Multimodal Automatic Coding of Client Behavior in Motivational Interviewing

Leili Tavabi
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA
ltavabi@usc.edu

Kalin Stefanov
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA
kstefanov@ict.usc.edu

Larry Zhang
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA
lzhang@ict.usc.edu

Brian Borsari
VA hospital San Francisco
University of California San Francisco
San Francisco, CA, USA
Brian.Borsari@va.gov

Joshua D Woolley
VA hospital San Francisco
University of California San Francisco
San Francisco, CA, USA
josh.woolley@ucsf.edu

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

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

ABSTRACT

Motivational Interviewing (MI) is defined as a collaborative conversation style that evokes the client's own intrinsic reasons for behavioral change. In MI research, the clients' attitude (willingness or resistance) toward change as expressed through language, has been identified as an important indicator of their subsequent behavior change. Automated coding of these indicators provides systematic and efficient means for the analysis and assessment of MI therapy sessions. In this paper, we study and analyze behavioral cues in client language and speech that bear indications of the client's behavior toward change during a therapy session, using a database of dyadic motivational interviews between therapists and clients with alcohol-related problems. Deep language and voice encoders, *i.e.*, BERT and VGGish, trained on large amounts of data are used to extract features from each utterance. We develop a neural network to automatically detect the MI codes using both the clients' and therapists' language and clients' voice, and demonstrate the importance of semantic context in such detection. Additionally, we develop machine learning models for predicting alcohol-use behavioral outcomes of clients through language and voice analysis. Our analysis demonstrates that we are able to estimate MI codes using clients' textual utterances along with preceding textual context from both the therapist and client, reaching an F1-score of 0.72 for a speaker-independent three-class classification. We also report initial results for using the clients' data for predicting behavioral outcomes, which outlines the direction for future work.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Information extraction; Neural networks; • **Human-centered computing** → Laboratory experiments.

KEYWORDS

motivational interviewing; mental health; machine learning; human behavior

ACM Reference Format:
Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal Automatic Coding of Client Behavior in Motivational Interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3382507.3418853

1 INTRODUCTION

Motivational Interviewing (MI) is a client-centered counseling approach for eliciting behavior change, through exploring and resolving ambivalence. MI focuses on natural language with strategies for eliciting change toward a positive outcome [21]. MI has broad applications in different domains including substance abuse and health-related behaviors [14, 15].

Iterations of the Motivational Interviewing Skill Code (MISC) [20] have been used to code therapist and client language during MI sessions, permitting the analysis of the link between the client's language and subsequent behavioral outcome. The client utterances are distinguished based on language suggesting willingness or resistance to change and are divided into three main categories: (i) *Change Talk (CT)* signaling willingness to change, (ii) *Sustain Talk*

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418853

(ST) indicating a desire not to change, or preserve the status quo and (iii) *Follow/Neutral (FN)* which is language unrelated to change (*e.g.*, commenting on politics). It has been shown that the in-session client language encompassing the mentioned types are prominent indicators for subsequent changes in a client's behavior [16, 22].

In this paper, we analyze the client's and therapist's speech via linguistic and prosodic modalities to develop models for estimating codings of client utterances throughout the session. We additionally use the same multimodal data to determine its ability to predict the client's behavioral outcome (desired or undesired change of behavior) following the session. Although no causal relationship between the clients' in-session data and their subsequent behavior change should be inferred due to the limited available information in addition to many uncontrolled factors, we believe this analysis provides valuable insights on possible associations between the clients' verbal and speech behaviors regarding change as obtained from their in-session data and their actual behavior change. To this end, we leverage a real-world dataset of client-therapist MI sessions from college students dealing with alcohol-related issues. We obtain their "Change in Typical Blood Alcohol Content" and "Change in Alcohol Related Problems" as two target behavioral outcomes after the session. We focus on two problems: (i) Estimating MI codes using textual and speech utterances and (ii) Predicting subsequent behavioral changes from clients' in-session data, which we formulate as 3- and 2-class classifications respectively.

The major contributions of this paper are as follows,

- analysis of the associations between client language and speech prosody with the corresponding MI codes;
- providing a machine learning framework for detecting MI codes using language and speech in real-world MI sessions; and
- analysis for prediction of the behavioral outcomes using clients' data corresponding to individual MI codes.

2 BACKGROUND

2.1 Behavioral Codings in Therapy

Previous research has shown that doctor-patient communication during a therapy session can provide strong insights on patient symptoms, effectiveness of the therapy toward the target outcome, and the patients' future adherence to the treatment. Multiple studies have investigated the associations of MI codings with target behavioral outcomes, and behavior change toward a specific goal. In a meta-analysis of 12 individual studies, sustain talk was shown to be a strong indicator of negative behavioral outcomes. A further sub-analysis on studies analyzing composite client language (*e.g.*, no. of Change Talk / (no. of Change Talk + no. of Sustain Talk)) has shown positive association with behavioral change [18].

A large body of work has focused on using the client's and therapist's linguistic indicators throughout the sessions to predict the behavioral outcome of the treatment. Howes et al. [11] examines the predictive power of automatically extracted topics using Latent Dirichlet Allocation (LDA) in predicting therapy outcomes for schizophrenia. They show that automatic and manual codings of topics are each effective for prediction of a different target variable (*e.g.*, manual topics allow for prediction of symptoms, etc). Using task-related behavioral codings of the client/therapist language has shown to be effective in estimating the final outcomes of the therapy sessions in Motivational Interviewing, Cognitive Behavioral Therapy, etc [17, 29]. Multiple efforts have therefore focused on the automated prediction of such behavioral codings in order to avoid the costly and time-consuming manual effort of annotation. Chen et al. [5] uses automatically-tagged behavioral codings adapted from MI, along with dialogue acts, word-level and utterance-level linguistic features of the therapist for an end-to-end assessment of Cognitive Behavioral Therapy (CBT) sessions. Ewbank et al. [7] also proposes a neural network classification model for MI behavioral codings, used in CBT sessions. They focus on five categorical behavioral codings from MI (change-talk active, change-talk explore, follow/neutral, sustain talk and describing problems) for a multi-label classification of client utterances. They encode both client and therapist utterances as a sequence of word embeddings obtained by word2vec [19], along with an added dimension to represent the speaker role, and feed the sequence of utterance embeddings to a bidirectional Long Short-Term Memory (LSTM). They further look into associations of the selected behavioral codings with the desired outcome of the CBT sessions by running a logistic regression, and observe that the quantity of sustain talk was negatively associated with reliable improvement. Xiao et al. [30] leverages bi-directional Gated Recurrent Units (GRU) on sequences of word embeddings pretrained on in-domain data to predict therapist and client codings. Huang et al. [12] combines the topic- and word-level content of the current utterance, verbal context (five previous therapist utterances), and codes (ten previous codes), along with a domain adaptation mechanism on topic embeddings. They use this data for code classification of individual MI sessions, and examine how the distribution of content changes across time stages within sessions. In other domains of therapy, Tseng et al. [28] approaches human behavior estimation in couple's therapy, in which couples with real marital issues discuss selected topics. They first extract semantic information using seq2seq models into deep sentence embeddings, which are then fed into a Recurrent Neural Network (RNN) for estimating the behavioral codings of each speaker. The models are trained for automated coding of negative sentiment by attributing the codings from the entire session to all the utterances within that session, as a form of data augmentation.

Compared to the existing literature for behavioral code prediction using linguistic content, the multimodal domain remains relatively less explored. Black et al. [2] uses speech prosody features toward measuring different emotional cues within sessions of married couples partaking in problem-solving interactions. They use prosodic, spectral, and voice quality features to capture global acoustic properties for each spouse and trained gender-specific and gender-independent classifiers for classification of the extreme instances into six selected codes (*e.g.*, "low" versus "high" blame). Singla et al. [27] uses a multimodal approach, combining prosodic information, speech pause information, and lexical information, to classify and predict CBT codes using LSTM models. The work demonstrates improved results when using attention mechanism on multimodal data.

Aswamenakul et al. [1] utilizes speech features using COVAREP [6], linguistic features using LIWC [23], and GloVe [24] word embeddings to predict MI client codes. The statistical analysis of COVAREP speech quality features shows statistically significant differences between change talk versus follow/neutral utterances and sustain talk versus follow/neutral utterances, while minimal differences between change talk versus sustain talk utterances. This suggests it may be more difficult for models to distinguish change talk and sustain talk based on acoustic features. Their multimodal fusion model combining verbal and non-verbal behavior outperforms their unimodal models, while showing text as the significantly more powerful data stream.

Existing work mostly focus on the linguistic content of the therapy sessions, leaving the multimodal aspects open for exploration. Additionally there is still much room for exploring associations of MI codings with behavioral outcomes. In this paper, we investigate the predictability of MI codings using multimodal interaction data from real-world MI sessions. Additionally, we explore the predictability of behavioral outcomes using multimodal in-session data to gain more insights into their possible associations.

3 DATA

In this work, we utilized two clinical datasets [3, 4] from real-world motivational interviewing sessions with college students involving alcohol-related issues. The datasets consist of audio recordings, manual transcriptions, and MISC codes. The study has been IRB-approved, and the data collection has been performed with the consent of the participating volunteers. The transcriptions include the sessions' metadata including manual MISC codings for both the therapist and client utterances, along with the speaker tag (client/therapist) for each utterance. Table 1 shows an example segment of a client-therapist dialogue.

To obtain the start and end timestamps of the utterances for speech processing, we use Speechmatics, which is an automatic tool using individual sessions' text and audio files as input to obtain word-level timestamps. Using this dataset, we have access to real-world sessions from 219 individual clients with 12 unique therapists. The clients involved in this dataset have an average age of 18.8 years with a 40:60 female to male ratio. The average length of the utterances is 5.31 seconds, and the average duration of the sessions is 49.85 minutes. The dataset consists of a total of 41,494 client utterances and 51,802 therapist utterances. In this work, we primarily focus on the classification of client utterances into three main categories of MI codings, while taking into account the preceding utterances from both the therapist and client as part of the history context. A subset of the sessions also include behavioral measures related to the therapy's desired outcome, reduced alcohol consumption. For each session, we have two behavioral measures: Change in Typical Blood Alcohol Content (CTBAC), and Change in Alcohol-Related Problems (CPROB). Blood Alcohol Content (BAC) and Alcohol-Related Problems inventory was administered during the MI session, and at a 6-month follow-up. The change of these measures over the course of this 6-month period is used in our analysis as the behavioral outcome measure. A positive value for both CTBAC and CPROB indicates an increase in blood alcohol content or alcohol-related problems and therefore indicates an undesirable outcome, and vice versa. Additionally zero change in both of these measures is also considered as undesired outcome, since it suggests the therapy session had not been fully effective. Based on the value of CTBAC and CPROB measures, we divide the sessions into two main categories, subsequent desired or undesired outcome. Out of the entire 219 sessions, we have 166 sessions with behavioral outcomes. Tables 2 and 3 show the distribution of the data across the MI codes and behavioral outcomes which indicates an imbalanced dataset in both cases.

Table 1: Example interaction segment from the dataset

Speaker	Transcript	MISC Code
Therapist	I mean, it sounds to me like, when you tell me that your average is like five to ten drinks, so that's already a heavy-drinking episode.	Complex Reflection
Client	Yeah, yeah. I guess that contributes to heavy-drinking.	Follow/Neutral
Client	But like, and like I'm sure you hear this all the time—but like, I have a couple friends that are like way past me, and they drink a lot more, so yeah I wouldn't consider myself like a heavy drinker.	Sustain Talk
Client	And I realize like according to this, I am, but I do realize its bad	Change Talk
Therapist	So you're comparing yourself to the people you're around.	Simple Reflection
Client	Yeah, right, like my friends that are like most like me.	Follow/Neutral

Table 2: MI codes class distribution.

Sustain Talk	Follow/Neutral	Change Talk
0.13	0.59	0.28

Table 3: Outcomes class distribution

	Undesired	Desired
Blood Alcohol Content	0.55	0.45
Alcohol-Related Problems	0.70	0.30

4 METHODOLOGY

4.1 Multimodal Feature Extraction

4.1.1 Textual Features. We used two feature-sets for statistical analysis and representation of the text modality for the prediction model.

LIWC. The Linguistic Inquiry Word Count (LIWC) is a dictionary-based tool that assigns scores to documents in psychologically meaningful categories including social, affective and cognitive processes [23]. We used LIWC for our statistical analysis due to its interpretability and for the purpose of identifying important textual features in separating utterances with different MI codes.

BERT. For the text representation in our classification models, we extract embeddings from the pretrained language model Bidirectional Encoder Representations from Transformers (BERT). BERT is an unsupervised language representation model, pre-trained on large corpora of text, and it has provided significant advancements to different tasks in Natural Language Processing (NLP) including text classification. We therefore use BERT to take advantage of its powerful pre-trained representations. We extract BERT embeddings (using bert-base-uncased) per utterance, for both clients and therapists, and obtain 768-dimensional representational vectors.

4.1.2 Speech Features. Two different feature-sets were used for statistical analysis and representation of speech prosody.

eGeMAPS. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) provides a set of interpretable acoustic features obtained by speech processing experts for their potential to detect affect in speech. eGeMAPS has been widely used in literature due to its performance in emotion recognition tasks, as well as theoretical significance [8]. This feature set consists of 23 features such as fundamental frequency and loudness. We use this feature set in our statistical analysis to gain insights about the most significant features in our task.

VGGish. For the audio representation in our classification models, we extract features from a pre-trained deep convolutional neural network, inspired by the VGG networks used for image classification. VGGish is pre-trained on audio spectograms extracted from a large database of audio event categories, for example, vehicle noise, music genre and human locomotion [9, 10]. VGGish, like other models pre-trained on large datasets, has shown to be able to generate powerful represention, and was selected in this work due to convincing evidence from literature for its effectiveness in emotion recognition tasks [13, 25]. To extract the pre-trained embeddings, the audio files were first converted to log-mel spectrogram and the resulting images were passed to a modified VGG deep convolutional neural network [26] for recognizing Audioset classes. We obtained the 128-dimensional embeddings, generated by VGGish after dimensionality reduction with Principal Component Analysis (PCA)[1]. We used a hop size of 0.96s, meaning a 128-dimensional vector was extracted for every 0.96s of the audio signals. As a result, each audio sequence is represented by a $T \times 128$ matrix, where T is the number of timesteps.

4.2 Statistical Behavior Analysis

4.2.1 MI Code Analysis. To study the verbal and nonverbal indicators associated with types of behavioral codes in MI sessions, we used interpretable feature sets from both text and speech to investigate the associations. We used eGeMAPS [8] for speech, and LIWC [23] for language.

To analyze speech and linguistic features, we executed hierarchical Analysis of Variance (ANOVA) across the features, with type of speech being nested under subjects. We observed the F-statistic for statistical significance and reported the most significant features in Tables 4 and 5 for text and speech respectively. All p-values are very close to zero ($< 10^{-4}$) and therefore not reported. The significant results in this test indicate that the mean of a feature is significantly different in at least one out of the three classes.

Among the significant features obtained using LIWC are "informal" terms including "assent", *i.e.*, words like "agree", "yes", "ok" which are highly frequent in utterances categorized as follow/neutral. Additionally words per sentence tends to be lower for follow/neutral instances compared to change talk and sustain talk, which is likely the reason it is a strong discriminant.

The effect sizes for acoustic features are relatively smaller. Significant features include loudness of voice, pitch, spectral flux (the rate of spectral change in speech), harmonics-to-noise-ratio and the first MFCC coefficient.

Table 4: Five most statistically significant features from text obtained using hierarchical ANOVA

Feature	F-Statistic
Informal	7.880
Function	7.215
Assent	7.160
Words per sentence	6.966
Analytic	6.443

Table 5: Five most statistically significant features from speech obtained using hierarchical ANOVA

Feature	F-Statistic
Loudness	3.662
Spectral Flux	2.898
Pitch	2.818
Harmonics-to-noise ratio (HNR)	2.768
Mel-Frequency Cepstral Coefficient 1 (MFCC1)	2.476

4.3 MI Code Prediction Models

For the unimodal text models, we examine the model's prediction performance using two sets of data: 1) Taking only the client's current utterance and 2) Using the client's current utterance and the history context from preceding client and therapist utterances. For the unimodal speech model, we focus on the client's speech from the current utterance.

4.3.1 Client Utterances. For both unimodal text and speech models, an encoder first maps the BERT and VGGish embeddings from client utterances to fixed-size vector embeddings. An instance-based encoder maps the BERT input vectors to fixed-size representations. For speech, sequences of input VGGish embeddings are fed to a single-layer GRU, taking the last layer as the speech embedding. The embeddings from both text and speech are fed into a final classification fully-connected (FC) layer.

[1] https://github.com/tensorflow/models/tree/master/research/audioset

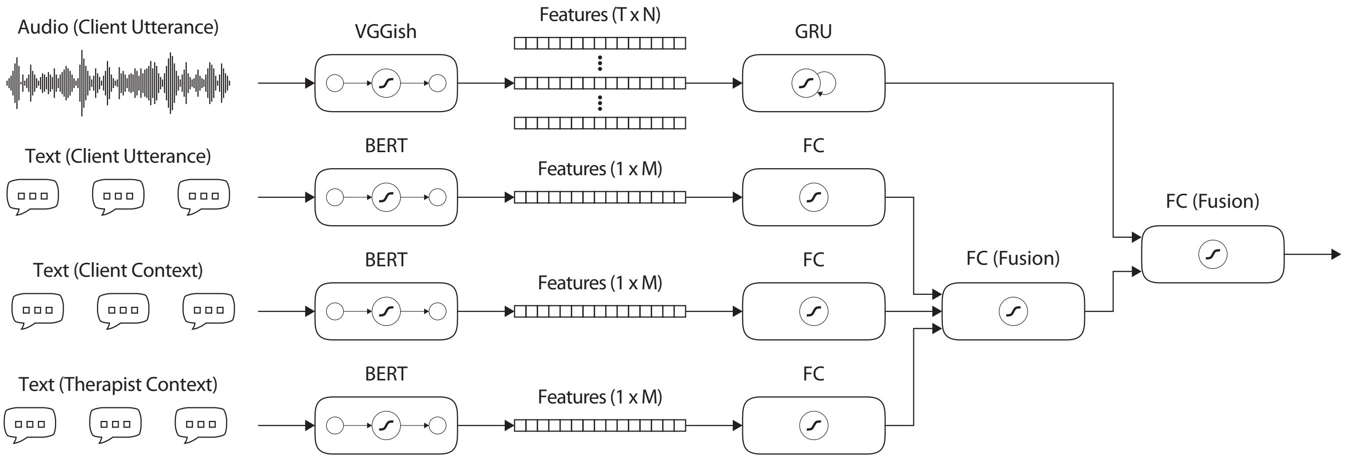

Figure 1: Bimodal fusion network used for prediction of MI codes.

4.3.2 Client Utterance and History Context. Three client-therapist dialogue turns preceding the current client utterance are extracted to represent context. The size of the preceding context is selected based on empirical analyses. The contextual client and therapist utterances are encoded separately to account for the inherent differences of the used language based on their roles, by being fed into single FC layers. Current client utterance is also encoded by using an FC layer of equal size. The current utterance as well as contextual utterances are encoded in three fixed-sized representation vectors. These vectors are concatenated and passed through a hidden linear layer, before being fed to the final classification layer.

4.3.3 Bimodal Fusion. For the fusion model, the individually trained text and speech models are loaded with their last classification layers removed. The obtained representation vectors from the two modalities are concatenated and passed through an FC layer, which is ultimately fed into the final classification layer. An illustration of the model architecture is provided in Figure 1.

4.4 Outcome Prediction Models

In this section, we investigate the associations and predictive power of clients' language and speech content pertaining to individual MI codings with the behavioral outcomes. Toward this goal, we develop a model that takes as input, the sequence of client utterances pertaining to specific codes. For example we extracted all the utterances labeled as 'change talk' from each session, obtaining a stream of data per code type per session. We also compare the predictive power of individual codes with the entire sequence of client utterances from the session. We train individual models per modality per stream of data using a universal architecture, to gain an understanding of the importance of each code type in prediction. The sessions' sequential data is passed to a fixed-sized GRU for obtaining the representation vectors taking the last state, which is subsequently fed into the final classification layer. Similar to the fusion model used for MI codes, we leverage the individually trained models for text and speech to obtain vector embeddings from the last hidden layer. These text and speech embeddings are concatenated and passed through a hidden FC layer before being fed into the final classification layer.

4.5 Experimental Setup

In this paper, we work on a dataset of 219 real-world MI sessions with 219 different clients. We extract the client utterances as our data points, with their corresponding manually-coded MISC codes. The datasets of client utterances amount to a size of 41,494 data points, on which we perform a 3-class classification, with a one-subject-out cross validation. The dataset is imbalanced between the three classes with 'Sustain Talk' as the minority class. To handle the data imbalance, a cross-entropy loss is used with a weight vector, where the weight of each class is inversely proportional to its frequency. The weights are learned from the train set within each fold, where 10% of the train data is held out as the validation set. The evaluation results for the 3-class classification of MISC codes are computed using F1-score and the model with the best performance on the validation set is selected for each fold. We optimize the network using Adam, with a batch size of 256 and a learning rate of 10^{-3} for unimodal and 10^{-5} for fusion models. The fusion models load individually trained text and speech models without freezing to obtain the embeddings.

For text and speech, we use BERT and VGGish pre-trained models to take advantage of the powerful embeddings obtained from pre-training on large amounts of data. The text and speech embeddings consists of 1×768 and $T \times 128$ embeddings respectively, where T represents the number of time steps. For both modalities, we designed an encoder network mapping the input feature space to a 256-d embedding space. The speech inputs are fed into a 1-layer GRU taking the output vector, and the text inputs are fed into an FC layer due to their latent temporal dimension. The 256-d representational vectors are then fed to the classification layer for unimodal models. The multimodal model takes the same 256-d embeddings from the previously trained text and speech models, concatenates the two vectors and feeds them to the last 256-d FC layer before classification.

For prediction of our two target behavioral outcomes, we use the same data on the session level. We take the subjects with either positive or negative behavioral changes (including those with no change at all). We have 166 subjects with Change in Alcohol-Related Problem labels and Change in Blood Alcohol Level, which is the data our behavior prediction models are trained with. We extract the client utterances from each session as a sequence, removing the utterances from the therapist. We approach this 2-class classification of behavioral changes following different perspectives: First, to identify whether data pertaining to specific MISC codes have higher prediction power therefore, looking at sequences of data pertaining to change talk, sustain talk and follow/neutral codes individually; and second, to use the client data from the session holistically for the prediction. Similar to the models used in MI code prediction, the outcome prediction models also consist of 256-d encoding and fusion layers.

5 RESULTS AND DISCUSSION

Table 6 demonstrates the classification results for our unimodal and multimodal estimation of MI codes. The unimodal results demonstrate the superiority of text compared to speech in predicting MISC codes and content representation. This is not surprising due to the amount of meaningful content inherently carried through the text modality, which is amplified by powerful representation models such as BERT. The significant difference between the classification performance between text and speech is also possibly extenuated by low-quality speech files. The noise in the recordings also makes it difficult to capture pauses, which are significant indicators of sincerity in speech. We compare our classification performance with the results from the most relevant previous work using a similar dataset and problem formulation [1]. They take a multimodal approach for a 3-class classification of client utterance codes, similar to our work. They use pretrained word embeddings GloVe (Global Vectors for Word Representation) and LIWC for the text modality, and use COVAREP [6] for speech. They train logistic regressions models by using different combinations of the three feature sets, and show that using all three feature sets obtains the highest classification performance. Although results from different combinations show that speech makes a minor improvement over the text-only model. Our models outperform this baseline from previous work, reaching F1 score of 0.721 compared to the previous 0.566, by encoding historical context as well the client utterance, while also taking advantage of more advanced encoders like BERT and VGGish.

Our results demonstrate that adding history context to our text model obtains statistically significant improvement compared to using only the current client utterance. There is a large performance gap between our text and speech models, although our speech model still significantly outperforms the speech model provided by the baseline. The multimodal results slightly underperform the text results, which we believe is mainly due to low-quality speech files. The speech data is subject to further investigation for de-noising and pre-processing for improvement of the multimodal results.

Table 7 shows the model performance across the three classes. It can be seen that the hardest class for the model to recognize is "sustain talk", which could be partly due to its very low frequency in the dataset. In future work, we will use oversampling for a more better representation of the minority class.

Table 6: MISC codes estimation results for three-class classification; multimodal baseline obtained from previous work with similar dataset [1]. Average micro F1-scores and their standard deviations (in parentheses) are given.

Modality	Data/Model	Micro F1-score
Text	Utterance	0.701 (0.065)
	Utterance + Context	**0.721** (0.062)
Speech	Utterance	0.531 (0.086)
Multimodal	Our model	0.714 (0.063)
	Late Fusion	0.702 (0.064)
Multimodal Baseline [1]		0.566

Table 7: Precision and recall for the code prediction model

	Precision	Recall	F1
Sustain Talk	0.43	0.53	0.47
Follow/Neutral	0.85	0.77	0.81
Change Talk	0.60	0.66	0.63

The confusion matrix from the code prediction can be found in Table 8. It can be seen that the misclassification of "sustain talk" instances is mostly due its confusion with "change talk", which is aligned with observations from previous work indicating that sustain talk and change talk are more difficult to discriminate. [1].

Table 8: Confusion matrix for the three-class code prediction model (ST: Sustain Talk, FN: Follow/Neutral, CT: Change Talk)

	ST	FN	CT
ST	0.53	0.19	0.28
FN	0.08	0.77	0.15
CT	0.15	0.19	0.66

Table 9 shows the results from the binary outcome prediction models for both target behaviors. The results are reported for individual MI codes as well as the entire sequence of client utterances (referred to as 'all'), for assessing their influence on the behavioral outcome. We observe that using the entire sequence provides the best prediction for both outcomes, while noting that the multimodal model obtains the highest performance for CTBAC where the text-only model outperforms other models for CPROB. Among the individual talk types, there is no consistent pattern of one talk type significantly outperforming others in prediction. Further investigation is needed to evaluate the predictive power of each MI code with the outcome. Overall, the outcome prediction results show marginal improvement over the chance baseline, which further validates the challenge associated with predicting human behavior by accessing only a small interaction window. Incorporating personal information like age, gender and family history with alcohol abuse can better facilitate this behavior prediction in future work.

Table 9: Outcome estimation results for two-class classification. Random baseline (mean F1-score over 1000 trials) and F1-scores over the entire dataset are given. (CTBAC: Change in Typical Blood Alcohol Content, CPROB: Change in Alcohol-Related Problems)

Modality	Data	F1-score	
		CTBAC	CPROB
Text	CT	0.517	0.494
	FN	0.507	0.491
	ST	0.545	0.416
	ALL	0.495	**0.535**
Speech	CT	0.563	0.403
	FN	0.560	0.444
	ST	0.565	0.364
	ALL	0.548	0.362
Multimodal	CT	0.494	0.524
	FN	0.512	0.506
	ST	0.482	0.482
	ALL	**0.578**	0.518
Random Baseline		0.473	0.376

Even though the proposed method is far from being able to predict the therapy outcome, predicting one's alcohol abuse risk raises ethical concerns. There are a number of scenarios where this information might be abused to discriminate against individuals with high risk. Such predictions are also far from perfect which might exacerbate the risk of their deployment. It is imperative that the future development and deployment of such systems would fully inform the users about the errors and implications of predictions. Patients should have the full authority to choose who would access such information, and these tools should be empowering the patients to make decisions based on the machine-based prediction of their therapeutic outcome.

6 CONCLUSIONS

In this paper, we reported on our work for automatic estimation of MI codes. To this end, we analyzed and modeled MI codes using text and speech data from real-word MI therapy sessions for alcohol-related issues. We developed and evaluated a neural network model for estimation of MI codes in both unimodal and multimodal fashion. Our analysis indicates the superiority of text achieved by encoding language with a pre-trained transformer network (BERT), providing the best unimodal results. We also modeled the context of client utterances by encoding the preceding utterances which resulted in improved performance. The bimodal fusion of text and speech slightly underperformed the unimodal text models. The unimodal text and bimodal text fusion models obtained similar micro F1 classification scores of 0.721 for a leave-one-subject-out 3-class classification.

We also experimented with using clients' text and speech data in predicting the subsequent behavioral outcome. We used the sequences of client utterances with specific MI codes individually (*e.g.*, extracting all CT utterances) and compared the predictive power across different MI codes and also the entire client data regardless of codes. Our results show marginal improvement over the chance baseline, which speaks to the challenge of the task in hand. Future work will focus on exploring the possibility of using client's personal information such as age, gender, family history of alcohol-related issues in addition to the interaction data for building informed models for prediction.

Automated estimation of MISC codes using machine learning shows great promise in providing an objective and cost-effective means for the analysis and assessment of MI and other psychotherapy sessions. It also enables the use of such codings for providing the therapists with analytical means of better understanding their clients' behavioral outcomes and and the efficacy of their therapeutic strategies. With this work, we aim to facilitate therapists with the tools for better assisting clients in reaching their desired behavioral outcomes.

ACKNOWLEDGMENTS

This work was supported by National Institute on Alcohol Abuse and Alcoholism grant R01 AA027225, R01 AA017427 and R01 AA12518. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health, or the Department of Veterans Affairs or the United States Government. The authors would like to thank the clients and therapists who allowed their audiotapes to be utilized for this study.

REFERENCES

[1] Chanuwas Aswamenakul, Lixing Liu, Kate B. Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Multimodal Analysis of Client Behavioral Change Coding in Motivational Interviewing. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 356–360. https://doi.org/10.1145/3242969.3242990

[2] Matthew P. Black, Athanasios Katsamanis, Brian R. Baucom, Chi-Chun Lee, Adam C. Lammert, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2013. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication* 55, 1 (Jan. 2013). https://doi.org/10.1016/j.specom.2011.12.003

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

[4] Kate B Carey, James M Henson, Michael P Carey, and Stephen A Maisto. 2009. Computer versus in-person intervention for students violating campus alcohol policy. *Journal of consulting and clinical psychology* 77, 1 (2009), 74.

[5] Zhuohao Chen, Nikolaos Flemotomos, Victor Ardulov, Torrey A Creed, Zac E Imel, David C Atkins, and Shrikanth Narayanan. 2020. Feature Fusion Strategies for End-to-End Evaluation of Cognitive Behavior Therapy Sessions. *arXiv preprint arXiv:2005.07809* (2020).

[6] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[7] MP Ewbank, R Cummins, V Tablan, A Catarino, S Buchholz, and AD Blackwell. 2020. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research* (2020), 1–13.

[8] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (apr 2016), 190–202.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[11] Christine Howes, Matthew Purver, and Rose McCabe. 2013. Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical informatics insights* 6 (2013), BII–S11661.

[12] Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling Temporality of Human Intentions by Domain Adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 696–701. https://doi.org/10.18653/v1/D18-1074

[13] Wei Jiang, Zheng Wang, Jesse S Jin, Xianfeng Han, and Chunguang Li. 2019. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors* 19, 12 (2019), 2730.

[14] Brad Lundahl and Brian L Burke. 2009. The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses. *Journal of clinical psychology* 65, 11 (2009), 1232–1245.

[15] Brad W Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, and Brian L Burke. 2010. A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on social work practice* 20, 2 (2010), 137–160.

[16] Molly Magill, Timothy R. Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca E.F. Gordon, J. Scott Tonigan, and Theresa Moyers. 2018. A Meta-Analysis of Motivational Interviewing Process: Technical, Relational, and Conditional Process Models of Change. *Journal of consulting and clinical psychology* 86, 2 (Feb. 2018), 140–157. https://doi.org/10.1037/ccp0000250

[17] Molly Magill, Timothy R Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca EF Gordon, J Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of consulting and clinical psychology* 86, 2 (2018), 140.

[18] Molly Magill, Jacques Gaume, Timothy R. Apodaca, Justin Walthers, Nadine R. Mastroleo, Brian Borsari, and Richard Longabaugh. 2014. The Technical Hypothesis of Motivational Interviewing: A Meta-Analysis of MI's Key Causal Model. *Journal of consulting and clinical psychology* 82, 6 (Dec. 2014), 973–983. https://doi.org/10.1037/a0036833

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[20] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2003).

[21] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change.* Guilford press.

[22] Brian T. Pace, Aaron Dembe, Christina S. Soma, Scott A. Baldwin, David C. Atkins, and Zac E. Imel. 2017. A Multivariate Meta-Analysis of Motivational Interviewing Process and Outcome. *Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors* 31, 5 (Aug. 2017), 524–533. https://doi.org/10.1037/adb0000280

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

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

[25] Balasubramanian Raman and Partha Pratim Roy. [n.d.]. A Segment Level Approach to Speech Emotion Recognition using Transfer Learning. ([n. d.]).

[26] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[27] Karan Singla, Zhuohao Chen, Nikolaos Flemotomos, James Gibson, Dogan Can, David Atkins, and Shrikanth Narayanan. 2018. Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy. In *Interspeech 2018*. ISCA, 3413–3417. https://doi.org/10.21437/Interspeech.2018-2551

[28] Shao-Yen Tseng, Brian R Baucom, and Panayiotis G Georgiou. 2017. Approaching Human Performance in Behavior Estimation in Couples Therapy Using Deep Sentence Embeddings.. In *INTERSPEECH*. 3291–3295.

[29] Henny A Westra. 2011. Comparing the predictive capacity of observed in-session resistance to self-reported motivation in cognitive behavioral therapy. *Behaviour research and therapy* 49, 2 (2011), 106–113.

[30] Bo Xiao, Doğan Can, James Gibson, Zac E. Imel, David C. Atkins, Panayiotis Georgiou, and Shrikanth S. Narayanan. 2016. Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks. 908–912. https://doi.org/10.21437/Interspeech.2016-1560