A technology prototype system for rating therapist empathy from audio recordings in addiction counseling

Bo Xiao[1], Chewei Huang[1], Zac E. Imel[2], David C. Atkins[3], Panayiotis Georgiou[1] and Shrikanth S. Narayanan[1]

[1] Department of Electrical Engineering, University of Southern California, Los Angeles, CA, United States
[2] Department of Educational Psychology, University of Utah, Salt Lake City, UT, United States
[3] Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA, United States

ABSTRACT

Scaling up psychotherapy services such as for addiction counseling is a critical societal need. One challenge is ensuring quality of therapy, due to the heavy cost of manual observational assessment. This work proposes a speech technology-based system to automate the assessment of therapist empathy—a key therapy quality index—from audio recordings of the psychotherapy interactions. We designed a speech processing system that includes voice activity detection and diarization modules, and an automatic speech recognizer plus a speaker role matching module to extract the therapist's language cues. We employed Maximum Entropy models, Maximum Likelihood language models, and a Lattice Rescoring method to characterize high *vs.* low empathic language. We estimated therapy-session level empathy codes using utterance level evidence obtained from these models. Our experiments showed that the fully automated system achieved a correlation of 0.643 between expert annotated empathy codes and machine-derived estimations, and an accuracy of 81% in classifying high *vs.* low empathy, in comparison to a 0.721 correlation and 86% accuracy in the oracle setting using manual transcripts. The results show that the system provides useful information that can contribute to automatic quality insurance and therapist training.

Submitted 19 November 2015
Accepted 29 March 2016
Published 20 April 2016

Corresponding author
Bo Xiao, xiaobothu@gmail.com

Academic editor
Julita Vassileva

Additional Information and Declarations can be found on page 20

DOI 10.7717/peerj-cs.59

Subjects Human–Computer Interaction, Computational Linguistics, Emerging Technologies, Natural Language and Speech
Keywords Prototype system, Speech processing, Language modeling, Addiction counseling, Empathy modeling, Automatic speech recognition

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. A national survey in the United States by the Substance Abuse and Mental Health Services Administration showed that there were 23.9 million illicit drug users in 2012. However, only 2.5 million persons received treatment at a specialty facility (*Substance Abuse and Mental Health Services Administration, 2013*). 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 (*Moyers et al., 2005*). The lack of human and time resources prohibits the evaluation of psychotherapy in large scale; 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 work, 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 (*Miller & Rollnick, 2012*). MI has been proved effective in various clinical trials; and theories about its mechanisms have been developed (*Miller & Rose, 2009*). 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. Empathy mainly encompasses two aspects in the MI scenario: the therapist's internalization of a patient's thoughts and feelings, i.e., taking the perspective of the patient; and the therapist's response with the sensitivity and care appropriate to the suffering of the patient, i.e., feeling for the patient (*Batson, 2009*). Empathy is an evolutionarily acquired basic human ability, and a core factor in human interactions. Physiological mechanisms on single-cell and neural-system levels lend support to the cognitive and social constructs of empathy (*Preston & De Waal, 2002*; *Iacoboni, 2009*; *Eisenberg & Eggum, 2009*). Higher ratings of therapist empathy are associated with treatment retention and positive clinical outcomes (*Elliott et al., 2011*; *Miller & Rose, 2009*; *Moyers & Miller, 2013*). Therefore, we choose to computationally quantify empathy.

The study of the techniques that support the measurement, analysis, and modeling of human behavior signals is referred to as Behavioral Signal Processing (BSP) (*Narayanan & Georgiou, 2013*). 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 (*Black et al., 2013*; *Georgiou et al., 2011*; *Xiao et al., 2014*), and the use of vocal prosody and facial expressions in understanding behavioral characteristics in Autism Spectrum Disorders (*Lee et al., 2014*; *Bone et al., 2014*; *Metallinou, Grossman & Narayanan, 2013*; *Guha et al., 2015*). Closely related to BSP, Social Signal Processing studies modeling, analysis and synthesis of human social behavior through multimodal signal processing (*Vinciarelli et al., 2012*).

Computational models of empathy essentially explore the relation between low level behavior signals and high level human judgments, with the guidance of domain theories and data from real applications. Our previous work on empathy modeling has used lexical, vocal similarity, and prosodic cues. We have shown that lexical features derived from empathic *vs.* generic language models are correlated with expert annotated therapist empathy ratings (*Xiao et al., 2012*). We modeled vocal entrainment between the therapist and patient through acoustic similarity measures, and demonstrated that there is a link between similarity measures and empathy ratings (*Xiao et al., 2013*). We also quantified prosodic features of the therapist and patient, and classified high *vs.* low empathy ratings using features derived from the prosodic pattern distributions (*Xiao et al., 2014*). In related work, Kumano et al. have employed multimodal behavior cues including facial expression, gaze, speech activity, head gestures, and response timing information, with Bayesian probabilistic models to predict the perceived empathy level in group conversations (*Kumano et al., 2011*; *Kumano et al., 2013*). These works demonstrate the feasibility of computationally modeling empathy through multimodal cues.

However, empathy estimation in the aforementioned work 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 work, we propose a prototype system that satisfies this requirement. This paper focuses on the computational aspect; for more discussion in a psychological aspect, please refer to *Xiao et al. (2015)*.

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 Fig. 1. The Voice Activity Detection (VAD) module separates speech from non-speech (when they speak). The diarization module separates speakers in the interaction (who is speaking). The Automatic Speech Recognition (ASR) system decodes spoken words from the audio (what they say). And we 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 paper, 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.

In the rest of the paper, '*Automatic speech recognition*' describes the modules and methods in the automatic transcription system. Then 'Therapist Empathy Models using Language Cues' describes the lexical modeling of empathy. 'Data Corpora' introduces the real application data utilized in this work. 'System Implementation' describes the

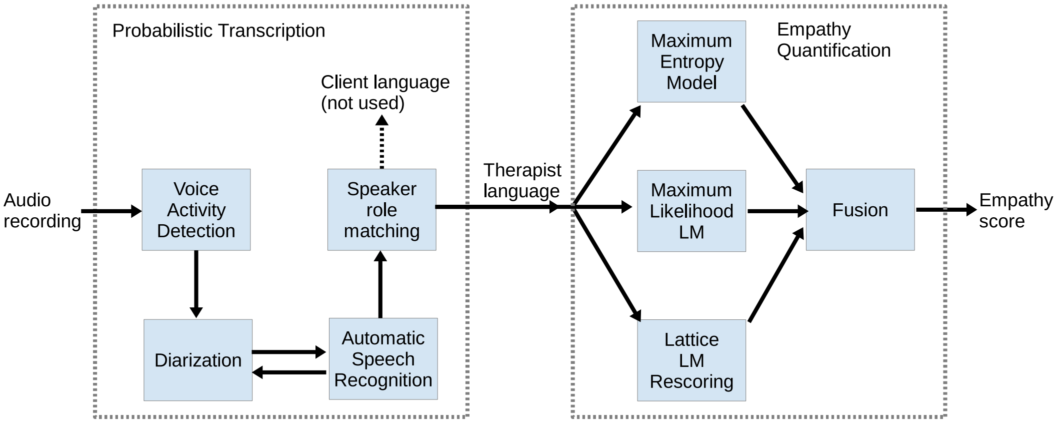

Figure 1 **Overview of modules in the system, including VAD, Diarization, ASR, speaker role matching, and therapist language modeling for empathy prediction.**

system implementation; and 'Experiment and Results' reports the experimental results. 'Discussion' discusses the findings of this work. Finally, 'Conclusion' concludes the paper.

AUTOMATIC SPEECH RECOGNITION

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, Tsiartas & Narayanan (2013)*. 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. This binarization threshold increases from 0.5, until that all segments are shorter than an upper bound of segment length (e.g., 60 s). 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., 0.1 s) and the combined segment does not exceed the upper bound of segment length (e.g., 60 s). After the merging we drop segments that are too short (e.g., less than 1 s).

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 *Wang, Lu & Yan (2008)*. The second method adopts GLR speaker segmentation and Riemannian manifold method for speaker clustering, as implemented in *Huang et al. (2014)*. This method slices each GLR derived segment into short-time segments (e.g., 1 s), so as to increase the number of samples in the manifold space for more robust clustering (see *Huang et al. (2014)* 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 'System Implementation.'

ASR

We decided to train an ASR using speech recordings from in-domain data corpora that were collected in real psychotherapy settings. These recordings may best match the acoustic conditions (possibly noisy and heterogeneous) in the target application. In this work, a large vocabulary, continuous speech recognizer (LVCSR) is implemented using the Kaldi library (*Povey et al., 2011*).

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

Dictionary: We combine the lexicon in Switchboard (*Godfrey, Holliman & McDaniel, 1992*) and WSJ (*Paul & Baker, 1992*) 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 out of vocabulary 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 (*Stolcke, 2002*). 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 'Data Corpora') 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.

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 possible way to automatically match roles to the speakers with minimal assumptions about the data collection procedures is by 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 (*Stolcke, 2002*).
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) (1) holds—we match S_1 to therapist and S_2 to patient; (ii) (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.

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

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

5. Outliers: When the diarization module outputs highly biased results 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.

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, and test the models on clean or ASR decoded noisy text. One approach is based on Maximum Likelihood N-gram Language Models (LMs) of high *vs.* low empathy respectively, previously employed in *Xiao et al. (2012)* and in *Georgiou et al. (2011)* for a similar problem; we adopt this method for its simplicity and effectiveness. Additional modeling approaches may be complementary to increase the accuracy of empathy prediction; for this reason we adopt a widely applied method—Maximum Entropy model, which has shown good performance in a variety of natural language processing tasks. Moreover, in order to improve the test performance on ASR decoded text, it is possible to evaluate an ensemble of noisy text hypotheses through rescoring the decoding lattice with high *vs.* low empathy LMs. In this way empathy relevant words in the decoding hypotheses gain more weights so that they become stronger features. Without rescoring, it is likely that these words do not contribute to the modeling due to their absence in the best paths of lattices. In this work, we employ the above three approaches and their fusion in the system.

For each session, we first infer therapist empathy at the utterance level, then integrate the local evidence toward session level empathy estimation. We discuss more about the modeling strategies in 'Empathy modeling strategies.' The details of the proposed methods are described as follows.

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 (*Berger, Pietra & Pietra, 1996*; *Rosenfeld, 1996*). We train a two-class (high *vs.* low empathy) MaxEnt model on utterance level data using the MaxEnt toolkit in *Zhang (2013)*.

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 jth 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 (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 (*Liu & Nocedal, 1989*).

$$P_n(Y|u) = \frac{1}{Z(u)} \exp\left(\sum_j \lambda_n^j f_n^j(u,Y)\right). \tag{3}$$

Based on the trained MaxEnt model, averaging utterance level evidence $P_n(H|u)$ gives the session level empathy score α_n, as shown in (4), where $\mathbf{U}_T$ is the set of K therapist utterances.

$$\alpha_n(\mathbf{U}_T) = \frac{1}{K}\sum_{i=1}^{K} P_n(H|u_i), \quad \mathbf{U}_T = \{u_1, u_2, \ldots, u_K\},\ n = 1,2,3. \tag{4}$$

Maximum likelihood based model

Maximum Likelihood language models (LM) based on N-grams 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, the posterior probability $P(H|u)$ is formulated by the likelihoods as in (5), where we assume equal prior probabilities $P(H) = P(L)$.

$$P(H|u) = \frac{P(u|H)P(H)}{P(u|H)P(H) + P(u|L)P(L)} = \frac{P(u|H)}{P(u|H) + P(u|L)}. \tag{5}$$

We train the high empathy LM ($\mathtt{LM}_H$) and low empathy LM ($\mathtt{LM}_L$) 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 (*Stolcke, 2002*). Next, for robustness we mix a large in-domain LM (e.g., the final LM in ASR) to $\mathtt{LM}_H$ and $\mathtt{LM}_L$ with a small weight (e.g., 0.1). Let us denote the mixed LMs as $\mathtt{LM}'_H$ and $\mathtt{LM}'_L$.

For the inference of $P(H|u)$, we first compute the log-likelihoods $l_n(u|H)$ and $l_n(u|L)$ by applying $\mathtt{LM}'_H$ and $\mathtt{LM}'_L$, where $n = 1,2,3$ are the utilized n-gram orders. Then $P_n(H|u)$ is obtained as in (6).

$$P_n(H|u) = \frac{e^{l_n(u|H)}}{e^{l_n(u|H)} + e^{l_n(u|L)}}. \tag{6}$$

We compute session level empathy score β_n as the average of utterance level evidence as shown in (7), where $\mathbf{U}_T$ is the same as in (4).

$$\beta_n(\mathbf{U}_T) = \frac{1}{K}\sum_{i=1}^{K} P_n(H|u_i). \tag{7}$$

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 it 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 'Robustness of empathy modeling methods'). An illustration of the lattice paths re-ranking effect is shown in Fig. 2.

0. **Input**: ASR decoded lattice $\mathcal{L}$, high and low empathy LMs $\mathtt{LM}'_H$, $\mathtt{LM}'_L$ as described in 'Maximum likelihood based model'.
1. Update the LM scores in $\mathcal{L}$ by applying $\mathtt{LM}'_H$ and $\mathtt{LM}'_L$ 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, $1 \le r \le R$, respectively.

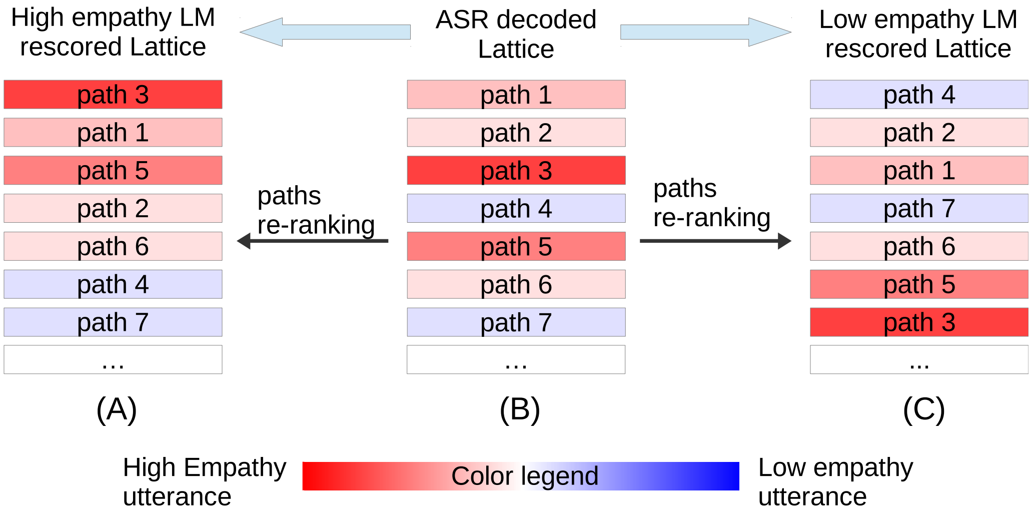

Figure 2 **Illustration of lattice rescoring by high/low empathy LMs.** (B) represents the ASR decoded lattice, where each row denotes a path in the lattice, ranked by their scores. Each path is color-coded to show the empathy degree. High/low empathy LM rescoring produce two new lattices shown in (A) and (C), respectively. The paths are re-ranked based on their new scores.

4. Compute the utterance level empathy score $S_H(\mathcal{L})$ as in (8)

$$S_H(\mathcal{L}) = \frac{\exp\left(\frac{1}{R}\sum_{r=1}^{R} s_H(r)\right)}{\exp\left(\frac{1}{R}\sum_{r=1}^{R} s_H(r)\right) + \exp\left(\frac{1}{R}\sum_{r=1}^{R} s_L(r)\right)}. \tag{8}$$

5. Compute the session level empathy score γ as in (9), where $\mathcal{U}_T$ is the set of K lattices of therapist utterances.

$$\gamma(\mathcal{U}_T) = \frac{1}{K}\sum_{i=1}^{K} S_H(\mathcal{L}_i). \tag{9}$$

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

Note that the Lattice Rescoring method is a natural extension of the Maximum Likelihood LM method in 'Maximum likelihood based model'. When the score $s_H(r)$ denotes log-likelihood and $R = 1$, (8) becomes equivalent to (6). In that case $S_H(\mathcal{L})$ represents a 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 paths explicitly. It also allows more efficient averaging of the evidence from the top hypotheses.

DATA CORPORA

This section introduces the three data corpora used in the study.

- "TOPICS" corpus—153 audio-recorded MI sessions randomly selected from 899 sessions in five psychotherapy studies (*Roy-Byrne et al., 2014*; *Tollison et al., 2008*;

Table 1 Details about the data corpora employed, including counts of session, talk turn, and word token, and also total time duration.

Corpus	No. sessions	No. talk turns	No. word tokens	Duration
TOPICS	153	3.69×10^4	1.12×10^6	104.2 h
Gen. Psyc.	1,200	3.01×10^5	6.55×10^6	–
CTT	200	2.40×10^4	6.24×10^5	68.6 h

Neighbors et al., 2012; *Lee et al., 2013*; *Lee et al., 2014*), 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, sigh, and other nonverbal vocalizations. Session length ranges from 20 min to 1 h.

- "General Psychotherapy" corpus—transcripts of 1,200 psychotherapy sessions in MI and a variety of other treatment types (*Imel, Steyvers & Atkins, 2015*). 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) (*Baer et al., 2009*). 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 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 (*Moyers et al., 2007*). 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* (SPs), whose behaviors were regulated for therapist training and evaluation purposes. For example, SP sessions had pre-scripted situations. Sessions involving an SP or a *Real Patient* (RP) were about the same size in the entire corpus. The 200 sessions used in this study

Table 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%

Table 3 Summary of data corpora usage in the training and test.

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

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 4.5–7, with mean of 5.90 ± 0.58. Table 2 shows the counts of high *vs.* low empathy and SP *vs.* RP sessions. Moreover, the selected sessions are diverse in the therapists involved. There are 133 unique therapists, and any therapist has no more than three sessions.

SYSTEM IMPLEMENTATION

In this section, we describe the system implementation in more detail. Table 3 summarizes the usage of data corpora in various modeling and application steps.

VAD: We construct the VAD training and development sets by sampling from the TOPICS corpus. The total length of the two sets are 5.2 h and 2.6 h, 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 'Voice activity detection,' 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 *Wang, Lu & Yan (2008)*.
2. Result D_2: apply the Riemannian clustering method in *Huang et al. (2014)*.
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 *Huang et al. (2014)* 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 'Speaker diarization', 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. 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 'Maximum likelihood rescoring on ASR decoded lattices,' we employ the top 100 paths ($R = 100$).

Empathy model fusion: The three methods in 'Therapist Empathy Models using Language Cues' 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.

EXPERIMENT AND RESULTS

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 'Automatic Speech Recognition' and empathy modeling in 'Therapist Empathy Models using Language Cues'.

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.

Table 4 Session-wise average performance of VAD and diarization modules.

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

Table 5 Overall 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

ASR system performance

Table 4 reports false alarm, miss, speaker error rate (for diarization only), and total error rate for the VAD and diarization modules. These results are the averages of session-wise values. 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.

Table 5 reports averaged ASR performance in terms of substitution, deletion, insertion, and total Word Error Rate (WER) for the case of ORA-D and 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.

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.

Empathy code estimation performance

Table 6 shows 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 'Maximum Entropy model'. 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 0–1; the RMSE σ is in the space of empathy codes (1–7); and the classification accuracy Acc is in percentage.

Table 6 **Empathy code estimation performance using the 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

Table 7 **Empathy code estimation performance using Maximum Likelihood LM.**

	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

Similarly, Table 7 shows the results by the fusion of empathy scores β_n, $n = 1,2,3$, derived by the n-gram LMs in 'Maximum likelihood based model'. From the results in Tables 6 and 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 in the condition of sparse training data. Thus, the influence of noisy input is also heavier for the MaxEnt model.

Table 8 shows the results using the empathy score γ that is derived by the lattice LM rescoring method in 'Maximum likelihood rescoring on ASR decoded lattices', 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 9 shows 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. The fully automatic system achieves 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 Services takes kid away from mother who then comes to psychotherapy), while RP sessions are not scripted, and the topics tend to be diverse.

Table 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

Table 9 Empathy code estimation performance by the fusion of the MaxEnt, Maximum Likelihood LM, 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

It is interesting and surprising that in the ORA-D and AUTO cases, the system performs relatively well given a WER above 40%. One reason might be that the distribution of session level WER is skewed to the lower end—the median WER (39.9%) is lower than the mean WER (43.6%). More importantly, ASR errors are likely independent to the high *vs.* low representation of empathy, i.e., noises in the transcripts are probably not biased towards higher or lower empathy in general. Even though the domain information in the observation is attenuated, its polarity of high *vs.* low empathy remains unchanged. Nevertheless, this does not mean that ASR errors have no effect on the performance. We found that the dynamic range of the predicted empathy scores is smaller and more centered in the ORA-D and AUTO cases, showing a reduced discriminative power.

There are some seemingly counter-intuitive results regarding *Acc*; e.g., in Table 9 ORA-D outperforms ORA-T in *Acc* for the RP sessions. Firstly, due to the small sample size of RP sessions, the difference of prediction accuracies in this case is not statistically significant ($p > 0.05$). This means the comparison is likely to be influenced by random effects. Moreover, as discussed above, though noisy text in the ORA-D case attenuates the representation of empathy, such effect is less critical for binary classification since it only concerns the polarity of high *vs.* low empathy rather than the actual degree. In Table 9, ORA-D has slightly higher σ than ORA-T. This shows that ORA-D does not exceed ORA-T in the estimation of empathy code values, possibly lending support to the decrease of estimation accuracy by the noisy text in the ORA-D case.

DISCUSSION

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 (*Barrett-Lennard, 1981*): (i) patient expression of

Table 10 **Count of human coder disagreement on high *vs.* low empathy coding.**

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

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 Xiao et al. (2013) and Xiao et al. (2014).

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. (2012) 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 work 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 toward an *overall* assessment. In our work, 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.

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 10 we list the counts of coder disagreement.

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 9).

Table 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 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

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.

Intuition about the discriminative power of lexical cues

We analyze the discriminative power of N-grams to provide some intuition on what the model captures regarding empathy. We train $\mathtt{LM}'_H$ and $\mathtt{LM}'_L$ similarly to 'Maximum likelihood based model' on the CTT corpus. Let us denote n-gram terms as w, the log-likelihood derived from $\mathtt{LM}'_H$ and $\mathtt{LM}'_L$ as $l_n(w|H)$ and $l_n(w|L)$, respectively. Let $\mathtt{cnt}(w)$ be the count of w in the CTT corpus. We define the discriminative power δ of w as in (10).

$$\delta(w) = (l_n(w|H) - l_n(w|L)) * \mathtt{cnt}(w). \tag{10}$$

Tables 11 and 12 show the bigrams and trigrams with extreme δ values, i.e., phrases strongly indicating high/low empathy. 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.

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 1,000-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, ..., 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.

Figure 3 illustrates the results. Figure 3A shows the corresponding WER by the sampled paths from lattice $\mathcal{L}$. Figures 3B–3D show the performances by the three methods regarding ρ, σ, and Acc, 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 $1, 6, \ldots, 46$). Meanwhile, the asterisks show the performance when using the 1-best decoded paths.

In Fig. 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 tends to be more stable, while the other two methods suffer from large deviations in performance. This demonstrates the gain of robustness by re-ranking the paths 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.

CONCLUSION

In this paper we have proposed a prototype of a fully automatic system to rate therapist empathy from language use 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. Experimental 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 experimental results

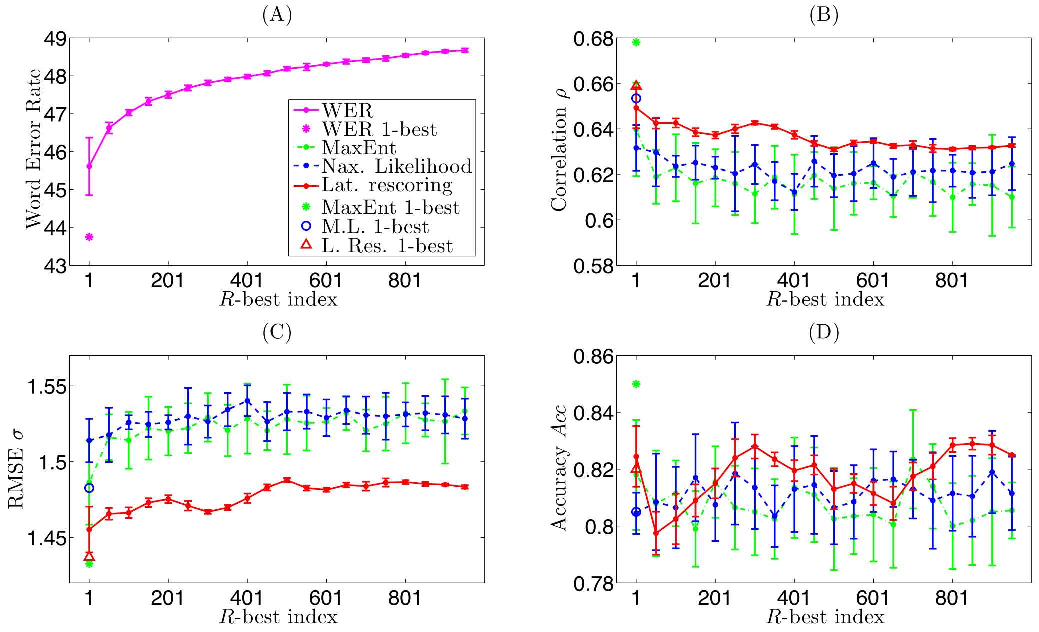

Figure 3 **Comparison of robustness by MaxEnt, Maximum Likelihood LM, and Lattice Rescoring methods.**

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, such as implementing more accurate VAD, diarization with overlapped speech detection, and a more robust ASR system. We would also like to acquire more and better training data such as by using close-talking microphones in collections. The use of close-talking microphones may fundamentally improve the accuracy of speaker diarization. As a result acoustic and prosodic cues may be integrated into the system, which relies on robust speaker identification. The system may be augmented by incorporating other behavioral modalities such 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.

APPENDIX. NOTE ON DATA SHARING

Restrictions would apply to release the data corpora we used in the experiments for two reasons. First, our work is a secondary study analyzing data of archived recordings of counseling sessions, which cannot be fully anonymized. Thus, the data cannot be released to the public. The only exception is the "General psychotherapy corpus" as a collection of psychotherapy transcripts. We obtained this data via library subscription from Alexander Street Press, Counseling and Psychotherapy Transcripts, Client Narratives, and Reference Works (http://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series). Second, all of the available original audio recordings were from third parties. The primary authors were not responsible for the collection of the original data, which was pulled from 6 different clinical trials. We list these specific studies and PI information as the following.

- Alcohol Research Collaborative: Peer Programs; *Tollison et al. (2008)*: Christine M. Lee; leecm@uw.edu
- Event Specific Prevention: Spring Break; *Lee et al. (2014)*; Christine M. Lee; leecm@uw.edu
- Event Specific Prevention: Twenty First Birthday; *Neighbors et al. (2012)*; Clayton Neighbors; cneighbors@uh.edu
- Brief Intervention for Problem Drug Use and Abuse in Primary Care; *Krupski et al. (2012)*; Peter Roy-Byrne; roybyrne@u.washington.edu
- Indicated Marijuana Prevention for Frequently Using College Students. *Lee et al. (2013)*; Christine M. Lee; leecm@uw.edu
- Context Tailored Training (CTT). *Baer et al. (2009)*; John Baer; jsbaer@uw.edu

We would like to point out that despite the constrains on data, the methods proposed in this work and the system we described are generally applicable to empathy estimation in Motivational Interviewing. We expect the results to be reproducible on audio data that are in similar nature to the data in our experiments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is supported by National Science Foundation, National Institute of Health, and Department of Defense. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Science Foundation, National Institute of Health, and Department of Defense.

Competing Interests

Shrikanth S Narayanan is an Academic Editor for PeerJ Computer Science.

Author Contributions

- Bo Xiao conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Chewei Huang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, performed the computation work, reviewed drafts of the paper.
- Zac E. Imel and David C. Atkins conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Panayiotis Georgiou and Shrikanth S. Narayanan conceived and designed the experiments, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

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).

Data Availability

The following information was supplied regarding data availability:

The raw data are recordings of addiction counseling sessions which cannot be fully anonymized. Thus, the data cannot be released to the public.

REFERENCES

Baer JS, Wells EA, Rosengren DB, Hartzler B, Beadnell B, Dunn C. 2009. Agency context and tailored training in technology transfer: a pilot evaluation of motivational interviewing training for community counselors. *Journal of Substance Abuse Treatment* **37(2)**:191 DOI 10.1016/j.jsat.2009.01.003.

Barrett-Lennard GT. 1981. The empathy cycle: refinement of a nuclear concept. *Journal of Counseling Psychology* **28(2)**:91 DOI 10.1037/0022-0167.28.2.91.

Batson CD. 2009. These things called empathy: eight related but distinct phenomena. In: *The Social Neuroscience of Empathy*. Cambridge: MIT Press, 3–15.

Berger AL, Pietra VJD, Pietra SAD. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* **22(1)**:39–71.

Black MP, Katsamanis A, Baucom BR, Lee C-C, Lammert AC, Christensen A, Georgiou PG, Narayanan SS. 2013. Toward automating a human behavioral coding system for married couples interactions using speech acoustic features. *Speech Communication* **55(1)**:1–21 DOI 10.1016/j.specom.2011.12.003.

Bone D, Lee C-C, Black MP, Williams ME, Lee S, Levitt P, Narayanan S. 2014. The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research* **57(4)**:1162–1177 DOI 10.1044/2014_JSLHR-S-13-0062.

Can D, Georgiou P, Atkins D, Narayanan SS. 2012. A case study: detecting counselor reflections in psychotherapy for addictions using linguistic features. In: *Proceedings of interspeech*, 2254–2257.

Eisenberg N, Eggum ND. 2009. Empathic responding: sympathy and personal distress. In: *The Social Neuroscience of Empathy*. Cambridge: MIT Press, 71–83.

Elliott R, Bohart AC, Watson JC, Greenberg LS. 2011. *Empathy, Psychotherapy* **48(1)**:43 DOI 10.1037/a0022187.

Georgiou P, Black M, Lammert A, Baucom B, Narayanan S. 2011. "That's aggravating, very aggravating": is it possible to classify behaviors in couple interactions using automatically derived lexical features? In: *Affective Computing and Intelligent Interaction*, 87–96.

Godfrey JJ, Holliman EC, McDaniel J. 1992. Switchboard: telephone speech corpus for research and development. In: *Proc. ICASSP*. Vol. 1. Piscataway: IEEE, 517–520.

Guha T, Yang Z, Ramakrishna A, Grossman R, Hedley D, Lee S, Narayanan S. 2015. On quantifying facial expression-related atypicality of children with autism spectrum disorder. In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Piscataway: IEEE, 803–807.

Huang CW, Xiao B, Georgiou P, Narayanan S. 2014. Unsupervised speaker diarization using riemannian manifold clustering. In: *Proceedings of interspeech*, 567–571.

Iacoboni M. 2009. Imitation, empathy, and mirror neurons. *Annual Review of Psychology* **60**:653–670 DOI 10.1146/annurev.psych.60.110707.163604.

Imel ZE, Steyvers M, Atkins DC. 2015. Computational psychotherapy research: scaling up the evaluation of patient–provider interactions. *Psychotherapy* **52(1)**:19–30 DOI 10.1037/a0036841.

Krupski A, Joesch JM, Dunn C, Donovan D, Bumgardner K, Lord SP, Ries R, Roy-Byrne P. 2012. Testing the effects of brief intervention in primary care for problem drug use in a randomized controlled trial: rationale, design, and methods. *Addiction Science & Clinical Practice* **7(1)**:27 DOI 10.1186/1940-0640-7-27.

Kumano S, Otsuka K, Matsuda M, Yamato J. 2013. Analyzing perceived empathy/antipathy based on reaction time in behavioral coordination. In: *Automatic face and gesture recognition*. Piscataway: IEEE, 1–8.

Kumano S, Otsuka K, Mikami D, Yamato J. 2011. 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*. Piscataway: IEEE, 43–50.

Lee C-C, Katsamanis A, Black MP, Baucom BR, Christensen A, Georgiou PG, Narayanan SS. 2014. Computing vocal entrainment: a signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech & Language* **28(2)**:518–539 DOI 10.1016/j.csl.2012.06.006.

Lee CM, Kilmer JR, Neighbors C, Atkins DC, Zheng C, Walker DD, Larimer ME. 2013. Indicated prevention for college student marijuana use: a randomized controlled trial. *Journal of Consulting and Clinical Psychology* **81(4)**:702 DOI 10.1037/a0033285.

Lee CM, Neighbors C, Lewis MA, Kaysen D, Mittmann A, Geisner IM, Atkins DC, Zheng C, Garberson LA, Kilmer JR, Larimer ME. 2014. Randomized controlled trial of a spring break intervention to reduce high-risk drinking. *Journal of Consulting and Clinical Psychology* **82(2)**:189 DOI 10.1037/a0035743.

Liu DC, Nocedal J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* **45(1–3)**:503–528 DOI 10.1007/BF01589116.

Metallinou A, Grossman RB, Narayanan S. 2013. Quantifying atypicality in affective facial expressions of children with autism spectrum disorders. In: *2013 IEEE international conference on multimedia and expo (ICME)*. Piscataway: IEEE, 1–6.

Miller WR, Rollnick S. 2012. *Motivational interviewing: helping people change*. New York: Guilford Press.

Miller WR, Rose GS. 2009. Toward a theory of motivational interviewing. *American Psychologist* **64(6)**:527 DOI 10.1037/a0016830.

Moyers TB, Martin T, Manuel JK, Hendrickson SM, Miller WR. 2005. Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment* **28(1)**:19–26 DOI 10.1016/j.jsat.2004.11.001.

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

Moyers TB, Miller WR. 2013. Is low therapist empathy toxic? *Psychology of Addictive Behaviors* **27(3)**:878 DOI 10.1037/a0030274.

Narayanan s, Georgiou P. 2013. Behavioral signal processing: deriving human behavioral informatics from speech and language. *Proceeding of the IEEE* **101(5)**:1203–1233 DOI 10.1109/JPROC.2012.2236291.

Neighbors C, Lee CM, Atkins DC, Lewis MA, Kaysen D, Mittmann A, Fossos N, Geisner IM, Zheng C, Larimer ME. 2012. A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations. *Journal of Consulting and Clinical Psychology* **80(5)**:850 DOI 10.1037/a0029480.

Paul DB, Baker JM. 1992. The design for the Wall Street Journal-based CSR Corpus. In: *DARPA speech and language workshop*, Stroudsburg: ACL, 357–362.

Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K. 2011. The kaldi speech recognition toolkit. In: *Proceedings of the ASRU*. Piscataway: IEEE.

Preston SD, De Waal F. 2002. Empathy: its ultimate and proximate bases. *Behavioral and Brain Sciences* **25(01)**:1–20.

Rosenfeld R. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language* **10(3)**:187–228 DOI 10.1006/csla.1996.0011.

Roy-Byrne P, Bumgardner K, Krupski A, Dunn C, Ries R, Donovan D, West II, Maynard C, Atkins DC, Graves MC, Joesch JM, Zarkin GA. 2014. Brief intervention for problem drug use in safety-net primary care settings: a randomized clinical trial. *Jama* **312(5)**:492–501 DOI 10.1001/jama.2014.7860.

Stolcke A. 2002. Srilm—an extensible language modeling toolkit. In: *Proceedings of interspeech*, 901–904.

Substance Abuse and Mental Health Services Administration. 2013. Results from the 2012 national survey on drug use and health: summary of national findings. In: *NSDUH Series H-46, HHS Publication No. (SMA) 13-4795*. Rockville: Substance Abuse and Mental Health Services Administration.

Tollison SJ, Lee CM, Neighbors C, Neil TA, Olson ND, Larimer ME. 2008. Questions and reflections: the use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior Therapy* **39(2)**:183–194 DOI 10.1016/j.beth.2007.07.001.

Van Segbroeck M, Tsiartas A, Narayanan SS. 2013. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice. In: *Proceedings of interspeech*, 704–708.

Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schröder M. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* **3(1)**:69–87 DOI 10.1109/T-AFFC.2011.27.

Wang W, Lu P, Yan Y. 2008. An improved hierarchical speaker clustering. *ACTA ACUSTICA* **33(1)**:9.

Xiao B, Bone D, Van Segbroeck M, Imel ZE, Atkins D, Georgiou P, Narayanan S. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In: *Proceedings of interspeech*, 213–217.

Xiao B, Can D, Georgiou PG, Atkins DC, Narayanan SS. 2012. Analyzing the language of therapist empathy in Motivational Interview based psychotherapy. In: *Signal & information processing association annual summit and conference (APSIPA ASC)*, 1–4.

Xiao B, Georgiou PG, Baucom BR, Narayanan SS. 2014. Power-spectral analysis of head motion signal for behavioral modeling in human interaction. In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Piscataway: IEEE, 4593–4597.

Xiao B, Georgiou PG, Imel ZE, Atkins DC, Narayanan SS. 2013. Modeling therapist empathy and vocal entrainment in drug addiction counseling. In: *Proceedings of interspeech*, 2861–2865.

Xiao B, Imel ZE, Georgiou P, Atkins DC, Narayanan SS. 2015. "Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE* **10(12)**:1–15 DOI 10.1371/journal.pone.0143055.

Zhang L. 2013. Maximum entropy modeling toolkit for Python and C++. *Available at https://github.com/lzhang10/maxent*.