Analyzing Speech Rate Entrainment and Its Relation to Therapist Empathy in Drug Addiction Counseling

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

[1]SAIL, 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.
[1]http://sail.usc.edu [2]zac.imel@utah.edu [3]datkins@u.washington.edu

Abstract

A key quality index in drug addiction counseling such as Motivational Interviewing is the degree of therapist's empathy towards the client. Empathy ratings are meant to evaluate the therapist's understanding of the patient's feelings, through their sensitivity and care of response. Empathy is also associated with the manifestation of behavioral entrainment in the interaction. In this paper, we compute a measure of entrainment in speech rate during dyadic interactions, and investigate its relation to perceived empathy. We show that the averaged absolute difference of turn-level speech rates between the therapist and the patient correlates with the ratings of therapist empathy. We also present the correlation of empathy to the statistics of speech and silence durations. Finally we show that in the task of automatically predicting high or low empathy, speech rate cues provide complementary information to previously proposed prosodic cues. These findings suggest speech rate as an important behavioral cue that is modulated by entrainment and contributes to empathy modeling.

Index Terms: Speech rate; Empathy; Entrainment; Speech-text alignment; Behavioral Signal Processing

1. Introduction

Empathy generally encompasses two aspects: (1) one's internalization of another's thoughts and feelings (taking the perspective of others), and (2) one's response with the sensitivity and care appropriate to the suffering of another (feeling for the other) [1]. Extensive multidisciplinary studies on empathy have established it as a core factor in human interaction [2–5].

In particular, the empathy level expressed by the therapist is an essential quality index in psychotherapy, including in drug addiction counseling. Clinical studies show that higher ratings of therapist empathy are associated with treatment retention and positive outcomes [6–8]. Such ratings are usually assigned by trained human coders. Human perception of empathy is often multimodal, where information in various communication channels is integrated in an implicit manner towards an overall judgment [9]. This paper considers the computational modeling of empathy, especially in illuminating specific underlying behavioral cues of its expression. Our overarching goal is to provide a computational ancillary of empathy to support human expert analysis and decision making [10].

In previous work, we have modeled empathy through lexical cues based on empathy-specific language models [11], vocal entrainment cues based on measures of vocal similarity between the therapist and patient [12], and prosodic cues based on the joint distributions of a group of quantized prosody features [13]. We showed positive correlations between these behavioral signal cues and expert annotated empathy codes, as well as successful prediction of high *vs.* low empathy codes using these cues. Moreover, Kumano *et al.* have studied the perceived empathy level in group conversations using Bayesian probabilistic models and a variety of behavioral cues, including facial expression, gaze, speech activity, head gestures, and response timing information [14, 15].

In this work, we follow the track of analyzing the connection between entrainment and empathy [12], by extending the dyadic patterning in speech rates. Entrainment refers to the phenomenon that the behaviors of the interlocutors becoming more similar during the interaction, possibly in multiple communication channels or biometrical states [16]. In the literature, theoretical relations between entrainment and empathy have been extensively studied [2, 17–19]. Some computational models of entrainment have also been reported, *e.g.,* Lee *et al.* have modeled the vocal entrainment of couples in conversations and its relation to the couples' affective behavioral characteristics [20]. Delaherche *et al.* have surveyed the emerging methods for capturing multimodal entrainment from behavior signals, and summarized them into three types: correlation based, phase and spectrum comparison, and bags-of-instances comparison [21].

Speech rate, *i.e.,* the number of words, syllables, or phonemes a subject utters in a unit of time, reflects many internal states of the subject. Entrainment in speech rate has been reported. Guitar *et al.* have shown that children slow their speech rate when the mothers speak slower [22]. Manson *et al.* have shown that the degree of speech rate entrainment may predict the outcome of a collaborative task by two interlocutors [23]. However, little work has focused on computational models of the link between speech rate entrainment and empathy, which is the aim of the current study.

In this paper, we first introduce the data sets used for the study in Sec. 2. We show a computational means for examining speech rate entrainment in Sec. 3. In Sec. 4 we investigate how the dynamics of speech rate entrainment are related to therapist empathy. In Sec. 5 we study the relation between speech/silence durations and empathy. We examine the performance of classifying perceived high *vs.* low empathy using the proposed rate cues in Sec. 6. We discuss the robustness of the cues in Sec. 7, and conclude the study with future directions in Sec. 8.

2. Dataset and speech alignment

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

2.1. Switchboard corpus

Switchboard [24] is a large collection of two-sided telephone conversation from the United States. A robot operates the con-

This work is supported by NSF, NIH and DoD.

nection between the interlocutors and introduces a topic to discuss. It also ensures no two speakers would converse together more than once.

In our analysis we employ 2438 sessions from the corpus. We use the Automatic Speech Recognizer (ASR) generated, and manually corrected word level alignment [1] of speech and transcript to compute speech rates for each session and speaker.

2.2. Motivational interviewing data

Motivational Interviewing (MI) is a type of addiction counseling, which helps people to resolve ambivalence and emphasizes the intrinsic motivation of changing addictive behaviors [25]. The effectiveness of MI has been shown in various clinical trials; and theories about its mechanisms have been developed [7]. In this work we use the recordings of MI sessions from two data corpora, the TOPICS set and the CTT set.

The TOPICS set contains 899 MI sessions from five different psychotherapy studies [26–30], including intervention of college student drinking and marijuana use, as well as mental health care for drug use in clinics. Audio data are available as single-channel far-field recordings. Session length is from 20 min to 1 hour. Due to resource constraints, we randomly selected 153 sessions and transcribed them with annotations of speaker and start/end time of each turn.

The CTT (Context Tailored Training) set contains 200 sessions chosen from 826 MI sessions in a therapist training study [31]. The recording and transcription schemes are the same as the TOPICS set. Each session is about 20 min.

The CTT corpus was annotated for therapist's *overall* empathy code. Three coders reviewed the 826 audio recordings, and assigned discrete code values from 1 to 7 for each session, following a specially designed coding system — the "Motivational Interviewing Treatment Integrity" (MITI) manual [32]. Intra-Class Correlations (ICC) of 0.67±0.16 for inter-coder and 0.79±0.13 for intra-coder prove coder reliability in the annotation. We use the mean value of empathy codes if the session is coded twice. No session was coded more than twice.

The selected 200 sessions are 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 of 2.16±0.55; while that for the high empathy class is from 4.5 to 7, with mean of 5.90±0.58. The set contains 133 unique therapists and no one has more than three sessions.

2.3. Automatic speech-text alignment for MI data

The available manual segmentation only marks speaking turns; for more precise timing between and within turns, we adopt an approach of force-aligning speech to transcripts based on ASR. We implemented a large vocabulary, continuous speech recognizer using the Kaldi toolkit [33]. In order to match the spontaneous, informal speaking style in MI, we constructed an Acoustic Model (AM) that is adapted to the data conditions. We employed the TOPICS set (total about 104h) to train the AM, using standard MFCC + Δ + $\Delta\Delta$ features, feature Maximum Likelihood Linear Regression (fMLLR), Speaker Adaptive Training (SAT), and Deep Neural Network (DNN) in the final stage. For the lexicon, we combined words in the Switchboard and WSJ dictionaries and manually added high frequency, domain-specific words and fillers, *e.g., vicodin* in the drug topic and *mm* as a filler word.

We employ the Viterbi algorithm for phoneme level forced-alignment which we transform into word level alignment for further analysis. To evaluate the AM which determines the alignment quality, we take an indirect approach by decoding the CTT set. We train the Language Model (LM) on the TOPICS set and mix it with a large background LM, then apply the ASR on the CTT set. We obtained an average Word Error Rate of 43.1%, which is in an acceptable range for the case of spontaneous conversation and challenging acoustic conditions. We provide further discussion about alignment reliability in Sec. 7.

[1] http://www.isip.piconepress.com/projects/switchboard/

3. Matching of average speech rate

We first investigate the proposed computational measure for entrainment in session-level, average speech rates of the interlocutors in the Switchboard corpus. We employ the Switchboard corpus since it is a standard database that contains a large number of interactions, therefore strengthens the statistical power of our hypothesis tests in addition to that obtained on the MI data. We define the average word rate R_w as in (1), where N is the total count of words (w_i) by a subject in the conversation. t_{begin} and t_{end} are the beginning and ending time of a word [2]. Similarly, we obtain the average syllable rate R_s and phoneme rate R_p in (2), (3). Note that we exclude partial words and nonverbal units such as hesitations and laughters.

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

$$R_s = \frac{\sum_{i}^{N} syllable\_cnt(w_i)}{\sum_{i=1}^{N} \left(t_{\text{end}}(w_i) - t_{\text{begin}}(w_i) \right)} \quad (2)$$

$$R_p = \frac{\sum_{i}^{N} phoneme\_cnt(w_i)}{\sum_{i=1}^{N} \left(t_{\text{end}}(w_i) - t_{\text{begin}}(w_i) \right)} \quad (3)$$

We hypothesize that if entrainment exists in interlocutor speech rates, they should correlate higher for pairs of true interlocutors than any randomly shuffled pairing of speakers. Such a benchmarking approach is standard in dyadic analyses [21].

Firstly, in Fig. 1 we show the distribution (the darker the higher density) of R_w by all speakers, where we see a clear trend of matching between pairs of interlocutors (labeled as speaker *A* and *B* in each pair). We compute the correlation of R_w (and R_s, R_p) over conversing speaker pairs to capture this trend of matching speech rates. In Table 1 we show the results. Due to the large number of samples (2438 sessions), these correlations are significant ($p < 10^{-19}$ in *t*-test) though the values are small. The correlations do not rely on the order of speaker labels *A* or *B*; the variance of the correlations obtained with random speaker labels is below 10^{-3}.

Meanwhile, we compute the correlation of the average speech rates between "randomly paired" *pseudo*-interlocutors that are not drawn from the same interaction. We repeat this process 1000 times. In Table 2 we report the mean value, most significant p-value, and maximum absolute value of the above correlations. We see that the lowest p-values under random pairings are dramatically larger than those in the cases of true interactions. The mean values are close to zero, suggesting there is no correlation under random conditions. These results lend further support to the existence of entrainment in speech rates during interactions.

Table 1: *Correlations of average speech rates by pairs of interlocutors, and the significance in t-test*

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

[2] We eliminate silence time to avoid the influence of line delay and interruption in phone conversation.

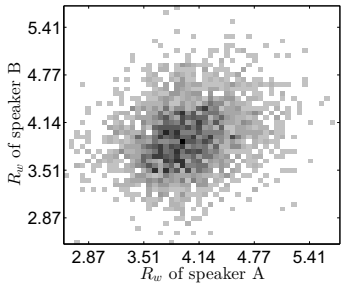

Figure 1: *Distribution of R_w by pairs of interlocutors*

Table 2: *Statistics of correlations of average speech rates by randomly shuffled pairs of pseudo-interlocutors*

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

Similarly, we conduct the analysis on the combination of the TOPICS and CTT sets using forced-alignment based speech rates. We exclude nonverbal and out-of-vocabulary words in computing the speech rates. As a result, we find significant correlations of average speech rates between the therapist and the patient, shown in Table 1. We also see that such correlations are not obtained in random pairings of therapists and patients, as shown in Table 2.

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

4. Relating speech rate entrainment dynamics and empathy

In Sec. 3 we showed evidence that speech rates are part of the cues exemplifying behavioral entrainment. In this section we study if the degree of such entrainment contributes to the perceived therapist's empathy level in MI. We consider the turn-by-turn differences in speech rates as a computational measure for entrainment, where a turn is a period that a single speaker holds the speaking floor.

We segment the audio based on the forced alignment. We keep intra-speaker silence (defined as pause) that is longer than 0.2 seconds, while merge the others with the speech segments. In this way we retain inter-word short pauses, while keeping longer pauses separate from the calculation of speech rate. For inter-speaker silence (defined as gap), we retain all measured values without any flooring/ceiling. Overlapping speech segments exist in the corpus, but are not accessible from the alignment, so that they are left out from the current analysis. We use speech utterances longer than 0.5 seconds and discard the rest to improve the robustness of speech rate estimation. We obtain the turn level speech rate r by counting on the unit of utterances u_i ($1 \leq i \leq N_u$), as in (4).

$$r = \frac{\sum_{i=1}^{N_u} symbol\_cnt(u_i)}{\sum_{i=1}^{N_u} \left(t_{\texttt{end}}(u_i) - t_{\texttt{begin}}(u_i)\right)} \tag{4}$$

We compute the averaged absolute differences of speech rates between each patient's turn and the therapist's turn that follows. This is because our focus is on the therapist's reaction to the patient's behavior. Let $r_w(k)$ and $r_w(k+1)$ be the word rate of turns k and $k+1$ that belong to the patient and therapist, respectively. r_w for the patient and the therapist are zero mean separately, *i.e.,* subtracted the mean of the raw turn-wise speech rate, so as to remove the bias of individual speech rate baseline. We define the averaged absolute difference D_w as in (5), assuming the session contains K turns, K being an even number. We also assume the session begins with the patient's turn (index odd — patient, even — therapist); otherwise one can chop the first and/or the last turn to fit the above assumptions. Moreover, we compute DD_w as in (6) that represents the averaged absolute difference of the change in speech rate within the same individual. This can be viewed as comparing the acceleration of speech rates.

$$D_w = \frac{1}{K/2} \sum_{k=1}^{K/2} |r_w(2k-1) - r_w(2k)| \tag{5}$$

$$DD_w = \frac{1}{\frac{K}{2}-1} \sum_{k=1}^{\frac{K}{2}-1} \Big| \Big(r_w(2k+1) - r_w(2k-1)\Big) - \Big(r_w(2k+2) - r_w(2k)\Big) \Big| \tag{6}$$

We derive D_s, D_p and DD_s, DD_p in a similar manner. We hypothesize that these cues, which reflect the degree of entrainment by the therapist, should correlate with therapist's empathy level. We show the obtained correlations in Table 3. All correlations are significant (based on *t*-test) at $p < 0.001$ except D_p with $p < 0.003$, and are in negative values meaning that higher rate-differences associate with lower perceived empathy. This lends support to our hypothesis that the degree of entrainment is linked to therapist's empathy level.

Table 3: *Correlations between averaged absolute differences of speech rates and therapist empathy*

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

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

5. Analysis of speech and silence durations

The durations of speech and silence are also related to the behavioral states of the interlocutors. We segment the audio as in Sec. 4, but retain short speech utterances under 0.5 seconds. We conduct the analysis on the CTT set.

In [12] the ratio of patient utterances correlated with therapist empathy. Here we expand this to include the segment types summarized in Table 4. Let the segment durations of a particular type be denoted d_i, $i = 1, 2, \cdots, S$. Let the total duration of the session be T, which contains $N_{\texttt{seg}}$ segments. For each

type we consider four cues: (i) $\sum_{i=1}^{S} d_i/T$, (ii) S/N_{seg}, (iii) mean of d_i, (iv) standard deviation of d_i.

We show the correlations between these cues and empathy in Table 4. First, we verify that the ratios of therapist and patient speech are negatively and positively correlated with therapist empathy, respectively, as reported in [12]. Second, we find that the ratios of pause have similar correlations to empathy. Since pauses are within speaking turns, one possible interpretation is that therapist who tends to stop then grab the floor more often may seem less empathic. Third, the mean and standard deviation of therapist's pause durations are negatively correlated with empathy, while that for the speech utterances are correlated positively. This suggests that long pauses and short speech utterances may be part of negative behaviors for showing empathy. Short speech utterances like backchannels are mostly annotated as overlapped speech and not analyzed here. In addition, we see that the ratios of gap in both directions are negatively correlated with empathy. This may suggest that high frequency of speaking turn exchange is associated with low empathy.

Table 4: *Correlations between speech/silence duration cues and therapist empathy: (a) therapist's speech, (b) patient's speech, (c) therapist's pause, (d) patient's pause, (e) gap from therapist to patient, (f) gap from patient to therapist, (g) all pauses, (h) all gaps.* ***Bold***—$p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, *based on t-test*

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

6. Experiment of empathy classification

We examine if the cues proposed in this work serve as complementary features to the prosodic features introduced in [13] for classifying high *vs.* low empathy codes. The prosodic features are joint distributions of various combinations of quantized speech segment duration, energy, pitch, jitter, and shimmer cues. We select the 100 top-performing features from these in terms of their correlation with empathy codes, based on the training set. We employ the 12-dim cues of speech rate (D_x, DD_x, σ_x^T, σ_x^P, for $x \in \{w, s, p\}$) and 32-dim inter-word and inter-turn duration cues in Table 4 as additional features. For the 200 sessions in the CTT set (See Sec. 2.2), we conduct a leave-one-therapist-out cross-validation for the 133 unique therapists in the corpus. We use linear SVM as the classifier.

Table 5: *Accuracies of empathy code classification*

Feat.	Chance	Prosody	Rate	Duration	Fusion
Acc.	60.5%	72.5%	64.5%	72.0%	77.0%

In Table 5 we report the accuracies of empathy code classification (chance level baseline is 60.5%). The fusion of features improves upon each individual feature set, which suggests that the speech rate and speech/pause/gap duration features provide additional information about empathy.

7. Discussion: reliability regarding noise in speech alignment

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

To check speech rate entrainment, we add zero mean, σ_z^2 variance Gaussian noise to utterance boundaries in the Switchboard corpus. To check the correlation of speech rate difference and empathy, we add zero mean, σ_z^2 Gaussian noise to the utterance length in the CTT set. Like in Sec. 4, we eliminate utterances shorter than 0.5 seconds after adding the noise. For both cases, we sample σ_z from 0 to 1 second with a step size of 0.02 seconds. We repeat the simulation 100 times and take the averaged correlation values.

In Fig. 2 and Fig. 3 we plot the correlations. We see that the results are still significant near $\sigma_z = 0.5$, and the degradations of correlations are negligible for $\sigma_z < 0.2$. These demonstrate that the above hypotheses are robust to alignment precision.

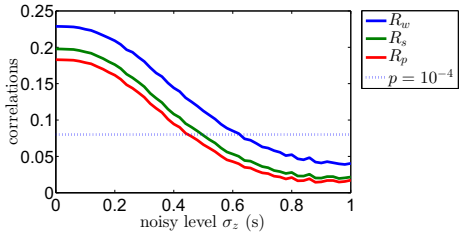

Figure 2: *Correlations of interlocutors' speech rates in simulation of noisy utterance boundaries*

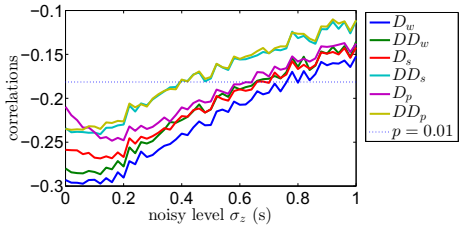

Figure 3: *Correlations of speech rate differences and empathy in simulation of noisy utterance lengths*

8. Conclusion

In this work we extracted word, syllable, and phoneme rates for interlocutors engaged in telephone conversation and addiction-counseling spoken interactions. Through statistical analyses, we showed the entrainment of interlocutors' speech rates by their positive session-wise correlations. The degree of entrainment — captured by the averaged absolute differences of turn-level speech rates of the therapist and patient — correlates with therapist's empathy rating. These relations were further verified to be robust in a simulation of noisy speech-text alignment. Moreover, we tested the correlation of ratio and duration statistics of speech, pause, and gap segments, with therapist's empathy rating. Furthermore, we employed these cues in an experiment classifying high *vs.* low empathy codes. Results showed speech rate, inter-word pause and inter-turn gap provided useful information, complementing previous prosodic cues for empathy modeling.

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

9. References

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

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

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

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

[5] P. Bellet and M. Maloney, "The importance of empathy as an interviewing skill in medicine," *Journal of the American Medical Association*, vol. 266, no. 13, pp. 1831–1832, 1991.

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

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

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

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

[10] S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceeding of IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[11] B. Xiao, D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *APSIPA ASC*, Dec. 2012.

[12] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling," in *Proc. Interspeech*, Sep. 2013.

[13] B. Xiao, D. Bone, M. Van Segbroeck, Z. E. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Modeling therapist empathy through prosody in drug addiction counseling," in *Proc. Interspeech*, Sep 2014.

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

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

[16] T. Wheatley, O. Kang, C. Parkinson, and C. Looser, "From mind perception to mental connection: Synchrony as a mechanism for social understanding," *Social and Personality Psychology Compass*, vol. 6, no. 8, pp. 589–606, 2012.

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

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

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

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

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

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

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

[24] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1. IEEE, 1992, pp. 517–520.

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

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

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

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

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

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

[31] 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.

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

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, Dec. 2011.