is equivalent to weighting shorter utterances higher, such that treating an utterance as a basic unit of expression.

We use a simple feature selection scheme to reduce complexity and avoid overfitting in the classification, by thresholding on the p-value of one-factor ANOVA [64] test (*i.e.,* a test of different mean values in two groups) on the training samples for each feature. We set the threshold to 10^{-3} for P_U, while we loosen the threshold to 10^{-2} for the baseline functionals as we observe that their significances are in general lower.

In Table 3.3 we list the classification accuracies by the different approaches with the same data and cross-validation method. The P_U features yield the best performance that is higher than chance level (and statistically significant; binomial test $p < 10^{-3}$) and higher than the result in [46] (but not statistically significant). The performance of the baseline method is higher than chance level but not statistically significant. We further discuss the results in Sec. 3.6.

3.6 Discussion

An interesting scientific question is whether the prosodic patterns of the therapist can themselves, out of contextualization of the patient behavior, provide important information regarding the therapist empathy. To address this, we compute the conditional distribution $P_U(F_n|r_n = \texttt{T})$. In comparison to the upper half of

Table 3.1, the prominent correlations ($|\rho| \geq 0.3$) between $P_U(F_n|\mathtt{T})$ and empathy are listed in Table 3.2. We can see the effect of high energy and high pitch is still negative, but the statistical significance is reduced; similarly for the other therapist prosodic patterns in Table 3.1. In addition, low energy patterns show positive correlation to empathy, which is consistent with the empirical finding [43].

Table 3.2: Prominent prosodic patterns for correlations ρ between E and $P_U(F_n|\mathtt{T})$: L — Low, M — Medium, H — High

f_n^1	f_n^2	f_n^3	ρ	p-value
$\hat{d}_n$ =M	$\hat{p}_n$ =H	$\hat{e}_n$ =H	-0.33	2×10^{-4}
$\hat{d}_n$ =L	$\hat{e}_n$ =L	$\hat{s}_n$ =H	0.31	6×10^{-4}
$\hat{e}_n$ =L	—	—	0.30	1×10^{-3}

Table 3.3: Therapist empathy $\hat{E}$ classification accuracies

Approach	Accuracy
Chance level	0.61
Vocal similarity and speech ratio [46]	0.70
Distribution of prosodic patterns P_U	0.75
Functionals of prosodic features	0.67

In Sec. 3.5.2 we find that the functionals of prosodic features are less effective to infer empathy than the distribution of prosodic patterns. The most significant correlation between the functionals and E is -0.3 by the median of therapist energy. This trend of higher energy implying lower empathy is consistent with the results by P_U; however, it is less discriminative. The quantized prosodic patterns proposed in this Chapter on the other hand, may only focus on part of the interaction. For example, the most significant pattern of (d_n =M, p_n=H, e_n=H) represents only 6% (range 1% to 15%) of therapist utterances in average. This suggests that it is important to study salient behavior patterns for high level summative behavioral characteristics like empathy. Such high level judgments are often a non-trivial

integration of local evidences, where some cues may be more important than others. In addition, it may be beneficial to jointly model multiple aspects of behavior (*e.g.,* multiple features from prosody).

The other interest is on the order of quantization Q. We tested the choices of $Q = 2, 4, 5$ in addition to $Q = 3$. In general we observe a similar trend compared to the findings in Sec. 3.5, however, the significances and accuracies are in general lower than the case of $Q = 3$. We believe that having more quantization bins may cause sparsity, even though fewer bins may reduce the discriminative power of the feature set.

3.7 Conclusion

In this Chapter we have extracted, quantized and modeled the distribution of prosodic cues in order to infer therapist empathy in Motivational Interviewing based psychotherapy. We found salient prosodic patterns that are significantly correlated with empathy, which was used to classify "high" and "low" empathy ratings achieving an accuracy of 75%. The results suggest that the use of high energy and pitch by the therapist is a negative sign of empathy. The quantization of prosodic features enabled the capture of salient patterns that led to more accurate inference of high level behavior like empathy, and outperformed the approach based on functionals of prosodic features.

In the future, we aim to validate empirical settings applied in this Chapter on larger-scale data, and in the end automate the parameter adaptation for robust analysis in practical use. For the inference of empathy, it would be useful to jointly model the lexical and prosodic information, in order to have a complete account of both "what they say" and "how they say it".

Chapter 4

Modeling Empathy through Lexical Cues and the Automatic Rating System

4.1 Introduction

Addiction counseling is a type of psychotherapy, where the therapist aims to support changing the patient's addictive behavior through face-to-face conversational interaction. Mental health care toward drug and alcohol abuse is essential to society. In the United States, a national survey by SAMHSA [65] showed that there were 23.9 million illicit drug users in 2012. However, only 2.5 million persons received treatment at a specialty facility [65]. Further to the gap between the provided addiction counseling and what is needed, it is also challenging to evaluate millions of counseling cases regarding the quality of the therapy and the competence of the therapists.

Unlike pharmaceuticals whose quality can be assessed during design and manufacturing, psychotherapy is essentially an interaction where multimodal communicative behaviors are the means of treatment, hence the quality is at best unknown until after the interaction takes place. Traditional approaches of evaluating the quality of therapy and therapist performance rely on manual observational coding of the therapist-patient interaction, *e.g.,* reviewing tape recordings and annotating

them with performance scores. This kind of coding process often takes more than five times real time, including initial human coder training and reinforcement [66]. The lack of human and time resources prohibits the evaluation of psychotherapy in large scale; and moreover, it limits deeper understanding of how therapy works due to the small number of cases evaluated. Similar issues exist in many human centered application fields such as education and customer service.

In this Chapter, we propose computational methods for evaluating therapists performance based on their behaviors. We focus on one type of addiction counseling called *Motivational Interviewing* (MI), which helps people to resolve ambivalence and emphasizes the intrinsic motivation of changing addictive behaviors [24]. MI has been proved effective in various clinical trails; and theories about its mechanisms have been developed [25]. Notably, *Therapist empathy* is considered essential to the quality of care, in a range of health care interactions including MI, where it holds a prominent function.

The study of the techniques that support the measurement, analysis, and modeling of human behavior signals is referred to as Behavioral Signal Processing (BSP) [17]. The primary goal of BSP is to inform human assessment and decision making. Other examples of BSP applications include the use of acoustic, lexical, and head motion models to infer expert assessments of married couples' communicative behavioral characteristics in dyadic conversations [67–69], and the use of vocal prosody and facial expressions in understanding behavioral characteristics in Autism Spectrum Disorders [55,70–72]. Closely related to BSP, Social Signal Processing studies modeling, analysis and synthesis of human social behavior through multimodal signal processing [73].

However, empathy estimation in previous work (see Chapter 2) requires manual annotations of behavioral cues not only for training the empathy model, but also

for application on new observations. Manual annotation on new observation data prohibits large scale deployment of therapist assessment, as it costs a large amount of time and manual labor. A fully automatic empathy estimation system would be very useful in real applications, even though manual annotations are still required for training the system. The system should, for example, take the audio recording of the interaction as input, and return the therapist empathy rating as its output, and no manual intervention would be needed in the process. In this Chapter, we propose a prototype system that satisfies this requirement.

We build the system by integrating state-of-the-art speech and language processing techniques. The top level diagram of the system is shown in Figure 4.1. We employ a Voice Activity Detection (VAD) module to separate speech from non-speech (when they speak); we employ a diarization module to separate speakers in the interaction (who is speaking). We setup an Automatic Speech Recognition (ASR) system to decode spoken words from the audio (what they say); and employ role-specific language models (*i.e.,* therapist *vs.* patient) to match the speakers with their roles (who is whom). The above four parts comprise an automatic transcription system, which takes audio recording of a session as input, and provides time-segmented spoken language as output. For therapist empathy modeling in this chapter, we focus on the spoken language of the therapist only. We propose three methods for empathy level estimation based on language models representing high *vs.* low empathy, including using the Maximum Entropy model, the Maximum likelihood based model trained with human-generated transcripts, and a Maximum likelihood approach based on direct ASR lattice rescoring.

Given the access to a collection of relatively large size, well annotated databases of MI transcripts, we train various models for each processing step, and evaluate

the performance of intermediate steps as well as the final empathy estimation accuracies by different models.

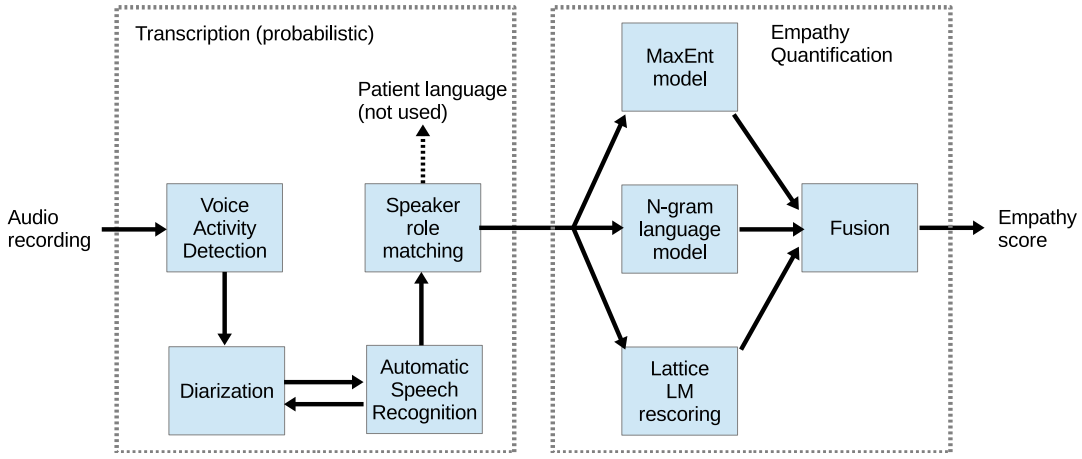

Figure 4.1: Overview of modules in the automatic empathy code prediction system

In the rest of this Chapter, we first describe the modules and methods in the automatic transcription system in Sec. 4.2. We then describe the lexical modeling of empathy in Sec. 4.3. We introduce the real application data utilized in this Chapter in Sec. 4.4. We describe the system implementation in Sec. 4.5, and report experiment results in Sec. 4.6. In Sec. 4.7 we discuss the findings in this Chapter We conclude the chapter in Sec. 4.8.

4.2 Automatic Speech Recognition

4.2.1 Voice Activity Detection

Voice Activity Detection (VAD) separates speech from non-speech, *e.g.,* silence and background noises. It is the first module in the system, which takes the audio recording of a psychotherapy session as input.

We employ the VAD system developed by Van Segbroeck *et al.* [61]. The system extracts four types of speech features: (i) spectral shape, (ii) spectro-temporal modulations, (iii) periodicity structure due to the presence of pitch harmonics, and (iv) the long-term spectral variability profile. In the next stage, these features are normalized in variance; and a three-layer neural network is trained on the concatenation of these feature streams.

The neural network outputs the voicing probability for each audio frame, which requires binarization to determine the segmentation points. We use an adaptive threshold on the voicing probability to constrain the maximum length of speech segments. We increase the binarization threshold beginning from 0.5, until that all segments are shorter than an upper bound of segment length (*e.g.,* 60s). Spoken segment longer than that is infrequent in the target dyadic interactions, and not memory efficient to process in speech recognition. We merge neighboring segments on condition that the gap between them is shorter than a lower bound (*e.g.,* 0.1s) and the combined segment does not exceed the upper bound of segment length (*e.g.,* 60s). After the merging we drop segments that are too short (*e.g.,* less than 1s).

4.2.2 Speaker Diarization

Speaker diarization is a technique that provides segmentation of the audio with information about "who spoke when". Separating the speakers facilitates speaker adaptation in ASR, and identification of speaker roles (patient, therapist in our application). We assume the number of speakers is known a priori in the application — two speakers in addiction counseling. Therefore, the diarization process mainly includes a segmentation step (dividing speech to speaker homogeneous segments) and a clustering step (assigning each segment to one of the speakers).

We employ two diarization methods as follows, and both of them take VAD results and Mel-Frequency Cepstrum Coefficient (MFCC) features as inputs. The first method uses Generalized Likelihood Ratio (GLR) based speaker segmentation, and agglomerative speaker clustering as implemented in [74]. The second method adopts GLR speaker segmentation and Riemannian manifold method for speaker clustering, as implemented in [75]. This method slices each GLR derived segment into short-time segments (*e.g.,* 1s), so as to increase the number of samples in the manifold space for more robust clustering (see [75] for more detail).

After obtaining the diarization results we compute session-level heuristics for outlier detection: *e.g.,* (i) percentage of speaking time by each speaker, (ii) longest duration of a single speaker's turn. These statistics can be checked against their expected values; and we define an outlier as a value that is more than three times of standard deviation away from the mean. For example, a 95%/5% split of speaking time in the two clusters may be a result of clustering speech *vs.* silence due to imperfect VAD. We use the heuristics and a rule based scheme to integrate the results from different diarization methods as described further in Sec. 4.5.

4.2.3 ASR

We employ a large vocabulary, continuous speech recognizer (LVCSR), implemented using the Kaldi library [76].

Feature: The input audio format is 16kHz single channel far-field recording. The acoustic features are standard MFCCs including Δ and $\Delta\Delta$ features.

Dictionary: We combine the lexicon in Switchboard [77] and WSJ [78] corpora, and manually add high frequency domain-specific words collected from the training corpus, *e.g., mm* as a filler word and *vicodin* as an in-domain word. We

ignore low frequency OOV words in the training corpus including misspellings and made-up words, which in total take less than 0.03% of all word tokens.

Text training data: We tokenize the training transcripts as follows. Overlapped speech regions of the two speakers are marked and transcribed; we only keep the longer utterance. Repetitions and fillers are marked and retained in the way they are uttered. We normalize non-verbal vocalization marks into either "[laughter]" or "[noise]". We also replace underscores by spaces, and remove punctuations and special characters.

Acoustic Model training: For the Acoustic Model (AM), we first train a GMM-HMM based AM, initially on short utterances with a monophone setting, and gradually expand it to a tri-phone structure using more training data. We then apply feature Maximum Likelihood Linear Regression (fMLLR) and Speaker Adaptive Training (SAT) techniques to refine the model. Moreover, we train a Deep Neural Network (DNN) AM with tanh nonlinearity, based on the alignment information obtained from the previous model.

Language Model training: For Language Model (LM) training, we employ SRILM to train N-gram models [79]. Initial LM is obtained from the text of the training corpus, using trigram model and Kneser-Ney smoothing. We further employ an additional in-domain text corpus of psychotherapy transcripts (see Sec. 4.4) to improve the LM. The trigram model of the additional corpus is trained in the same way and mixed with the main LM, where the mixing weight is optimized on heldout data.

4.2.4 Speaker Role Matching

The therapist and patient play distinct roles in psychotherapy interaction; knowing the speaker role hence is useful for modeling therapist empathy. The diarization

module only identifies distinct speakers but not their roles in the conversation. One way to automatically match roles to the speakers is by evaluating the styles of language use. For example, a therapist may use more questions than the patient. We expect a lower perplexity when the language content of the audio segment matches the LM of the speaker role, and *vice versa*. In the following we describe the role-matching procedure in detail.

0. **Input**: training transcripts with speaker-role annotated, two sets of ASR decoded utterances $\mathbf{U}_1$ and $\mathbf{U}_2$ for diarized speakers S_1 and S_2.

1. Train role-specific language models for (**T**)herapist and (**P**)atient separately, using corresponding training transcripts, *e.g.,* trigram LMs with Kneser-Ney smoothing, using SRILM [79].

2. Mix the final LM used in ASR to the role-specific LMs by a small weight (*e.g.,* 0.1), for vocabulary consistency and robustness.

3. Compute $ppl_{1,T}$ and $ppl_{1,P}$ as the perplexities for $\mathbf{U}_1$ over the two role-specific LMs. Similarly get $ppl_{2,T}$ and $ppl_{2,P}$ for $\mathbf{U}_2$.

4. Three cases: (i) (4.1) holds — we match S_1 to therapist and S_2 to patient; (ii) (4.2) holds — we match S_1 to patient and S_2 to therapist; (iii) in all other conditions, we take both S_1 and S_2 as therapist.

$$ppl_{1,T} \leq ppl_{1,P} \quad \& \quad ppl_{2,P} \leq ppl_{2,T} \tag{4.1}$$

$$ppl_{1,P} < ppl_{1,T} \quad \& \quad ppl_{2,T} < ppl_{2,P} \tag{4.2}$$

5. Outliers: When the diarization module outputs highly biased result in speaking time for two speakers, the comparison of perplexities is not meaningful.

If the total word count in $\mathbf{U}_1$ is more than 10 times of that in $\mathbf{U}_2$, we match S_1 to therapist; and *vice versa*.

6. **Output**: $\mathbf{U}_1$ and $\mathbf{U}_2$ matched to speaker roles.

When there is not a clear role match, *e.g.,* in step 4, case III and step 5, we have to make assumptions about speaker roles. Since our target is the therapist, we tend to oversample therapist language to augment captured information, and trade-off with the noise brought from patient language.

4.3 Therapist Empathy Models using Language Cues

We employ manually transcribed therapist language in MI sessions with high *vs.* low empathy ratings to train separate language models representing high *vs.* low empathy. Given the ASR extracted therapist language, we first infer therapist empathy at the utterance level, then integrate the local evidence towards session level empathy estimation. We discuss more about the modeling strategies in Sec. 4.7.1. The details of the proposed methods are described as follows.

4.3.1 Maximum Entropy Model

Maximum Entropy (MaxEnt) model is a type of exponential model that is widely used in natural language processing tasks, and achieves good performance in these tasks [80, 81]. We train a two-class (high *vs.* low empathy) MaxEnt model on utterance level data using the MaxEnt toolkit in [82].

Let high and low empathy classes be denoted H and L respectively, and $Y \in \{H, L\}$ be the class label. Let $u \in \mathbf{U}$ be an utterance in the set of therapist

utterances. We use n-grams ($n = 1, 2, 3$) as features for the feature function $f_n^j(u, Y)$, where j is an index of the n-gram. We define $f_n^j(u, Y)$ as the count of the j-th n-gram type that appears in u if $Y_u = Y$, otherwise 0.

MaxEnt model then formulates the posterior probability $P_n(Y|u)$ as an exponent of the weighted sum of feature functions $f_n^j(u, Y)$, as shown in (4.3), where we denote the weight and partition function as λ_n^j and $Z(u)$, respectively. In the training phase, λ_n^j is determined through the L-BFGS algorithm [83].

$$P_n(Y|u) = \frac{1}{Z(u)} exp\left(\sum_j \lambda_n^j f_n^j(u, Y)\right) \tag{4.3}$$

Based on the trained MaxEnt model, we compute the session level empathy score α_n as the average of utterance level evidence $P_n(H|u)$, as shown in (4.4), where $\mathbf{U}_T$ is the set of K therapist utterances.

$$\alpha_n(\mathbf{U}_T) = \frac{1}{K}\sum_{i=1}^{K} P_n(H|u_i),\ \mathbf{U}_T = \{u_1, u_2, \cdots, u_K\},\ n = 1, 2, 3. \tag{4.4}$$

4.3.2 Maximum Likelihood Model

Maximum Likelihood based N-gram language models (LM) can provide the likelihood of an utterance conditioned on a specific style of language, *e.g.,* $P(u|H)$ as the likelihood of utterance u in the empathic style. Following the Bayesian relationship, we model the posterior probability $P(H|u)$ by the likelihoods as in (4.5), where we assume equal prior probabilities $P(H) = P(L)$.

$$P(H|u) = \frac{P(u|H)P(H)}{P(u|H)P(H) + P(u|L)P(L)} = \frac{P(u|H)}{P(u|H) + P(u|L)} \tag{4.5}$$

We train the high empathy LM ($\texttt{LM}_H$) and low empathy LM ($\texttt{LM}_L$) using manually transcribed therapist language in high empathic and low empathic sessions, respectively. We employ trigram LMs with Kneser-Ney smoothing by SRILM in implementation [79]. Next, for robustness we mix a large in-domain LM (*e.g.,* the final LM in ASR) to $\texttt{LM}_H$ and $\texttt{LM}_L$ with a small weight (*e.g.,* 0.1). Let us denote the mixed LMs as $\texttt{LM}_H^{'}$ and $\texttt{LM}_L^{'}$.

For the inference of $P(H|u)$, we first compute the log-likelihoods $l_n(u|H)$ and $l_n(u|L)$ by applying $\texttt{LM}_H^{'}$ and $\texttt{LM}_L^{'}$, where $n = 1, 2, 3$ are the utilized N-gram orders. Then $P_n(H|u)$ is obtained as in (4.6).

$$P_n(H|u) = \frac{e^{l_n(u|H)}}{e^{l_n(u|H)} + e^{l_n(u|L)}} \tag{4.6}$$

We compute session level empathy score β_n as the average of utterance level evidences as shown in (4.7), where $\mathbf{U}_T$ is the same as in (4.4).

$$\beta_n(\mathbf{U}_T) = \frac{1}{K}\sum_{i=1}^{K} P_n(H|u_i) \tag{4.7}$$

4.3.3 Maximum Likelihood Rescoring on ASR Decoded Lattices

Instead of evaluating a single utterance as the best path in ASR decoding, we can evaluate multiple paths at once by rescoring the ASR lattice. The score (in likelihood sense) rises for the path of an highly empathic utterance when evaluated on the empathy LM, while drops on the low empathy LM. We hypothesize that rescoring the lattice would re-rank the paths so that empathy-related words may be picked up, which improves the robustness of empathy modeling when the decoding is noisy (more discussion in Sec. 4.7.4). In the following we describe the method

in more detail. An illustration of the lattice paths re-ranking effect is shown in Figure 4.2.

0. **Input**: ASR decoded lattice $\mathcal{L}$, high and low empathy LMs $\texttt{LM}_H^{'}$, $\texttt{LM}_L^{'}$ as described in Sec. 4.3.2

1. Update the LM scores in $\mathcal{L}$ by applying $\texttt{LM}_H^{'}$ and $\texttt{LM}_L^{'}$ as trigram LMs, denote the results as $\mathcal{L}_H$ and $\mathcal{L}_L$, respectively.

2. Rank the paths in $\mathcal{L}_H$ and $\mathcal{L}_L$ according to the weighted sum of AM and LM scores.

3. List the final scores of the R-best paths in $\mathcal{L}_H$ and $\mathcal{L}_L$ as $s_H(r)$ and $s_L(r)$ in the log field, $1 \leq r \leq R$, respectively.

4. Compute the utterance level empathy score $S_H(\mathcal{L})$ as in (4.8)

$$S_H(\mathcal{L}) = \frac{\texttt{exp}\left(\frac{1}{R}\sum_{r=1}^{R} s_H(r)\right)}{\texttt{exp}\left(\frac{1}{R}\sum_{r=1}^{R} s_H(r)\right) + \texttt{exp}\left(\frac{1}{R}\sum_{r=1}^{R} s_L(r)\right)} \tag{4.8}$$

5. Compute the session level empathy score γ as in (4.9), where $\mathcal{U}_T$ is the set of K lattices of therapist utterances.

$$\gamma(\mathcal{U}_T) = \frac{1}{K}\sum_{i=1}^{K} S_H(\mathcal{L}_i) \tag{4.9}$$

6. **Output**: Session level empathy score γ

Note that the lattice rescoring method is a natural extension of the Maximum Likelihood LM method in Sec. 4.3.2. When the score $s_H(r)$ denotes log-likelihood and $R = 1$, (4.8) becomes equivalent to (4.6). In that case $S_H(\mathcal{L})$ represents a

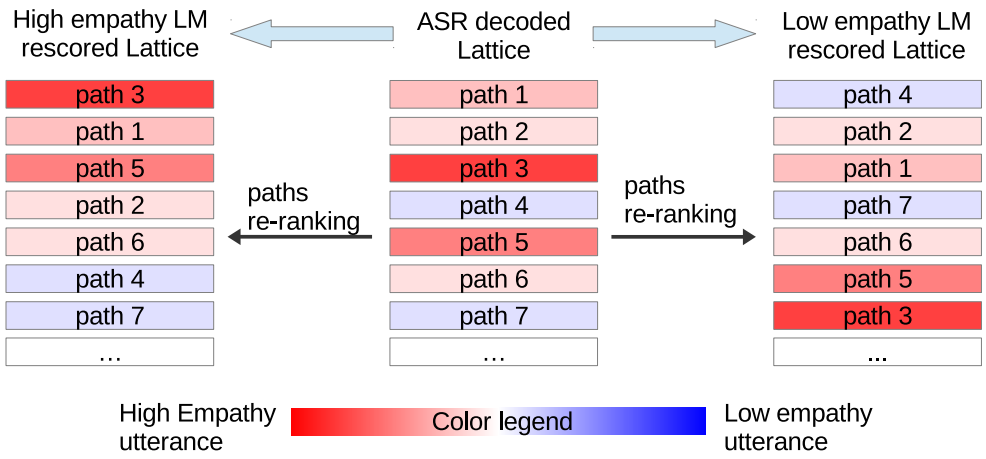

Figure 4.2: Illustration of rescoring lattice by high and low empathy LMs.

similar meaning to $P(H|\mathcal{L})$. The lattice is a more compact way of representing the hypothesized utterances since there is no need to write out the pathes explicitly. It also allows more efficient averaging of the evidence from the top hypotheses.

4.4 Data Corpora

In this section we introduce the three data corpora used in the study.

- "TOPICS" corpus — 153 audio-recorded MI sessions randomly selected from 899 sessions in five psychotherapy studies [84–88], including intervention of college student drinking and marijuana use, as well as clinical mental health care for drug use. Audio data are available as single channel far-field recordings in 16 bit quantization, 16 kHz sample rate. Audio quality of the recordings varies significantly as they were collected in various real clinical settings. The selected sessions were manually transcribed with annotations of speaker, start-end time of each turn, overlapped speech, repetition, filler words, incomplete words, laughter, sign, and other nonverbal vocalizations. Session length ranges from 20 min to 1 hour.

- "General Psychotherapy" corpus — transcripts of 1200 psychotherapy sessions in MI and a variety of other treatment types [89]. Audio data are not available.

- "CTT" corpus — 200 audio-recorded MI sessions selected from 826 sessions in a therapist training study (namely Context Tailored Training) [57]. The recording format and transcription scheme are the same as TOPICS corpus. Each session is about 20 min.

All research procedures for this study were reviewed and approved by Institutional Review Boards at the University of Washington (IRB_36949) and University of Utah (IRB_00058732). During the original trials all participants provided written consent. The UW IRB approved all consent procedures.

The details about the corpus sizes are listed in Table 4.1.

Table 4.1: Detail about size information of the data corpora

Corpus	No. sessions	No. talk turns	No. word tokens	Duration
TOPICS	153	3.69×10^4	1.12×10^6	104.2 hr
Gen. Psyc.	1200	3.01×10^5	6.55×10^6	-
CTT	200	2.40×10^4	6.24×10^5	68.6 hr

4.4.1 Empathy Annotation in CTT Corpus

Three coders reviewed the 826 audio recordings of the entire CTT corpus, and annotated therapist empathy using a specially designed coding system — the "Motivational Interviewing Treatment Integrity" (MITI) manual [26]. The empathy code values are discrete from 1 to 7, with 7 being of high empathy and 1 being of low empathy. 182 sessions were coded twice by the same or different coders, while no session was coded three times. The first and second empathy codes of the

sessions that were coded twice had a correlation of 0.87. Intra-Class Correlation (ICC) is 0.67±0.16 for inter-coder reliability, and 0.79±0.13 for intra-coder reliability. These statistics prove coder reliability in the annotation. We use the mean value of empathy codes if the session is coded twice.

In the original study, three psychology researchers acted as *Standardized Patient* (SP), whose behaviors were regulated for therapist training and evaluation purposes. For example, SP sessions had pre-scripted situations. Sessions involving a SP or a Real Patient (RP) were about the same size in the entire corpus. The 200 sessions used in this study are selected from the two extremes of empathy codes, which may represent empathy more prominently. The class of low empathy sessions has a range of code values from 1 to 4, with mean value of 2.16±0.55; while that for the high empathy class is 4.5 to 7, with mean of 5.90±0.58. We show the counts of high *vs.* low empathy and SP *vs.* RP sessions in Table 4.2. Moreover, the selected sessions are diverse in the therapists involved. There are 133 unique therapists, and any therapist has no more than three sessions.

Table 4.2: Counts of SP, RP, high and low empathy sessions in the CTT corpus

Patient	Low emp.	High emp.	Total	Ratio of high emp.
SP	46	78	124	62.9%
RP	33	43	76	56.6%
All	79	121	200	60.5%

4.5 System Implementation

In this section, we describe the system implementation in more detail. We summarize the usage of data corpora in various modeling and application steps in Table 4.3.

Table 4.3: Summary of data corpora usage

Corpus	Phase	VAD	Diar.	ASR-AM	ASR-LM	Role	Emp.
TOPICS	Train	X		X	X	X	
	Test						
Gen. Psyc.	Train				X	X	
	Test						
CTT	Train						X
	Test	X	X	X	X	X	X

VAD: We construct the VAD training and development sets by sampling from the TOPICS corpus. The total length of the two sets are 5.2h and 2.6h, respectively. We expect a wider coverage of heterogeneous audio conditions would increase the robustness of the VAD. We train the neural network as described in Sec. 4.2.1, and tune the parameters on the development set. We apply VAD on the CTT corpus.

Diarization: We run diarization on the CTT corpus as below.

1. Result D_1: apply the agglomerative clustering methods in [74].

2. Result D_2: apply the Riemannian clustering method in [75].

3. Run ASR using D_2 derived segmentation, obtain new VAD information according to the alignment in the decoding, disregard the decoded words.

4. Result D_3: based on the new VAD information, apply the method in [75] again, with a scheme of slicing speech regions into 1-minute short segments.

5. Result D_4: if D_3 is an outlier that is detected using the heuristics in Sec. 4.2.2, and D_2 or D_1 is not an outlier, then take D_2 or D_1 in turn as D_4; otherwise take D_3 as D_4. Such an integration scheme is informed by the performance on the training corpus.

ASR: We train the AM and the initial LM using the TOPICS corpus. We employ the General Psychotherapy corpus as a large in-domain data set and mix it in the LM for robustness. We observe that perplexity decreases on the heldout data after the mixing. The Deep Neural Network model is trained following the "`train_tanh.sh`" script in the Kaldi library. The ASR is used in finding more accurate VAD results as mentioned above. In addition, we apply the ASR to the CTT corpus under two conditions: (i) assuming accurate VAD and diarization conditions by utilizing the manually labeled timing and speaker information; (ii) using the automatically derived diarization results to segment the audio.

Role matching: We use the TOPICS corpus to train role-specific LMs for the therapist and patient. We also mix the final LM in ASR with the role-specific LMs for robustness.

Empathy modeling: We conduct empathy analysis on the CTT corpus. Due to data sparsity, we carry out a leave-one-therapist-out cross-validation on CTT corpus, *i.e.,* we use data involving all-but-one therapist's sessions in the corpus to train high *vs.* low empathy models, and test on that held-out therapist. For the lattice LM rescoring method in Sec. 4.3.3, we employ the top 100 paths ($R = 100$).

Empathy model fusion: The three methods in Sec. 4.3 and different choices of n-gram order n may provide complementary cues about empathy. This motivates us to setup a fusion module. Since we need to carry out cross-validation for empathy analysis, in order to learn the mapping between empathy scores and codes, we conduct an internal cross-validation on the training set in each round. For a single empathy score, we use linear regression and threshold search (minimizing classification error) for the mapping to the empathy code and the high or low class, respectively. For multiple empathy scores, we use support vector regression and linear support vector machine for the two mapping tasks, respectively.

4.6 Experiment and Results

4.6.1 Experiment Setting

We examine the effectiveness of the system by setting up the experiments in three conditions for comparison.

- ORA-T — Empathy modeling on manual transcriptions of therapist language (*i.e.,* using ORAcle Text).
- ORA-D — ASR decoding of therapist language with manual labels of speech segmentation and speaker roles (*i.e.,* using ORAcle Diarization and role labels), followed by empathy modeling on the decoded therapist language.
- AUTO — Fully automatic system that takes audio recording as input, carries out all the processing steps in Sec. 4.2 and empathy modeling in Sec. 4.3.

We setup three evaluation metrics regarding the performance of empathy code estimation: Pearson's correlation ρ, Root Mean Squared Error (RMSE) σ between expert annotated empathy codes and system estimations, and accuracy *Acc* of session-wise high *vs.* low empathy classification.

4.6.2 ASR System Performance

We report averaged false alarm, miss, speaker error rate (for diarization only), and total error rate for the VAD and diarization modules in Table 4.4. We can see that ASR derived VAD information dramatically improves the diarization results in D_4 compared to D_2 that is based on the initial VAD.

We report averaged ASR performance in terms of substitution, deletion, insertion, and total Word Error Rate (WER) in Table 4.5 for the case of ORA-D and

Table 4.4: VAD and diarization performance.

Results	False Alarm (%)	Miss (%)	Speaker error (%)	Total error (%)
VAD	5.8	6.8	-	12.6
D_2	6.9	8.7	13.7	29.3
D_4	4.2	6.7	7.3	18.1

AUTO. We can see that in the AUTO case there is a slight increase in WER, which might be a result of VAD and diarization errors, as well as the influence on speaker adaptation effectiveness. Using clean transcripts we were able to identify speaker roles for all sessions. For the AUTO case, due to diarization and ASR errors, we found a match of speaker roles in 154 sessions (78%), but failed in 46 sessions.

Table 4.5: ASR performance for ORA-D and AUTO cases.

Cases	Substitution (%)	Deletion (%)	Insertion (%)	WER (%)
ORA-D	27.1	11.5	4.6	43.1
AUTO	27.9	12.2	4.5	44.6

There are two notes about the speech processing results. First, due to the large variability of audio conditions in different sessions, the averaged results are affected by the very challenging cases. For example, session level ASR WER is in the range of 19.3% to 91.6%, with median WER of 39.9% and standard deviation of 16.0%. Second, the evaluation of VAD and diarization are based on speaking-turn level annotations, which ignore gaps, backchannels, and overlapped regions within turns. Therefore inherent errors exist in the reference data, but we believe they should not affect the conclusions significantly due to the relatively low ratio of such events.

4.6.3 Empathy Code Estimation Performance

In Table 4.6 we show the results of empathy code estimation using the fusion of empathy scores α_n, $n = 1, 2, 3$, which are derived by the MaxEnt model and n-gram features in Sec. 4.3.1. We compare the performance in ORA-T, ORA-D, and AUTO cases, for SP, RP and all sessions separately. Note that due to data sparsity, we conduct leave-one-therapist-out cross-validation on all sessions, and report the performance separately for SP and RP data. The correlation ρ is in the range of 0 to 1; the RMSE σ is in the space of empathy codes (1 to 7); and the classification accuracy *Acc* is in percentage.

Table 4.6: Empathy code estimation performance using MaxEnt model

	SP			RP			All sessions		
Cases	ρ	σ	*Acc*	ρ	σ	*Acc*	ρ	σ	*Acc*
ORA-T	0.747	1.27	87.9	0.653	1.49	80.3	0.707	1.36	85.0
ORA-D	0.699	1.38	85.5	0.651	1.51	84.2	0.678	1.43	85.0
AUTO	0.693	1.48	87.1	0.452	1.73	64.5	0.611	1.58	78.5

Similarly, in Table 4.7 we show the results by the fusion of empathy scores β_n, $n = 1, 2, 3$, derived by the n-gram LMs in Sec. 4.3.2. From the results in Table 4.6 and Table 4.7 we can see that the MaxEnt method and the Maximum Likelihood LM method are comparable in performance. The MaxEnt method suffers more from noisy data in the RP sessions than the Maximum Likelihood LM method as the performance decreases more in the AUTO case for RP, while it is more effective in cleaner condition like the ORA-D case. As a type of discriminative model, the MaxEnt model may overfit more than the Maximum Likelihood LM method in the condition of sparse training data. Thus the influence of noisy input is also heavier for the MaxEnt model.

Table 4.7: Empathy code estimation performance using Maximum Likelihood method

	SP			RP			All sessions		
Cases	ρ	σ	*Acc*	ρ	σ	*Acc*	ρ	σ	*Acc*
ORA-T	0.749	1.27	89.5	0.632	1.51	77.6	0.706	1.37	85.0
ORA-D	0.699	1.39	86.3	0.581	1.62	71.1	0.654	1.48	80.5
AUTO	0.693	1.51	87.1	0.510	1.72	73.7	0.628	1.59	82.0

In Table 4.8, we show the results using the empathy score γ that is derived by the lattice LM rescoring method in Sec. 4.3.3, for the case of ORA-D and AUTO that involves ASR decoding. Here we set the count of paths R for score averaging as 100. The lattice rescoring method performs comparably well in the ORA-D case. It performs well in the AUTO case for RP sessions, but suffers in SP sessions. For the latter, there might be a side effect that is influencing the performance — lattice path re-ranking may pick up words in patient language that are relevant to empathy, such that the noise (*i.e.,* patient language mixed in) is also "colored" and no longer neutral to empathy modeling. Since the SP sessions have similar story setup (hence shared vocabulary) but not for the RP sessions, such effect may be less for RP sessions.

Table 4.8: Empathy code estimation performance using lattice LM rescoring method

	SP			RP			All sessions		
Cases	ρ	σ	*Acc*	ρ	σ	*Acc*	ρ	σ	*Acc*
ORA-T	-	-	-	-	-	-	-	-	-
ORA-D	0.673	1.41	85.5	0.654	1.47	79.0	0.661	1.43	83.0
AUTO	0.557	1.58	79.0	0.516	1.64	76.3	0.543	1.60	78.0

In Table 4.9, we show the results by the fusion of the empathy scores including α_n, β_n, and γ, $n = 1, 2, 3$. The best overall results are achieved by such fusion except *Acc* in the AUTO case. With the fully automatic system, we achieve higher

than 80% accuracy in classifying high *vs.* low empathy, and correlation of 0.643 in estimation of empathy code. The performance for SP sessions is much higher than that for RP sessions. One reason might be that SP sessions are based on scripted situations (*e.g.,* Child Protective Serves takes kid away from mother who then comes to psychotherapy), while RP sessions are not scripted, and the topics tend to be diverse.

Table 4.9: Empathy code estimation performance by the fusion of the MaxEnt, Maximum Likelihood, and lattice LM rescoring (for ORA-D and AUTO cases) methods

	SP			RP			All sessions		
Cases	ρ	σ	*Acc*	ρ	σ	*Acc*	ρ	σ	*Acc*
ORA-T	0.758	1.24	90.3	0.667	1.45	79.0	0.721	1.32	86.0
ORA-D	0.717	1.33	87.9	0.674	1.46	86.8	0.695	1.38	87.5
AUTO	0.702	1.43	87.1	0.534	1.67	71.1	0.643	1.53	81.0

4.7 Discussion

4.7.1 Empathy Modeling Strategies

In this section we will discuss more about empathy and modeling strategies. Empathy is not an individual property but exhibited during interactions. More specifically, empathy is expressed and perceived in a cycle [15]: (i) patient expression of experience, (ii) therapist empathy resonation, (iii) therapist expression of empathy, and (iv) patient perception of empathy. The real empathy construct is in (ii), while we rely on (iii) to approximate the perception of empathy by human coders. This suggests one should model the therapist and patient jointly, as we have shown using the acoustic and prosodic cues for empathy modeling in [18, 46].

However, joint modeling in the lexical domain may be very difficult, since patient language is unconstrained and highly variable, which leads to data sparsity. Therapist language, as in (iii) above encodes empathy expression and hence provides the main source of information. Can *et al.* [90] proposed an approach to automatically identify a particular type of therapist talk style named *reflection*, which is closely linked to empathy. It showed that N-gram features of therapist language contributed much more than those of patient language. Therefore in this initial work we focused on the modeling of therapist language, while in the future plan to investigate effective ways of incorporating patient language.

Human annotation of empathy in this Chapter is a session level assessment, where coders evaluate the therapist's overall empathy level as a *gestalt*. In a long session of psychotherapy, the perceived therapist empathy may not be uniform across time, *i.e.,* there may be influential events or even contradicting evidence. Human coders are able to integrate such evidence towards an *overall* assessment. In this Chapter, since we do not have utterance level labels, in the training phase we treat all utterances in high *vs.* low empathy sessions as representing high *vs.* low empathy, respectively. We expect the model to overcome this since the N-grams manifesting high empathy may occur more often in high empathy sessions. In the testing phase, we found that scoring therapist language by utterances (and taking the average) exceeded directly scoring the complete set of therapist language. This demonstrates that the proposed methods are able to capture empathy on utterance level.

4.7.2 Inter-human-coder Agreement

62 out of 200 sessions in the CTT corpus were coded by two human coders. We binarize their coding with a threshold of 4.5. If the two coders annotated empathy

codes in the same class, we consider it as coder agreement. If they annotated the opposite, one (and only one) of them would have a disagreement to the class of the averaged code value. In Table 4.10 we list the counts of coder disagreement.

Table 4.10: Count of human coder disagreement

Coders	I	II	III	Total
Annotated sessions	43	47	34	$124 = 62 \times 2$
Disagreement	4	3	5	12
Agreement Ratio (%)	90.7	93.6	85.3	90.3

We see that the ratio of human agreement to the averaged code is around 90% on the CTT corpus. This suggests that human judgment of empathy is not always consistent, and the manual assessment of therapist may not be perfect. However, human agreement is still higher than that between the average code and automatic estimation (results in Table 4.9). In the future, we would like to investigate if computational methods can match human accuracy. Moreover, the computational assessment as an objective reference may be useful for studying the subjective process of human judgment of empathy.

4.7.3 Intuition about the Discriminative Power of Lexical Cues

Table 4.11: Bigrams associated with high and low empathy behaviors

High empathy			Low empathy		
sounds like	it sounds	kind of	okay so	do you	in the
that you	p s	you were	have to	your children	have you
i think	you think	you know	some of	in your	would you
so you	a lot	want to	at the	let me	give you
to do	sort of	you've been	you need	during the	would be
yeah and	talk about	if you	in a	part of	you ever
it was	i'm hearing	look at	have a	you to	take care

Table 4.12: Trigrams associated with high and low empathy behaviors

High empathy		Low empathy	
it sounds like	a lot of	during the past	please answer the
do you think	you think about	using card a	you need to
you think you	you think that	past twelve months	clean and sober
sounds like you	a little bit	do you have	have you ever
that sounds like	brought you here	some of the	to help you
sounds like it's	sounds like you're	little bit about	mm hmm so
p s is	you've got a	the past ninety	in your life
what i'm hearing	and i think	first of all	next questions using
one of the	if you were	you know what	you have to
so you feel	it would be	the past twelve	school or training

We analyze the discriminative power of N-grams to provide some intuition on what the model captures regarding empathy. We train $\texttt{LM}'_H$ and $\texttt{LM}'_L$ similarly to Sec. 4.3.2 on the CTT corpus. Let us denote n-gram terms as w, the log-likelihood derived from $\texttt{LM}'_H$ and $\texttt{LM}'_L$ as $l_n(w|H)$ and $l_n(w|L)$, respectively. Let $\texttt{cnt}(w)$ be the count of w in the CTT corpus. We define the discriminative power δ of w as in (4.10).

$$\delta(w) = (l_n(w|H) - l_n(w|L)) * \texttt{cnt}(w) \tag{4.10}$$

We show the bigrams and trigrams with extreme δ values, *i.e.,* strongly indicating high or low empathy, in Table 4.11 and Table 4.12, respectively. We see that high empathic words often express *reflective listening* to the patient, while low empathic words are often questioning or instructing the patient. This is consistent with the concept of empathy as "trying on the feeling" or "taking the perspective" of others.

4.7.4 Robustness of Empathy Modeling Methods

In this section we demonstrate the robustness of the lattice rescoring method in the ORA-D case (clean diarization), compared to MaxEnt and Maximum Likelihood LM methods. We examine how would each method perform when the WER increases. In order to simulate such conditions, we first generate the 1000-best lists of paths from the decoding lattice $\mathcal{L}$ and the high/low empathy LM rescored lattices $\mathcal{L}_H$, $\mathcal{L}_L$. We sample the lists at every 5 paths starting from the 1-best path, *i.e.,* in a sequence of 1, 6, 11, $\cdots$, 996, and treat them as the optimal paths from the decoding. If the sampling index exceeds the number of paths in the lattice, we take the last one in its N-best list. Based on every sampled path in $\mathcal{L}$, we carry out empathy code estimation by the MaxEnt and Maximum Likelihood LM methods. Based on the score of every sampled path in $\mathcal{L}_H$, $\mathcal{L}_L$, we carry out the lattice rescoring method. We set $R = 1$ for comparison, *i.e.,* taking the score of the first available path.

We show the results in Figure 4.3. In the upper left panel we plot the corresponding WER by the sampled paths from lattice $\mathcal{L}$. In the upper right and lower left/right panels, we plot the performance regarding ρ, σ, and *Acc* by the three methods, respectively. For figure clarity we display the mean and standard deviation for every 10 sample points (*e.g.,* the first point represents the statistics of sampling indices 1, 6, $\cdots$, 46). Meanwhile, we show the performances by using the 1-best decoded paths, denoted by asterisks.

In Figure 4.3, the WER increases by about 3%, while the performance in general drops accordingly. We observe that the lattice rescoring method outperforms the other two in degraded ASR conditions. Moreover, the lattice rescoring method tend to be more stable, while the other two methods suffer from large deviation in performance. This demonstrates the gain of robustness by re-ranking the paths

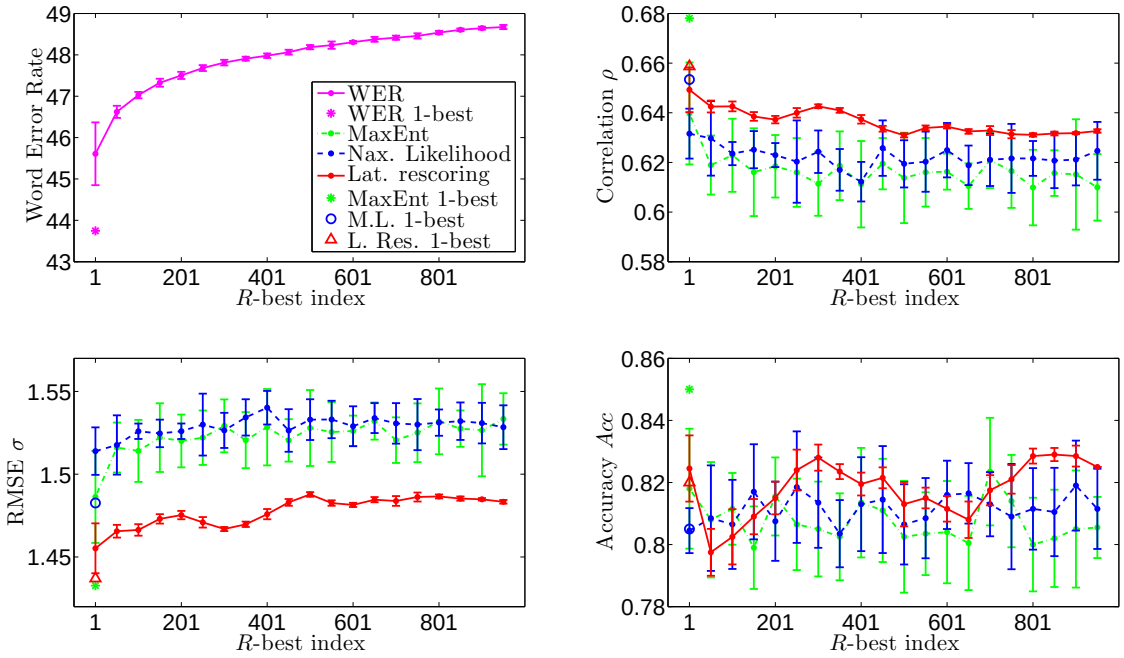

Figure 4.3: Comparison of robustness by MaxEnt, Maximum Likelihood, and lattice LM rescoring methods

according to their relevance to empathy, where the original lattice may have uncertain levels of empathy representation in the list of paths. In practice, if the empathy LM is rich enough, one can also decode the utterance directly using the high/low empathy LMs instead of rescoring the lattice.

4.7.5 Standard Patient and Real Patient Data

In Table 4.6 to 4.9 we have seen that the system is more effective for SP sessions than RP sessions. There may be several reasons. First, SP sessions are based on scripted situations (*e.g.,* Child Protective Serves takes kid away from mother who then comes to psychotherapy), while RP sessions are not scripted and the topics tend to be diverse. Second, the count of SP sessions is more than that of RP, hence more training data are available (see Table 4.2). As a result, data sparsity is less for SP sessions. Thirdly, SP sessions are recorded in a more controlled environment

such that the audio quality is better in average than RP sessions. This is reflected in the ASR WER: *e.g.,* in the ORA-D case, the mean session-wise WER for SP and RP are 34.5% and 57.6%, respectively.

In the current experiment, the classification accuracy *Acc* for RP sessions is still statistically significant with $p < 0.01$ in binomial test. We believe more sample data improvements in robust speech processing can improve performance in RP sessions.

4.8 Conclusion

In this chapter we have proposed a prototype of a fully automatic system to rate therapist empathy from language cues in addiction counseling. We constructed speech processing modules that include VAD, diarization, and a large vocabulary continuous speech recognizer customized to the topic domain. We employed role-specific language models to identify therapist's language. We applied MaxEnt, Maximum Likelihood LM, and lattice rescoring methods to estimate therapist empathy codes in MI sessions, based on lexical cues of the therapist's language. In the end, we composed these elements and implemented the complete system.

For evaluation, we estimated empathy using manual transcripts, ASR decoding using manual segmentation, and fully automated ASR decoding. Experiment results showed that the fully automatic system achieved a correlation of 0.643 between human annotation and machine estimation of empathy codes, as well as an accuracy of 81% in classifying high *vs.* low empathy scores. Using manual transcripts we achieve a better performance of 0.721 and 86% in correlation and classification accuracy, respectively. The experiment results show the effectiveness of the system in therapist empathy estimation. We also observed that the

performance of the three modeling methods are comparable in general, while the robustness varies for different methods and conditions.

In the future, we would like to improve the underlying techniques for speech processing and speech transcription. We would also like to acquire more and better training data such as by using close talking microphones in collections.

The system may be augmented by incorporating other behavioral modalities such as the acoustic and prosodic cues from the vocal channel, as well as gestures and facial expressions from the visual channel. A joint modeling of these dynamic behavioral cues may provide a more accurate quantification of therapist's empathy characteristics.

Chapter 5

Modeling Empathy through Speech Rate Entrainment

5.1 Introduction

In this Chapter, we follow the track of analyzing the connection between entrainment and empathy [46], by extending the dyadic patterning in speech rates. Entrainment refers to the phenomenon that the behaviors of the interlocutors becoming more similar during the interaction, possibly in multiple communication channels or biometrical states [91]. In the literature, theoretical relations between entrainment and empathy have been extensively studied [6, 8, 92, 93]. Some computational models of entrainment have also been reported, *e.g.,* Lee *et al.* have modeled the vocal entrainment of couples in conversations and its relation to the couples' affective behavioral characteristics [70]. Delaherche *et al.* have surveyed the emerging methods for capturing multimodal entrainment from behavior signals, and summarized them into three types: correlation based, phase and spectrum comparison, and bags-of-instances comparison [28].

Speech rate, *i.e.,* the number of words, syllables, or phonemes a subject utters in a unit of time, reflects many internal states of the subject. Entrainment in speech rate has been reported. Guitar *et al.* have shown that children slow their speech rate when the mothers speak slower [94]. Manson *et al.* have shown that the degree of speech rate entrainment may predict the outcome of a collaborative

task by two interlocutors [95]. However, little work has focused on computational models of the link between speech rate entrainment and empathy, which is the aim of this Chapter.

In this chapter, we first introduce the data sets in Sec. 5.2. We show a computational means for examining speech rate entrainment in Sec. 5.3. In Sec. 5.4 we investigate how the dynamics of speech rate entrainment are related to therapist empathy. In Sec. 5.5 we study the relation between speech/silence durations and empathy. We examine the performance of classifying perceived high *vs.* low empathy using the proposed rate cues in Sec. 5.6. We discuss the robustness of the cues in Sec. 5.7, and conclude the study with future directions in Sec. 5.8.

5.2 Dataset and Speech Alignment

To develop and test the ideas about speech rate entrainment, we consider two data sources: a standard telephonic human-human dialog, and a set of data drawn from a corpus of client-therapist interaction during drug addiction counseling.

5.2.1 Switchboard Corpus

Switchboard [77] is a large collection of two-sided telephone conversation from the United States. A robot operates the connection between the interlocutors and introduces a topic to discuss. It also ensures no two speakers would converse together more than once.

In our analysis we employ 2438 sessions from the corpus. We use the ASR generated, and manually corrected word level alignment of speech and transcript [77] to compute speech rates for each session and speaker.

5.2.2 Motivational Interviewing Data and Automatic Alignment

We employ the same TOPICS and CTT sets as in Section 4.4, which are recordings of Motivational Interviewing sessions. In total there are 353 sessions.

The available manual segmentation only marks speaking turns; for more precise timing between and within turns, we adopt an approach of force-aligning speech to transcripts based on ASR. In the experiment we employ the ASR trained in Section 4.2.3. We employ the Viterbi algorithm for phoneme level forced-alignment which we transform into word level alignment for further analysis. Further discussion about alignment reliability is in Sec. 5.7.

5.3 Matching of Average Speech Rate

We first investigate the proposed computational measure for entrainment in session-level, average speech rates of the interlocutors in the Switchboard corpus. We employ the Switchboard corpus since it is a standard database that contains a large number of interactions, therefore strengthens the statistical power of our hypothesis tests in addition to that obtained on the MI data. We define the average word rate R_w as in (5.1), where N is the total count of words (w_i) by a subject in the conversation. $t_{\texttt{begin}}$ and $t_{\texttt{end}}$ are the beginning and ending time of a word. We eliminate silence time to avoid the influence of line delay and interruption in phone conversation. Similarly, we obtain the average syllable rate R_s and phoneme rate R_p in (5.2), (5.3). Note that we exclude partial words and nonverbal units such as hesitations and laughters.

$$R_w = \frac{N}{\sum_{i=1}^{N} \left(t_{\texttt{end}}(w_i) - t_{\texttt{begin}}(w_i) \right)} \tag{5.1}$$

$$R_s = \frac{\sum_{i}^{N} syllable\_cnt(w_i)}{\sum_{i=1}^{N} \left(t_{\texttt{end}}(w_i) - t_{\texttt{begin}}(w_i) \right)} \tag{5.2}$$

$$R_p = \frac{\sum_{i}^{N} phoneme\_cnt(w_i)}{\sum_{i=1}^{N} \left(t_{\texttt{end}}(w_i) - t_{\texttt{begin}}(w_i) \right)} \tag{5.3}$$

We hypothesize that if entrainment exists in interlocutor speech rates, they should correlate higher for pairs of true interlocutors than any randomly shuffled pairing of speakers. Such a benchmarking approach is standard in dyadic analyses [28].

Firstly, in Figure 5.1 we show the distribution (the darker the higher density) of R_w by all speakers, where we see a clear trend of matching between pairs of interlocutors (labeled as speaker A and B in each pair). We compute the correlation of R_w (and R_s, R_p) over conversing speaker pairs to capture this trend of matching speech rates. In Table 5.1 we show the results. Due to the large number of samples (2438 sessions), these correlations are significant ($p < 10^{-19}$ in t-test) though the values are small. The correlations do not rely on the order of speaker labels A or B; the variance of the correlations obtained with random speaker labels is below 10^{-3}.

Meanwhile, we compute the correlation of the average speech rates between "randomly paired" *pseudo*-interlocutors that are not drawn from the same interaction. We repeat this process 1000 times. In Table 5.2 we report the mean value, most significant p-value, and maximum absolute value of the above correlations. We see that the lowest p-values under random pairings are dramatically larger than

those in the cases of true interactions. The mean values are close to zero, suggesting there is no correlation under random conditions. These results lend further support to the existence of entrainment in speech rates during interactions.

Table 5.1: Correlations of average speech rates by pairs of interlocutors, and the significance in t-test

Corpus	R_w	R_s	R_p	p-val
Switchboard	0.229	0.198	0.183	$< 10^{-19}$
TOPICS + CTT	0.279	0.314	0.311	$< 10^{-7}$

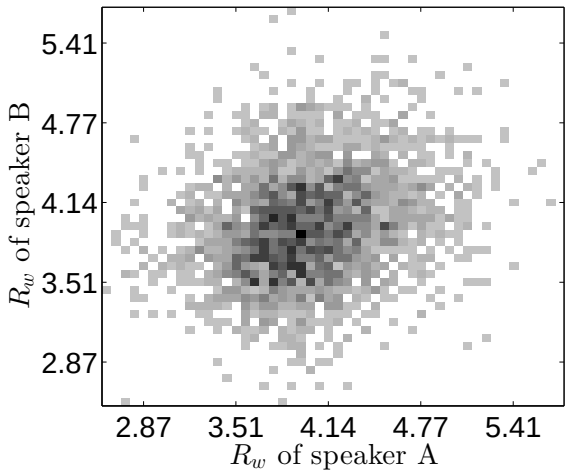

Figure 5.1: Distribution of average speech rates by pairs of interlocutors

Similarly, we conduct the analysis on the combination of the TOPICS and CTT sets using forced-alignment based speech rates. We exclude nonverbal and out-of-vocabulary words in computing the speech rates. As a result, we find significant correlations of average speech rates between the therapist and the patient, shown in Table 5.1. We also see that such correlations are not obtained in random pairings of therapists and patients, as shown in Table 5.2.

In conclusion, the results in this section demonstrate the entrainment in interlocutors' speech rates (*i.e.,* trend toward matching) in telephone conversation and addiction counseling scenarios.

Table 5.2: Statistics of correlations of average speech rates by randomly shuffled pairs of pseudo-interlocutors

Rate	Mean	Min. p-val.	Max. Abs.
Switchboard			
R_w	0.0005	0.0002	0.075
R_s	0.0004	0.0020	0.063
R_p	0.0006	0.0009	0.067
TOPICS + CTT			
R_w	0.0010	0.0011	0.173
R_s	0.0013	0.0001	0.212
R_p	-0.0023	0.0005	0.184

5.4 Relating Speech Rate Entrainment Dynamics and Empathy

In Sec. 5.3 we showed evidence that speech rates are part of the cues exemplifying behavioral entrainment. In this section we study if the degree of such entrainment contributes to the perceived therapist's empathy level in MI. We consider the turn-by-turn differences in speech rates as a computational measure for entrainment, where a turn is a period that a single speaker holds the speaking floor.

We segment the audio based on the forced alignment. We keep intra-speaker silence (defined as pause) that is longer than 0.2 seconds, while merge the others with the speech segments. In this way we retain inter-word short pauses, while keeping longer pauses separate from the calculation of speech rate. For inter-speaker silence (defined as gap), we retain all measured values without any flooring/ceiling. Overlapping speech segments exist in the corpus, but are not accessible from the alignment, so that they are left out from the current analysis. We use speech utterances longer than 0.5 seconds and discard the rest to improve the robustness of speech rate estimation. We obtain the turn level speech rate r by counting on the unit of utterances u_i $(1 \leq i \leq N_u)$, as in (5.4).

$$r = \frac{\sum_{i=1}^{N_u} symbol\_cnt(u_i)}{\sum_{i=1}^{N_u} \left(t_{\texttt{end}}(u_i) - t_{\texttt{begin}}(u_i)\right)} \tag{5.4}$$

We compute the averaged absolute differences of speech rates between each patient's turn and the therapist's turn that follows. This is because our focus is on the therapist's reaction to the patient's behavior. Let $r_w(k)$ and $r_w(k+1)$ be the word rate of turns k and $k+1$ that belong to the patient and therapist, respectively. r_w for the patient and the therapist are zero mean separately, *i.e.,* subtracted the mean of the raw turn-wise speech rate, so as to remove the bias of individual speech rate baseline. We define the averaged absolute difference D_w as in (5.5), assuming the session contains K turns, K being an even number. We also assume the session begins with the patient's turn (index odd — patient, even — therapist); otherwise one can chop the first and/or the last turn to fit the above assumptions. Moreover, we compute DD_w as in (5.6) that represents the averaged absolute difference of the change in speech rate within the same individual. This can be viewed as comparing the acceleration of speech rates.

$$D_w = \frac{1}{K/2} \sum_{k=1}^{K/2} |r_w(2k-1) - r_w(2k)| \tag{5.5}$$

$$\begin{aligned} DD_w = \frac{1}{\frac{K}{2}-1} \sum_{k=1}^{\frac{K}{2}-1} \Big| &\Big(r_w(2k+1) - r_w(2k-1)\Big) \\ &-\Big(r_w(2k+2) - r_w(2k)\Big)\Big| \end{aligned} \tag{5.6}$$

We derive D_s, D_p and DD_s, DD_p in a similar manner. We hypothesize that these cues, which reflect the degree of entrainment by the therapist, should correlate with therapist's empathy level. We show the obtained correlations in Table 5.3. All correlations are significant (based on t-test) at $p < 0.001$ except D_p with

$p < 0.003$, and are in negative values meaning that higher rate-differences associate with lower perceived empathy. This lends support to our hypothesis that the degree of entrainment is linked to therapist's empathy level.

Table 5.3: Correlations between averaged absolute differences of speech rates and therapist empathy

Cues	D_w	D_s	D_p
Corr.	-0.293	-0.259	-0.210
Cues	DD_w	DD_s	DD_p
Corr.	-0.280	-0.234	-0.235

Based on the zero mean turn level speech rates, we compute their standard deviations, *e.g.,* σ_w^T and σ_w^P (word rate deviations) for the therapist and patient respectively, and adopt these as additional behavioral cues. We found significant correlations of value -0.360, -0.311, -0.293 ($p < 10^{-4}$) between σ_w^P, σ_s^P, σ_p^P and empathy codes. However, interestingly, no significant relation was found between therapist's speech rate variations (σ_w^T, σ_s^T, σ_p^T) and empathy. This suggests that an empathic therapist is more capable of regulating a patient's behavioral states such that the conversation goes more smoothly. The mechanism of speech rate regulation in the MI scenario is topic for future in-depth research investigation.

5.5 Analysis of Speech and Silence Durations

The durations of speech and silence are also related to the behavioral states of the interlocutors. We segment the audio as in Sec. 5.4, but retain short speech utterances under 0.5 seconds. We conduct the analysis on the CTT set.

In [46] the ratio of patient utterances correlated with therapist empathy. Here we expand this to include the segment types summarized in Table 5.4. Let the segment durations of a particular type be denoted d_i, $i = 1, 2, \cdots, S$. Let the

total duration of the session be T, which contains $N_{\texttt{seg}}$ segments. For each type we consider four cues: (i) $\sum_{i=1}^{S} d_i/T$, (ii) $S/N_{\texttt{seg}}$, (iii) mean of d_i, (iv) standard deviation of d_i.

We show the correlations between these cues and empathy in Table 5.4. First, we verify that the ratios of therapist and patient speech are negatively and positively correlated with therapist empathy, respectively, as reported in [46]. Second, we find that the ratios of pause have similar correlations to empathy. Since pauses are within speaking turns, one possible interpretation is that therapist who tends to stop then grab the floor more often may seem less empathic. Third, the mean and standard deviation of therapist's pause durations are negatively correlated with empathy, while that for the speech utterances are correlated positively. This suggests that long pauses and short speech utterances may be part of negative behaviors for showing empathy. Short speech utterances like backchannels are mostly annotated as overlapped speech and not analyzed here. In addition, we see that the ratios of gap in both directions are negatively correlated with empathy. This may suggest that high frequency of speaking turn exchange is associated with low empathy.

Table 5.4: Correlations between speech/silence duration cues and therapist empathy: (a) therapist's speech, (b) patient's speech, (c) therapist's pause, (d) patient's pause, (e) gap from therapist to patient, (f) gap from patient to therapist, (g) all pauses, (h) all gaps. **Bold**—$p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, based on t-test

	Cue i	Cue ii	Cue iii	Cue iv
(a)	**−0.255**	**−0.361**	**0.192	**0.192
(b)	**0.305**	**0.362**	*0.141	*0.163
(c)	**−0.374**	**−0.323**	$^{**}-0.222$	**−0.239**
(d)	**0.310**	**0.382**	−0.010	−0.127
(e)	**−0.249**	**−0.236**	−0.081	−0.058
(f)	$^{**}-0.196$	**−0.237**	−0.015	−0.103
(g)	0.0420	**0.212	−0.025	$^{*}-0.164$
(h)	**−0.246**	**−0.237**	−0.052	−0.087

5.6 Experiment of Empathy Classification

We examine if the cues proposed in this Chapter serve as complementary features to the prosodic features introduced in [18] for classifying high *vs.* low empathy codes. The prosodic features are joint distributions of various combinations of quantized speech segment duration, energy, pitch, jitter, and shimmer cues. We select the 100 top-performing features from these in terms of their correlation with empathy codes, based on the training set. We employ the 12-dim cues of speech rate (D_x, DD_x, σ_x^T, σ_x^P, for $x \in \{w, s, p\}$) and 32-dim inter-word and inter-turn duration cues in Table 5.4 as additional features. Moreover, we check the fusion of the above features with lexical cues based on manual transcription, in order to examine the combination of multimodal cues. These lexical cues are those proposed in Chapter 4 based on Maximum Entropy and Maximum Likelihood models.

For the 200 sessions in the CTT set (See Sec. 5.2.2), we conduct a leave-one-therapist-out cross-validation for the 133 unique therapists in the corpus. We use linear SVM as the classifier.

Table 5.5: Accuracies of empathy code classification

Chance level	60.5%
Prosodic cues	72.5%
Speech rate entrainment cues	64.5%
Speaking turn duration cues	72.0%
Prosody + Speech rate + Duration	77.0%
Lexical cues	86.0%
Lexical + Prosody + Speech rate + Duration	91.0%

In Table 5.5 we report the accuracies of empathy code classification (chance level baseline is 60.5%). The fusion of features improves upon each individual feature set, where the differences are all statistically significant at $p < 0.05$. These

results suggest that the speech rate and speech/pause/gap duration features provide additional information about empathy. Fusion of the multimodal features achieved the highest performance.

5.7 Discussion: Reliability Regarding Noise in Speech Alignment

Speech-to-text alignment is important for our analysis, since it provides the various timing information based cues. We have empirically verified the accuracy of the alignment. Here we simulate noise in the alignment results, in order to check how robust our hypotheses are to alignment errors.

To check speech rate entrainment, we add zero mean, σ_z^2 variance Gaussian noise to utterance boundaries in the Switchboard corpus. To check the correlation of speech rate difference and empathy, we add zero mean, σ_z^2 Gaussian noise to the utterance length in the CTT set. Like in Sec. 5.4, we eliminate utterances shorter than 0.5 seconds after adding the noise. For both cases, we sample σ_z from 0 to 1 second with a step size of 0.02 seconds. We repeat the simulation 100 times and take the averaged correlation values.

In Figure 5.2 and Figure 5.3 we plot the correlations. We see that the results are still significant near $\sigma_z = 0.5$, and the degradations of correlations are negligible for $\sigma_z < 0.2$. These demonstrate that the above hypotheses are robust to alignment precision.

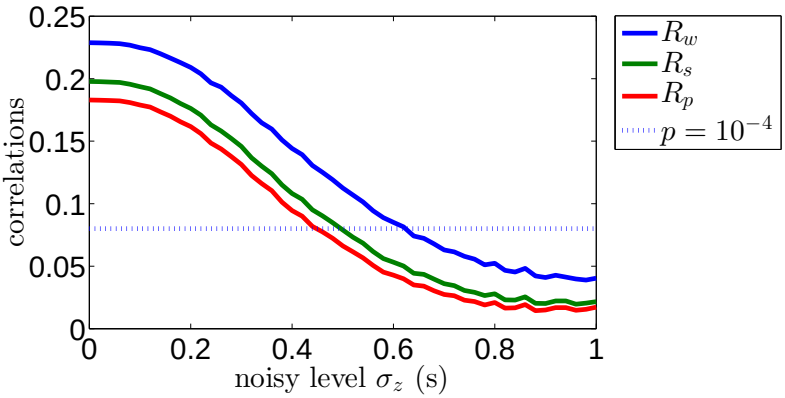

Figure 5.2: Correlations of interlocutors' speech rates in simulation of noisy utterance boundaries

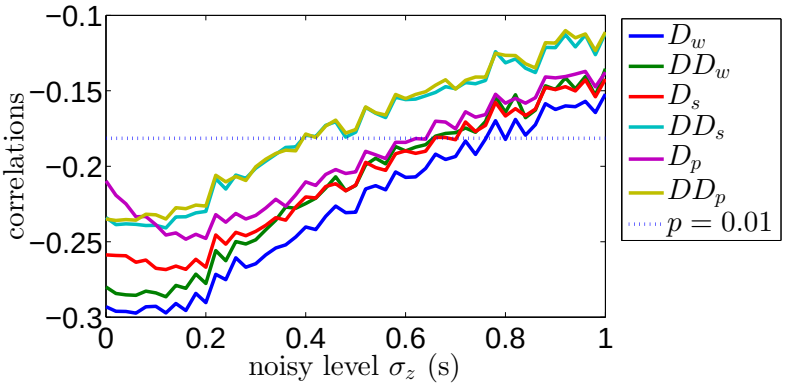

Figure 5.3: Correlations of speech rate differences and empathy in simulation of noisy utterance lengths

5.8 Conclusion

In this Chapter we extracted word, syllable, and phoneme rates for interlocutors engaged in telephone conversation and addiction-counseling spoken interactions. Through statistical analyses, we showed the entrainment of interlocutors' speech rates by their positive session-wise correlations. The degree of entrainment — captured by the averaged absolute differences of turn-level speech rates of the therapist and patient — correlates with therapist's empathy rating. These relations were further verified to be robust in a simulation of noisy speech-text alignment. Moreover,

we tested the correlation of ratio and duration statistics of speech, pause, and gap segments, with therapist's empathy rating. Furthermore, we employed these cues in an experiment classifying high *vs.* low empathy codes. Results showed speech rate, inter-word pause and inter-turn gap provided useful information, complementing previous prosodic cues for empathy modeling. Fusion of lexical, prosodic, entrainment, and turn taking cues achieved the best performance.

In the future we plan to model speech rate dynamics in more detail. This might require a joint consideration of entrainment with other factors including turn taking dynamics, and the interlocutor emotional state. For modeling of empathy, we will further investigate the role of vocal cues in both empathy expression and perception. We will also work on ways to effectively fuse the various cues for more accurate modeling.

Chapter 6

Conclusion and Future Work

This dissertation has studied prosodic, lexical, speech rate entrainment, and turn taking cues to model therapist empathy, and to predict expert assessment of empathy. Experiment results show that the above cues based on speech and language processing provide useful information about therapist empathy. Their relations to empathy are represented by the correlation to expert-provided empathy code values, as well as the accuracy of binary classification of high *vs.* low empathy codes. In general, lexical cues are the most prominent indicating empathy, followed by the prosodic cues and the entrainment cues. This may suggest that although entrainment links to empathy most broadly, it manifests in many ways of behavioral expressions, so that one type of feature is not enough to represent the relation of entrainment and empathy. Language may be more useful to evaluate empathy in a particular application, as it is an abstract form representing human interpretable semantic meanings; however, the model learned in one field may not directly be applicable to other fields since the language in other scenarios are different. On the contrary, entrainment cues, though not strongly correlated with empathy as lexical cues, may be more generic in other human interaction scenarios.

Findings in this dissertation point out that modeling and assessing therapist empathy through automatic signal processing is possible. Development of such a system may contribute to large scale evaluation of psychotherapy in an objective, evidence-driven manner. In addition, these findings may be useful in empathy simulation for a more human-like computer agent in human-computer interaction.

There are several directions to further develop the research on empathy modeling. Firstly, there are other behavioral modalities such as facial expression, gestures, and physiological measurements. Data in these modalities are not available for the current study, but may be included in future collection and study. Empathy is not constant along the session of interaction; the moments that the client needs empathic response may be identified as empathic opportunities. Locating these empathic opportunities and tracking the response by the care-provider may indicate a more authentic feeling of empathy that the client perceives. The study of empathy modeling in addiction counseling may be transfered to other mental or physical health care scenarios, and more broadly, human interactions such as education, customer care, and family interaction.

Reference List

[1] B. Xiao, Z. E. Imel, P. Georgiou, D. C. Atkins, and S. S. Narayanan, "Computational Analysis and Simulation of Empathic Behaviors — A Survey of Empathy Modeling with Behavioral Signal Processing Framework," *Accepted to Current Psychiatry Reports*, 2016.

[2] E. B. Titchener, *Lectures on the experimental psychology of the thought-processes*. Macmillan, 1909.

[3] M. L. Hoffman, *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, 2001.

[4] C. D. Batson, "These things called empathy: eight related but distinct phenomena," *The social neuroscience of empathy*, pp. 3–15, 2009.

[5] B. M. Cuff, S. J. Brown, L. Taylor, and D. J. Howat, "Empathy: a review of the concept," *Emotion Review*, 2014.

[6] J. Decety and P. Jackson, "The functional architecture of human empathy," *Behavioral and cognitive neuroscience reviews*, vol. 3, no. 2, pp. 71–100, 2004.

[7] R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg, "Empathy," *Psychotherapy*, vol. 48, no. 1, pp. 43–49, 2011.

[8] S. D. Preston and F. De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, vol. 25, no. 01, pp. 1–20, 2002.

[9] F. De Vignemont and T. Singer, "The empathic brain: how, when and why?" *Trends in cognitive sciences*, vol. 10, no. 10, pp. 435–441, 2006.

[10] M. Iacoboni, "Imitation, empathy, and mirror neurons," *Annual review of psychology*, vol. 60, pp. 653–670, 2009.

[11] S. Lelorain, A. Brdart, S. Dolbeault, and S. Sultan, "A systematic review of the associations between empathy measures and patient outcomes in cancer care," *Psycho-Oncology*, vol. 21, no. 12, pp. 1255–1264, 2012.

[12] F. Derksen, J. Bensing, and A. Lagro-Janssen, "Effectiveness of empathy in general practice: a systematic review," *British Journal of General Practice*, vol. 63, no. 606, pp. e76–e84, 2013.

[13] T. B. Moyers and W. R. Miller, "Is low therapist empathy toxic?" *Psychology of Addictive Behaviors*, vol. 27, no. 3, p. 878, 2013.

[14] E. T. van Berkhout and J. M. Malouff, "The Efficacy of Empathy Training: A Meta-Analysis of Randomized Controlled Trials." *Journal of Counseling Psychology*, 2015.

[15] G. T. Barrett-Lennard, "The empathy cycle: Refinement of a nuclear concept," *Journal of Counseling Psychology*, vol. 28, no. 2, p. 91, 1981.

[16] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[17] S. Narayanan and P. Georgiou, "Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language," *Proceeding of IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[18] B. Xiao, D. Bone, M. Van Segbroeck, Z. E. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Modeling Therapist Empathy through Prosody in Drug Addiction Counseling," in *Proc. Interspeech*, Sep. 2014, pp. 213–217.

[19] B. Xiao, Z. E. Imel, P. Georgiou, D. C. Atkins, and S. S. Narayanan, ""Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PLoS One*, vol. 10, no. 12, Dec. 2015.

[20] B. Xiao, Z. E. Imel, D. Atkins, P. Georgiou, and S. S. Narayanan, "Analyzing Speech Rate Entrainment and Its Relation to Therapist Empathy in Drug Addiction Counseling," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015.

[21] H. Riess, "Biomarkers in the psychotherapeutic relationship: the role of physiology, neurobiology, and biological correlates of E.M.P.A.T.H.Y." *Harvard Review of Psychiatry*, vol. 19, no. 3, pp. 162–174, 2011.

[22] C. Regenbogen, D. A. Schneider, A. Finkelmeyer, N. Kohn, B. Derntl, T. Kellermann, R. E. Gur, F. Schneider, and U. Habel, "The differential contribution of facial expressions, prosody, and speech content to empathy," *Cognition & emotion*, vol. 26, no. 6, pp. 995–1014, 2012.

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[24] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford Press, 2012.

[25] W. R. Miller and G. S. Rose, "Toward a theory of motivational interviewing," *American Psychologist*, vol. 64, no. 6, p. 527, 2009.

[26] T. Moyers, T. Martin, J. Manuel, W. Miller, and D. Ernst, *Revised global scales: Motivational Interviewing Treatment Integrity 3.0*, 2007.

[27] S. Kumano, K. Otsuka, M. Matsuda, and J. Yamato, "Analyzing perceived empathy/antipathy based on reaction time in behavioral coordination," in *Automatic Face and Gesture Recognition*. IEEE, 2013, pp. 1–8.

[28] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, 2012.

[29] S. W. McQuiggan and J. C. Lester, "Modeling and evaluating empathy in embodied companion agents," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 348–360, 2007.

[30] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, "A computational model of empathy: Empirical evaluation," in *Proc. ACII*. IEEE, 2013, pp. 1–6.

[31] M. Ochs, D. Sadek, and C. Pelachaud, "A formal model of emotions for an empathic rational dialog agent," *Autonomous Agents and Multi-Agent Systems*, vol. 24, no. 3, pp. 410–440, 2012.

[32] S. H. Rodrigues, S. Mascarenhas, J. a. Dias, and A. Paiva, "A Process Model of Empathy For Virtual Agents," *Interacting with Computers*, 2014.

[33] J. L. Coulehan, F. W. Platt, B. Egener, R. Frankel, C.-T. Lin, B. Lown, and W. H. Salazar, "Let Me See If I Have This Right : Words That Help Build Empathy," *Annals of Internal Medicine*, vol. 135, no. 3, pp. 221–227, 2001.

[34] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 181–184.

[35] B. Xiao, D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy," in *Proc. APSIPA ASC*, Dec. 2012, pp. 1–4.

[36] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[37] S. N. Chakravarthula, B. Xiao, Z. E. Imel, D. C. Atkins, and P. Georgiou, "Assessing Empathy using Static and Dynamic Behavior Models based on Therapists Language in Addiction Counseling," in *Proc. Interspeech*, Dresden, Sep. 2015.

[38] J. W. Pennebaker, R. J. Booth, and M. E. Francis, *Linguistic Inquiry and Word Count (LIWC)*, 2007. [Online]. Available: http://www.liwc.net/

[39] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2379–2392, 2013.

[40] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. Narayanan, "Predicting Therapist Empathy in Motivational Interviews using Language Features Inspired by Psycholinguistic Norms," in *Proc. Interspeech*, Dresden, Sep. 2015.

[41] S. P. Lord, E. Sheng, Z. E. Imel, J. Baer, and D. C. Atkins, "More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client," *Behavior therapy*, vol. 46, no. 3, pp. 296–303, 2015.

[42] L. Aziz-Zadeh, T. Sheng, and A. Gheytanchi, "Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability," *PLoS One*, vol. 5, no. 1, pp. 1–8, 2010.

[43] E. Weiste and A. Perkyl, "Prosody and empathic communication in psychotherapy interaction," *Psychotherapy Research*, pp. 1–15, 2014.

[44] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Prentice hall, Mar. 2010.

[45] Z. E. Imel, J. S. Barco, H. J. Brown, B. R. Baucom, J. S. Baer, J. C. Kircher, and D. C. Atkins, "The association of therapist empathy and synchrony in vocally encoded arousal." *Journal of counseling psychology*, vol. 61, no. 1, p. 146, 2014.

[46] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. S. Narayanan, "Modeling Therapist Empathy and Vocal Entrainment in Drug Addiction Counseling," in *Proc. Interspeech*, Sep. 2013, pp. 2861–2865.

[47] C. M. Bishop, *Pattern recognition and machine learning*. Springer, Oct. 2007.

[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[49] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn, "FERA 2015-second facial expression recognition and analysis challenge," *Proc. IEEE ICFG*, 2015.

[50] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," in *Automatic Face and Gesture Recognition*. IEEE, 2011, pp. 43–50.

[51] K. P. Murphy, "Dynamic bayesian networks: representation, inference and learning," Ph.D. dissertation, University of California, Berkeley, 2002.

[52] S. Kumano, K. Otsuka, M. Matsuda, and J. Yamato, "Analyzing Perceived Empathy Based on Reaction Time in Behavioral Mimicry," *IEICE Transactions on Information and Systems*, vol. 97, no. 8, pp. 2008–2020, 2014.

[53] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Analyzing Interpersonal Empathy via Collective Impressions," *IEEE Transactions on Affective Computing*, no. 99, 2015.

[54] A. McAllister, J. Sundberg, and S. R. Hibi, "Acoustic Measurements and Perceptual Evaluation of Hoarseness in Children's Voices," *Logopedics Phoniatrics Vocology*, vol. 23, 1998.

[55] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[56] B. Z. Pollermann, "A Place for Prosody in a Unified Model of Cognition and Emotion," in *Proc. Speech Prosody*, 2002.

[57] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of substance abuse treatment*, vol. 37, no. 2, p. 191, 2009.

[58] T. Moyers, T. Martin, J. Manuel, and W. Miller, "The motivational interviewing treatment integrity (MITI) code: Version 2.0," *University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions. Albuquerque, NM*, 2008.

[59] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[60] M. Brookes and others, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, 1997.

[61] M. Van Segbroeck, A. Tsiartas, and S. S. Narayanan, "A Robust Frontend for VAD: Exploiting Contextual, Discriminative and Spectral Cues of Human Voice," in *Proc. InterSpeech*, Lyon, France, Aug. 2013, pp. 704–708.

[62] H. Van hamme, "Robust Speech Recognition using Cepstral Domain Missing Data Techniques and Noisy Masks," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 213–216.

[63] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[64] R. V. Hogg and E. A. Tanis, "Probability and Statistical Inference." Pearson Prentice Hall, 2009, pp. 379–389.

[65] Substance Abuse and Mental Health Services Administration, *Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings. NSDUH Series H-46, HHS Publication No. (SMA) 13-4795. Rockville, MD, U.S.A.*, 2013.

[66] T. B. Moyers, T. Martin, J. K. Manuel, S. M. Hendrickson, and W. R. Miller, "Assessing competence in the use of motivational interviewing," *Journal of Substance Abuse Treatment*, vol. 28, no. 1, pp. 19–26, 2005.

[67] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[68] P. Georgiou, M. Black, A. Lammert, B. Baucom, and S. Narayanan, "That's aggravating, very aggravating: Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proc. ACII*, 2011, pp. 87–96.

[69] B. Xiao, P. G. Georgiou, B. R. Baucom, and S. S. Narayanan, "Power-spectral analysis of head motion signal for behavioral modeling in human interaction," in *Proc. ICASSP*, May 2014, pp. 4593–4597.

[70] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[71] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," in *Proc. ICME*. IEEE, 2013, pp. 1–6.

[72] T. Guha, Z. Yang, A. Ramakrishna, R. Grossman, D. Hedley, S. Lee, and S. Narayanan, "On Quantifying Facial Expression-related Atypicality of Children with Autism Spectrum Disorder," in *Proc. ICASSP*. Brisbane, Australia: IEEE, Apr. 2015, pp. 803–807.

[73] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schrder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[74] W. Wang, P. Lu, and Y. Yan, "An improved hierarchical speaker clustering," *ACTA ACUSTICA*, vol. 33, no. 1, p. 9, 2008.

[75] C. W. Huang, B. Xiao, P. Georgiou, and S. Narayanan, "Unsupervised Speaker Diarization Using Riemannian Manifold Clustering," in *Proc. Interspeech*, Sep. 2014, pp. 567–571.

[76] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, Dec. 2011.

[77] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1. IEEE, 1992, pp. 517–520, http://www.isip.piconepress.com/projects/switchboard/.

[78] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[79] A. Stolcke, "SRILM An Extensible Language Modeling Toolkit," in *Proc. Interspeech*, 2002.

[80] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[81] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modelling," *Computer Speech & Language*, vol. 10, no. 3, pp. 187–228, 1996.

[82] L. Zhang, *Maximum Entropy Modeling Toolkit for Python and C++*, 2013. [Online]. Available: https://github.com/lzhang10/maxent

[83] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[84] P. Roy-Byrne, K. Bumgardner, A. Krupski, C. Dunn, R. Ries, D. Donovan, I. I. West, C. Maynard, D. C. Atkins, M. C. Graves, and others, "Brief intervention for problem drug use in safety-net primary care settings: a randomized clinical trial," *JAMA*, vol. 312, no. 5, pp. 492–501, 2014.

[85] S. J. Tollison, C. M. Lee, C. Neighbors, T. A. Neil, N. D. Olson, and M. E. Larimer, "Questions and reflections: the use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students," *Behavior Therapy*, vol. 39, no. 2, pp. 183–194, 2008.

[86] C. Neighbors, C. M. Lee, D. C. Atkins, M. A. Lewis, D. Kaysen, A. Mittmann, N. Fossos, I. M. Geisner, C. Zheng, and M. E. Larimer, "A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations," *Journal of consulting and clinical psychology*, vol. 80, no. 5, p. 850, 2012.

[87] C. M. Lee, J. R. Kilmer, C. Neighbors, D. C. Atkins, C. Zheng, D. D. Walker, and M. E. Larimer, "Indicated prevention for college student marijuana use: a randomized controlled trial," *Journal of consulting and clinical psychology*, vol. 81, no. 4, p. 702, 2013.

[88] C. M. Lee, C. Neighbors, M. A. Lewis, D. Kaysen, A. Mittmann, I. M. Geisner, D. C. Atkins, C. Zheng, L. A. Garberson, J. R. Kilmer, and others, "Randomized controlled trial of a Spring Break intervention to reduce high-risk drinking," *Journal of consulting and clinical psychology*, vol. 82, no. 2, p. 189, 2014.

[89] Z. E. Imel, M. Steyvers, and D. C. Atkins, "Computational psychotherapy research: Scaling up the evaluation of patientprovider interactions," *Psychotherapy*, vol. 52, no. 1, pp. 19–30, 2015.

[90] D. Can, P. Georgiou, D. Atkins, and S. S. Narayanan, "A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using Linguistic Features," in *Proc. Interspeech*, Portland, Sep. 2012, pp. 2254–2257.

[91] T. Wheatley, O. Kang, C. Parkinson, and C. Looser, "From Mind Perception to Mental Connection: Synchrony as a Mechanism for Social Understanding," *Social and Personality Psychology Compass*, vol. 6, no. 8, pp. 589–606, 2012.

[92] T. Arizmendi, "Linking mechanisms: Emotional contagion, empathy, and imagery," *Psychoanalytic Psychology*, vol. 28, no. 3, p. 405, 2011.

[93] J. B. Bavelas, A. Black, C. R. Lemery, and J. Mullett, "Motor mimicry as primitive empathy," *Empathy and its Development*, p. 317, 1990.

[94] B. Guitar and L. Marchinkoski, "Influence of mothers' slower speech on their children's speech rate," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 4, pp. 853–861, 2001.

[95] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.