# Assessing Empathy using Static and Dynamic Behavior Models based on Therapist's Language in Addiction Counseling

*Sandeep Nallan Chakravarthula*[1], *Bo Xiao*[1],
*Zac E. Imel*[2], *David C. Atkins*[3], *Panayiotis Georgiou*[1]

[1]Dept. Electrical Engineering, University of Southern California, U.S.A.
[2]Dept. Educational Psychology, University of Utah, U.S.A.
[3]Dept. Psychiatry & Behavioral Sciences, University of Washington, U.S.A.
nallanch,boxiao@usc.edu, zac.imel@utah.edu, datkins@u.washington.edu,
georgiou@sipi.usc.edu

## Abstract

Empathy by the counselor is an important measure of treatment quality in psychotherapy. It is a behavioral process that involves understanding and sharing the experiences and emotions of a person over the course of an interaction. While a complex phenomenon, human behavior can at moments be perceived as strongly empathetic or non-empathetic. Currently, manual coding of behavior and behavioral signal processing models of empathy often pose the unnatural assumption that empathy is constant throughout the interaction. In this work we investigate two models: Static Behavior Model (SBM) that assumes a fixed degree of empathy throughout an interaction; and a context-dependent Dynamic Behavior Model (DBM), which assumes a Hidden Markov Model, allowing transitions between high- and low- empathy states. Through the non-causal human perception mechanisms, these states can be perceived and integrated as high- or low- gestalt empathy. We show that the DBM performs better than the SBM, while as a byproduct, generating local labels that may be of use to domain experts. We also demonstrate the robustness of both SBM and DBM to transcription errors stemming from ASR rather than human transcriptions. Our results suggest that empathy manifests itself in different forms over time and is best captured by context-dependent models.

**Index Terms**: Dynamic Behavior, Empathy, ASR, Language Model, Maximum Likelihood, Hidden Markov Model

## 1. Introduction

Human behavior expression is extremely complex and conveys a multitude of information, including information related to one's mental health. Psychologists have used observational methods to analyze human behavior and reach conclusions regarding the efficacy of various treatments for patients. Interpersonal interactions are complex and dynamic processes that contain significant heterogeneity and variability with respect to context, speaker traits and identity, *etc.*. Attaining complete insight into human interaction mechanisms is therefore challenging.

Engineering approaches offer a viable way to study human interactions. The emerging field of Behavioral Signal Processing [1, 2] offers encouraging results towards using computational tools and models of human interactions to inform research and practice across a variety of behavior-centered domains. The approach relies on integrating domain knowledge and engineering; *e.g.,* feature design and machine learning methods are guided by domain knowledge, and experimental results in turn validate the effectiveness of these multimodal features and algorithms on real datasets. Our work in this paper is a continuation of previous BSP work [3, 4] that has focused on empathy in patient-therapist interactions during counseling for drug addiction.

The presence of empathy generally involves taking the perspective of others, and responding with sensitivity and care appropriate to the suffering of another [5]. High empathy ratings are associated with positive outcomes in a variety of human interactions [6, 7]. Empathy is considered an essential quality of therapists in psychotherapy and drug abuse counseling in particular and is associated with positive clinical outcomes such as decreased substance use [8, 9]. In psychotherapy studies, therapists are often rated on their empathy levels by third-party observers (coders), based on multimodal behavioral cues expressed throughout the interaction.

Previous work on inferring empathy has used acoustic cues [10, 11] as well as lexical information stream using maximum likelihood models [4], while other work has focused on simulating and synthesizing empathy through computer agents [12, 13]. Xiao *et al.*[3] found that vocal similarity between client and therapist was significantly correlated with empathy and that it contributed to better classification when combined with speaking time features.

## 2. Overview of Proposed Work

Existing methods of inferring session-level behavioral descriptors, such as empathy, have an inherent drawback; they assume that a person, perceived by annotators to behave in a certain way during an interaction, occupies only that particular behavioral state for the entire duration. This effectively treats an interlocutor as a single-state generative model, referred to here as *Static Behavior Model* (SBM, Sec. 4.1), which is simplistic and does not account for the human tendency to dynamically react and adapt to stimuli. In this work, we present a more realistic model of human interaction that allows for the human subject to move through a range of behavioral states, that we call a *Dynamic Behavior Model* (DBM, Sec. 4.2). We provide methods of training these models using session labels in Sec. 5.

Further to exploring dynamic models that allow for behavioral transitions throughout the interaction we are also interested in understanding how humans integrate local information towards a gestalt rating. This is a highly non-linear process and is studied extensively in the psychology literature; *e.g.,* [14] investigate recency (thing observed last counts more) and primacy (thing observed first counts more).

Work supported by NSF, NIH and DoD

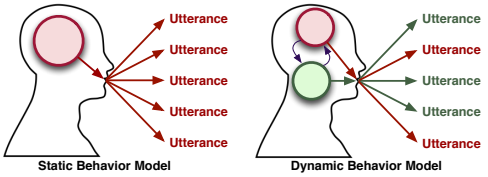


Figure 1: Conceptualizing the proposed graphical model on the right versus the baseline on the left

Integrating information requires an in-depth investigation of the algorithmic metrics of behavior (say log likelihood of an empathetic statement) versus the impact this has on the coding process. In our recent work we briefly touched on this [15] in investigating how human annotators employ isolated saliency or causal integration to make their decisions. In this work we will employ our DBM to investigate whether equal weight should be given to every turn (Sec. 4.2.1) or whether likelihood metrics should be employed (Sec. 4.2.2).

We use the dataset described in Sec 3. We evaluate our models in Sec. 6 on three different transcripts of different transcription accuracy.

## 3. Dataset

In this work, we employ the CTT (Context Tailored Training) data set collected in a therapist training study [16]. The study's focus was on Motivational Interviewing [17], which is a type of addiction counseling that helps people resolve ambivalence and emphasizes the intrinsic motivation of changing addictive behaviors. Each session in the set received an *overall* empathy code in the range of 1 to 7 given by human-experts according to a specific coding manual named "Motivational Interviewing Treatment Integrity" [18]. Intra-Class Correlations (ICC) for inter-coder and intra-coder comparisons were 0.67±0.16 and 0.79±0.13 respectively, which exhibit high coder reliability in the annotation.

We binarize the set of 200 sessions, based on the average ratings, into two classes: $C_0$ (1 to 4) representing low-empathy and $C_1$ (4.5 to 7) representing high-empathy. In total there are 133 unique therapists in the set. The audio recording of each session is about 20 min long with single channel, far-field microphone. These sessions were also manually transcribed with speaker labels and timing information (*i.e.,* manual diarization).

Moreover, we trained a large vocabulary Automatic Speech Recognizer (ASR) on additional in-domain data using the Deep Neural Network model implemented in the Kaldi library [19]. We applied the ASR to the CTT set in two settings: with manual and automatic diarization. For the latter, we conducted Voice Activity Detection and Diarization on the audio signal before decoding, plus speaker role identification on the decoded transcripts. In average, we obtained Word Error Rate of 43.1% and 44.6% for the manual and automatic diarization cases, respectively. We, therefore, have three versions of the dataset to test our models on: (1) Manual Diarization Manual Transcription (MDMT), (2) Manual Diarization Automatic Transcription (MDAT) and (3) Automatic Diarization Automatic Transcription (ADAT).

## 4. Behavioral Modeling

Georgiou *et al.* [20] have analyzed human behavior in a couple therapy setting by assuming a constant behavioral state of expression throughout the interaction. This is equivalent to modeling each interlocutor in the interaction as a single state generative model as shown in Fig. 1 (left). This is clearly limiting and does not reflect human behavior, that dynamically adapts based on the various stimuli, internal and external.

We expanded that work in [21] by introducing dynamic behavior models for the case of negativity in couples therapy. In this paper we further expand this work in the domain of addiction and motivational interviewing and evaluate noisy transcripts using the *Dynamic Behavior Model* (DBM) described below. DBM, in contrast to the SBM, allows for transitions between different states and makes turn-level decisions about empathy instead of session-level ones.

### 4.1. Static Behavior Model (SBM)

The *Static Behavior Model* (SBM) is in effect behavioral averaging as in [20]. Thus, all the utterances observed in that session are generated from the same behavioral state as in Fig. 1(left). Predicting the interlocutor's behavioral state becomes equivalent to identifying the class label $C_i = \{C_0, C_1\}$. We use a *Maximum-A-Posteriori* scheme where, for a set of *m* observed utterances from the entire interaction, $\bar{U}$=$\{U(1), \ldots, U(M)\}$, we want to find:

$$P(\text{Low-Empathy or High-Empathy}|\bar{U}) = P(C_0 \text{ or } C_1|\bar{U})$$

$$C_i = \arg\max_{C_j} P(\bar{U}|C_j)P(C_j) \quad (1)$$

### 4.2. Dynamic Behavior Model (DBM)

In the DBMs, the behavior of the interlocutor is modeled as changing across talk turns, but remaining constant within a turn. This turn-level variation across time is realized in the form of transitions between behavioral states. Fig. 1 (right) illustrates the DBM, where different utterances are generated from different states within the same session of interaction.

Since each utterance can now be generated by one of two possible states $S_i = \{S_0, S_1\}$, there is no one-to-one correspondence between the states and classes $C_0/C_1$ anymore. The resulting formulation is similar to (1) and estimates the turn-level state probabilities as:

$$P(S_i|U(m)) \propto P(U(m)|S_i)P(S_i) \quad (2)$$

Human perception is capable of integrating local events to arrive at an overall global-level impression, but this process is complex and not transparent. In our work, we have used perceptual methods that use local information to derive the same global decisions. For example, in [15] we evaluated whether global behavior could be effectively judged based on a locally isolated, yet highly informative event or by integrating information over time. Similarly, in this work, the DBM employs two information-integration techniques, one that assumes that each talk-turn conveys exactly the same amount of information and should be counted as such, and one that employs probabilistic measures for accumulating behavioral beliefs to reach global decisions.

#### 4.2.1. Activation-based

The Activation-based DBM (ADBM), shown in Fig. 2 (left), decides the global behavior using a majority-vote principle, thus assuming that all talk turns carry the same perceptual weight. The turn-level decision for utterance $U(m)$ about the behavioral state $S_i$ is:

$$S_i = \arg\max_{S_j} P(U(m)|S_j)P(S_j) \quad (3)$$

For the session-level decision, we require a mapping from the dominant behavioral state to the behavioral class, $S_i \rightarrow C_j, \forall\ i, j \in \{0, 1\}$. This mapping is learned at the training stage, as explained in Sec. 5.2.1.

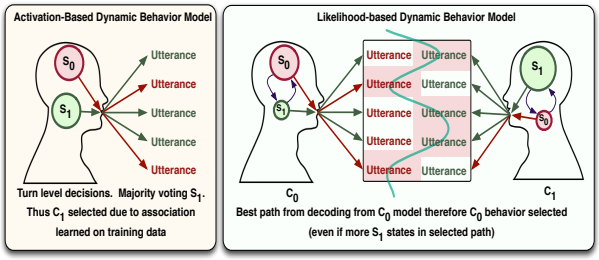


Figure 2: Activation-based DBM (left) versus the Likelihood-based DBM (right)

#### 4.2.2. Likelihood-based

The Likelihood-based DBM (LDBM), shown in Fig. 2 (right), uses a decision scheme based on Hidden Markov Modeling of the behavioral classes $C_j$. The states $S_0$ and $S_1$ that generate the utterances are now assumed to be hidden, and the HMM of each class, HMM($C_j$), is allowed to generate from both states. Thus, the underlying states must be deduced from the model that provides the best decoded sequence. This modeling is meant to reflect the real-world observation that the same utterance can be indicative of different types of behavior, depending on the context. Thus, the decision rule is:

$$C_i = \arg\max_{\lambda^j} P(\overline{S}_j|\overline{U}, \lambda^j) \quad (4)$$

Where $\lambda^j$ is the set of HMM parameters of class $C_j$, $\overline{U}$=$\{U(1), ..., U(M)\}$ is the set of utterances in that session, $\overline{S}$=$\{S(1), ..., S(M)\}$ is state sequence decoded by HMM($C_j$).

## 5. Training

This section contains the description for the training of the SBM and the two DBMs. In order to avoid data contamination between training and testing, we employ a leave-one-therapist-out cross-validation scheme. For practical implementation purposes, we represent probabilistic lexical structures with language models built using the SRILM toolkit [22]. We also use normalized perplexity as our probability measure. Since minimizing perplexity is equivalent to maximizing probability, we use a minimum perplexity scheme in place of maximum likelihood where valid.

### 5.1. Static Behavior Model

In the SBM, language models $L_0$/$L_1$ are trained for classes $C_0$/$C_1$ on the utterances from corresponding train-fold sessions respectively. We then smooth the models with a *Universal Background Model* (UBM) using an interpolation weight parameter $\lambda$=0.1. At the end of training, our SBM consists of language models for each interlocutor corresponding to low- and high-empathy behavior.

### 5.2. Dynamic Behavior Model

The DBM assumes that an interaction consists of multiple behavioral instantiations, even if the entire session is tied to one behavioral type. Since we only have session-level labels $C_0$/$C_1$, the utterance-level labels $S_0$/$S_1$ are latent. They are iteratively estimated through semi-supervised learning methods. The learning convergence is verified indirectly through the total training perplexity, since local label information is not present. The ADBM and LDBM tie utterance-level labels to the session-level behavior in different perceptual ways, as explained below:

#### 5.2.1. Activation-based DBM

In the ADBM, language models $L_0$/$L_1$ are built to represent behavioral states $S_0$/$S_1$, as opposed to classes $C_0$/$C_1$ in the SBM. However, they are initialized exactly in the same way as that of the SBM, but are then re-trained until the training perplexity converges. Each utterance is classified independently of the rest using an Expectation Maximization-like (EM) learning scheme as described in Algorithm 1.

**Algorithm 1** EM algorithm for state convergence in activation-based DBM

Initialize utterances in $C_0$ session $\in \{S_0\}$, $C_1 \in \{S_1\}$
Build language model $L_0$ from utterances $\in S_0$,
$L_1$ from utterances $\in S_1$
**while** training perplexity does not converge **do**
*E-step*: Classify every utterance $U(m)$ in $C_0$,$C_1$ classes
Get perplexities $PP_0$, $PP_1$ of $U(m)$ computed by $L_0$,$L_1$
$PP_0\{U(m)\} \underset{\text{state}=S_0}{\overset{\text{state}=S_1}{\gtrless}} PP_1\{U(m)\}; m \in \{0, 1, ..., M\}$
*M-step*: Build $L_0$ from $S_0$ utterances,& $L_1$ from $S_1$
**end while**

**Algorithm 2** Viterbi-EM algorithm for state and class parameter convergence in likelihood-based DBM

Initialize utterances in $C_0$ session $\in \{S_0\}$, $C_1 \in \{S_1\}$
Build language model $L_0$ from utterances $\in S_0$,
$L_1$ from utterances $\in S_1$
Initialize $\pi$,$\alpha^0$,$\alpha^1$
**while** training perplexity does not converge **do**
*E-step*: Decode $C_0$ utterances using $\alpha^0, L_0, L_1, \pi$
**for** every session utterance $U(m)$ **do**
Get probabilities $P_0$,$P_1$ of utterance $U(m)$ from $L_0$, $L_1$
Find probability that $U(m)$ was generated by state $S_k$
**if** m=1 (start of session) **then**
$\gamma_k(m)$ = $\pi_k$*$P_k\{U(m)\}$; k $\in \{0, 1\}$
**else**
$\gamma_k(m)$ = $\arg\max_j[\gamma_j$(m-1) * $\alpha^0$(j,k)] * $P_k\{U(m)\}$;
j,k $\in \{0, 1\}$
$\gamma_0$(m-1) * $\alpha^0$(0,k) $\underset{\theta_m(k)=S_1}{\overset{\theta_m(k)=S_0}{\gtrless}}$ $\gamma_1$(m-1) * $\alpha^0$(1,k);
k $\in \{0, 1\}$
**end if**
**end for**
Decode state sequence using $\gamma_k(M)$ and $\theta_m(k)$;
m $\in \{M, M-1, ...2\}$, k $\in \{0, 1\}$
Repeat E-step for $C_1$ utterances; replace $\boldsymbol{\alpha^0}$ with $\boldsymbol{\alpha^1}$
*M-step*: Re-estimate states, class parameters
Build $L_k$ from all $U(m)$ whose $state = S_k$; $k \in \{0, 1\}$
Update $\alpha^n(i,j) = \frac{\text{count}(<i,j> \text{ state pairs in class } C_n)}{\sum_k \text{count}(<i,k> \text{ state pairs in class } C_n)}$;
i,j,k $\in \{0, 1\}$, $n \in \{0, 1\}$
**end while**

After convergence, each session comprises both states. From the association of session with a class based on human coding ($C_0$=low- or $C_1$=high-empathy), and session with states ($S_0$ or $S_1$) based on the EM above, we can learn the mapping of states to classes. We do that by computing the proportions of state occupancies in each class and associating it with the dominant one. This, based on our initialization, usually results in an association of $S_0$ with $C_0$ and of $S_1$ with $C_1$.

#### 5.2.2. Likelihood-based DBM

In the LDBM, each behavioral class is represented by a HMM that describes the characteristics of transitions between different behavioral states. We initialize the transition matrix of $C_i$ to heavily favor $S_j \rightarrow S_i, \forall\ j$ transitions over the rest. We then use the Viterbi-EM algorithm to obtain converged estimates of model parameters and behavioral states, as described in Algorithm 2. At the end of training, the LDBM is associated

with HMMs that correspond to human-coded low- and high-empathy behaviors (HMM($C_i$)).

Each behavioral state $S_i$ is associated with a language model, while each class $C_i$ consists of a common initial-state probability vector $\boldsymbol{\pi}$ and a matrix $\boldsymbol{\alpha}^i$ that governs state transitions in that class. For example, $\boldsymbol{\alpha}^0(i,j)$ represents the probability of a $C_0$-rated therapist transitioning to state $j$, given that he/she was previously in state $i$.

# 6. Evaluation of Behavioral Models

Evaluation of the behavioral models is performed on the heldout set, with the methodology similar to that of the training. The evaluation results for all 3 models for all 3 transcription schemes are shown in Table 1.

### 6.1. Static Behavior Model

We use the overall perplexity of all utterances to estimate the behavioral class, as shown in (5).

$$C_i = \arg\min_j \sum_m PP_j\{U(m)\} \tag{5}$$

Where $PP_i\{\}$ represents perplexity score of utterance computed by language model $L_\text{i}$

### 6.2. Dynamic Behavior Model

At testing time, both ADBM and LDBM estimate the global session-level behavior based on the same schemes as those for training, as explained below:

#### 6.2.1. Activation-based DBM

Given a therapist's test session, state labels are assigned to each utterance $U(m)$ independently, based on the LM perplexities from $L_0$ and $L_1$. The most dominant state $S_k$ is identified and we choose the behavioral class that maximizes (6).

$$C_i = \arg\max_{C_j} P(S_k|C_j) \tag{6}$$

#### 6.2.2. Likelihood-based DBM

We decode the set of test utterances $\overline{U}$ using the HMMs of $C_0$/$C_1$, thereby obtaining the most likely state sequences $\overline{S}_0/\overline{S}_1$ and their corresponding likelihoods. The LDBM then picks the class with the highest likelihood as the global behavior label for the test session, as shown in (7)

$$C_i = \arg\max_{\lambda_j} P(\overline{S}_j|\overline{U},\lambda_j) \tag{7}$$

Where $\overline{U}$ is set of test utterances, $\{U(1),...,U(M)\}$
$\overline{S}_j$ is the most likely state sequence predicted by HMM($C_j$)
$\lambda_j$ is set of HMM parameters of class $C_j$, $\{\boldsymbol{\alpha}^\text{j}$,$L_0$,$L_1$,$\boldsymbol{\pi}\}$

# 7. Results and Discussion

Table 1 displays the classification accuracies of evaluation on all 3 behavioral models for each noisy version of our dataset.

As expected, the highest classification accuracy for all the models is obtained in the case of manual diarization and manual transcription. The Likelihood-DBM performs the best among all 3 models, with its highest classification accuracy close to 87%. Looking at the unigram column the ADBM performs worse than the LDBM, but better than the SBM. This is expected since the DBM can better capture behavior than the SBM and the LDBM weighs data according to their importance rather than equally weighting them as in the ADBM.

The results from bigram and trigram models demonstrate a drop in performance that can be attributed to data sparsity. This performance drop is visible in all models, while we would expect the context captured by the higher-order n-gram features to contribute to better performance if data sparsity issues were not prevalent.

| Dataset Type | Model | 1-gram Accuracy% | 2-gram Accuracy% | 3-gram Accuracy% |
|---|---|---|---|---|
| | Chance | 60.0 | 60.0 | 60.0 |
| MDMT | SBM | 79.0 | 78.0 | 75.5 |
| | ADBM | 81.5 | **82.5** | 80.0 |
| | LDBM | **86.5** | 80.5 | **81.0** |
| MDAT | SBM | 79 | 70.5 | 71 |
| | ADBM | 80.0 | **77.5** | **75.0** |
| | LDBM | **82** | 74.5 | 72.0 |
| ADAT | SBM | 73.5 | 66.5 | **68.5** |
| | ADBM | 74.0 | **69.0** | 68 |
| | LDBM | **78.5** | **69.0** | **68.5** |

Table 1: Classification Accuracy of Behavioral Models for different versions of the dataset

We see that, for the SBM and the ADBM, errors in diarization bring about a more significant drop in performance than errors in transcription. Errors in diarization result in parts of the interaction being attributed to the wrong speaker and hence create local, yet pronounced errors. In contrast, errors in transcription are distributed throughout the interaction. The SBM averages the whole interaction so it can handle errors better. The LDBM, due to its decoding, is more influenced by errors as they can propagate through the decoded sequence. Thus we observe that the LDBM is affected more by signal noise. Nevertheless, even with the high word error rates of such signals we see that LDBM still outperforms all other algorithms and does significantly better than chance.

# 8. Conclusions and Future Work

This work dealt with predicting therapist empathy in psychotherapy-based counseling. We proposed a Dynamic Behavior Model based scheme and contrasted it against the Static Behavioral Model. Our proposed work models a therapist as transitioning through multiple behavioral states, in order to obtain an overall perception of empathy. We tested two models for dynamically modeling the behavior: Activation-based DBM and Likelihood-based DBM which outperformed the SBM. Furthermore, the LDBM performed better than ADBM as it takes context into consideration while being a more realistic model of how human behavior evolves: an interlocutor is likely to change their behavior, but their past behavior is likely to play a factor on what their next behavior will be. We also tested the robustness of the 3 models to errors in diarization as well as transcription and found that the LDBM outperforms the other two models.

In this paper, we described only two methods of processing local information in order to derive a global behavioral description. We plan to investigate alternative information integration methods. We also intend to extend this model to further incorporate turn-level interaction between therapist and patient to model the behavioral influences of one over the other. In addition, we will explore using finer behavior stratification in the form of more behavioral states and observe its effect on the models' performance. Finally we want to investigate multimodal systems, starting from the integration of acoustic and lexical modalities.

## 9. References


[1] P. G. Georgiou, M. P. Black, and S. S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, ser. J-HGBU '11. New York, NY: ACM, 2011, pp. 7–12. [Online]. Available: http://doi.acm.org/10.1145/2072572.2072576

[2] S. S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceeding of the IEEE*, 2014.

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

[4] B. Xiao, P. G. Georgiou, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Proceedings of APSIPA*, 2012.

[5] C. D. Batson, "These things called empathy: eight related but distinct phenomena." 2009.

[6] N. D. Feshbach, "12 parental empathy and child adjustment/maladjustment," *Empathy and its development*, p. 271, 1990.

[7] P. S. Bellet and M. J. Maloney, "The importance of empathy as an interviewing skill in medicine," *JAMA*, vol. 266, no. 13, pp. 1831–1832, 1991.

[8] L. S. Greenberg, J. C. Watson, R. Elliot, and A. C. Bohart, "Empathy." *Psychotherapy: Theory, Research, Practice, Training*, vol. 38, no. 4, p. 380, 2001.

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

[10] 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 *In Proceedings of Interspeech*, 2014.

[11] 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 (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[12] 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.

[13] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, "A computational model of empathy: Empirical evaluation," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 1–6.

[14] D. D. Steiner and J. S. Rain, "Immediate and delayed primacy and recency effects in performance evaluation." *Journal of Applied Psychology*, vol. 74, no. 1, p. 136, 1989.

[15] C.-C. Lee, A. Katsamanis, P. G. Georgiou, and S. S. Narayanan, "Based on isolated saliency or causal integration? toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio test," in *Proceedings of InterSpeech*, Sep. 2012.

[16] 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.

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

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

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[20] P. G. Georgiou, M. P. Black, A. Lammert, B. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proceedings of Affective Computing and Intelligent Interaction*, Memphis, TN, USA, 2011.

[21] S. Nallan Chakravarthula, R. Gupta, B. Baucom, and P. G. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Acoustics, Speech, and Signal Processing, 2015. Proceedings.(ICASSP'15). 2015 IEEE International Conference on*. IEEE, 2015.

[22] A. Stolcke *et al.*, "SRILM-an extensible language modeling toolkit." in *INTERSPEECH*, 2002.