

# Analysis of Speech Features in Identifying Client's Change Talk in Motivational Interviewing

1<sup>st</sup> Shareef Babu Kalluri

*School of Computer Science*

*University of Petroleum and Energy Studies*

Dehradun, India, shareef@ddn.upes.ac.in

2<sup>nd</sup> Deepu Vijayasenan

*Dept. of ECE*

*National Institute of Technology Karnataka Surathkal*

Mangalore, India, deepuv@nitk.edu.in

**Abstract**—Motivational Interviewing (MI) is a frequently used and effective psychotherapy approach for treating behavioral problems. MI is a collaborative interaction for understanding the client's own reasoning for a change in behavior. In this study, we analyzed an MI corpus in the nutrition and fitness domains, in which counselor and client utterances were categorized using Motivational Interviewing Skill Code (MISC). During the interaction when the client expresses the need or willingness to change is the Change Talk (CT). We aimed to analyze the speech features and proposed a BiLSTM multimodal neural network model for detecting the change talk or not a change talk. Our approach using speech and language information in detecting the change talk is par with other multimodal approaches(language and facial information) with an F1-score of 0.573 for CT. The proposed set of speech features shows the statistical significance in identifying the change talk or not a change talk.

**Index Terms**—Motivational Interviewing, Speech features, Change Talk, Multimodal neural networks

## I. INTRODUCTION

Motivational Interviewing (MI) is a client-centered approach that focuses on exploring and resolving clients' ambivalence and facilitating a resolution for their behavioral change. The main goal of MI techniques is to elicit the clients' own internal motivations for altering their behavior in the desired direction [1]. MI was initially developed to address problems with substance misuse. But, its use has been expanded to address a wide range of psychological and physical health issues, including healthy lifestyle choices, drug and alcohol misuse, and gambling [2], [3]. The occurrence of MI is rising, demanding enhanced therapeutic quality and reliability [4]. Motivational interviewing is practiced in conjunction with a perceptive, knowledgeable therapist or counselor who may assist a client in accessing their inherent motivation before reflecting and reinforcing those ideas to the client. MI seeks to motivate individual Change Talk (CT), by encouraging the client's willingness to make changes through discourse. The Motivational Interviewing Skill Code (MISC) was created to assess the quality of MI [5]. MISC provides different categories for counselor and client utterances to assess the quality of MI. All the client utterance subgroups are categorized into the following three broader categories.

- Change Talk (CT) – refers to utterances that reflect positive reactions toward identifying and changing target behavior.

- Sustain Talk (ST) – refers to the utterances that reflect a desire to avoid the behavior change or to preserve the present state.
- Follow Neutral (FN) – refers to the utterances that show neither client inclination toward or away from the desired behavior change.

It is harmful to our well-being to live an unhealthy lifestyle that fails to acknowledge eating habits and sleep deprivation. However, changing ingrained lifestyle habits is difficult. Motivational interviewing could help in refining nutritional diet and physical activities for individuals who were in an ambivalent stage to change themselves. In this work, we adopted the MI dataset which is collected from Japanese nationals who attend MI sessions in a non-medical setting.

The MISC categories of client and counselor are analyzed by the MI trainers and trainees from the audio/video sessions. Manually analyzing these data requires both time and resources. Consequently, it is anticipated that automatic utterance categorization in MI will benefit the development of MI skills. It helps MI trainers, to assist trainees in efficient and effective approaches. Additionally, it aids trained counselors in comprehending how clients' behaviors alter during treatment sessions.

There are many attempts in the literature to classify the client change talk, sustain talk and follow neutral categories in an MI session besides automatically categorizing the utterances using MISC. In this task, linguistic features are important and are reliably helpful [6]. Cao *et.al* [7] used the language information in GRU models in classifying the counselor and client categories. Tavabi *et.al* also used the language information in categorizing the client and counselor utterances using the RoBERTa pre-trained model [8]. Few other studies that use only language information for classifying the MI client utterance categories using GRU [8], [9], LSTM with attention mechanism [10] and with RNN models [11].

However, there have been very few studies that address the utilization of speech features for detecting and classifying client categories in MI sessions. Aswamenakul *et.al* used the 74-dimensional speech descriptors using COVAREP API and analysed a set of audio features, and investigated how they are contributing in classifying the change talk and sustain talk, along with these audio features, Aswamenakul *et.al* uses language information obtained from pre-trained word

embedding (GloVe) and linguistic inquiry word count (LIWC) to train a logistic regression model and reported that the audio descriptors from COVAREP are not greatly contributing to classifying CT and ST in MI sessions rather than language embedding features [4]. Tavabi *et.al* used BERT and VGGish pre-trained models for obtaining the language and speech embeddings. The client speech embedding is encoded with GRU model and language embedding is obtained from BERT for both client and counselor, and they are fused to classify the client categories [12].

Another recent work [13], uses language and facial information in classifying the change talk or non-change talk. Authors have used the context information from the previous 5 utterances for both language and facial information for training the BiLSTM and GRU models, and also reported that the facial information from the counselor also contributes to classifying client utterances.

As discussed above, in the literature, there have been successful attempts in detecting the change talk using language and facial information. However, facial information is a sensitive detail, where one's identity could be revealed from the data, which could be not favorable for every client. Therefore, in this work, we attempted to explore speech and language information in detecting the change talk from the client's utterances, where this information is less sensitive and easier to record compared to facial information. In this work, we made a successful attempt on par with the state-of-the-art results using language and face information in detecting the client's change talk (CT) or not with the help of speech and language information using bidirectional long and short-term memory (BiLSTM) neural network.

## II. DATASET

We have adopted the MI dataset which focuses on changing the diet plan of the client [13]. This dataset is not from a medical setting rather it was collected with Zoom remote meeting platform in concern with COVID-19 spread. The counselors participated in the counseling session from their site with proper internet connection. The dataset focused on recording both the client and counsellor speech and facial expressions while having an MI session for diet. The speakers both client and counselors are native Japanese speakers and the data is collected in Japanese. From the dataset description of [13], four clinicians who were psychotherapists and healthcare professionals with professional MI skills have participated as counselors. There are 12 to 13 sessions for each counselors with different clients for each session. In this dataset, there are 52 clients who are intended to change/improve their diet plans. There are 27 male and 25 female clients whose average age is around 35 years. This dataset has chosen 48 counseling sessions out of 52 because of technical issues encountered while recording. And the counseling session lasts for around 20 minutes each with an average of 21 min 57 sec. There are 23,643 utterances from both client (12,346) and the counselor (11,297) from 48 counseling sessions with an average of 2.33 sec for the client and 2.75 sec for the counselor. The dataset has language transcripts obtained from Google ASR for each

TABLE I  
PERCENTAGE OF EACH SPEECH CATEGORY IN CLIENT

Category	CT	ST	FN
Percentage	15	9	76

utterance of both the client and counselor. It also has utterance annotations using MISC version 2.1 [5]. More details of the dataset can be read from [13].

In this study, our task is to detect client's change talk (CT) or non-change talk (STFN) (non-change talk includes the Sustain Talk and Follow Neutral category utterances). Table I shows the percentage of client utterance categories.

## III. FEATURE EXTRACTION

### A. Language features:

From prior studies, RoBERTa outperformed BERT in classifying client utterances in MI sessions [8], [13], [14]. In our case, RoBERTa is used to acquire language embedding for each utterance. We used Hugging Face japanese-roberta-base pre-trained model, which results in obtaining a 768-dimensional vector for each utterance or token [15].

### B. Speech Features:

Psychologists and communication researchers have been investigating the ability of the voice to convey emotional signals. The measurement of emotion-differentiating parameters such as subglottal pressure, excitation strength at glottal closure instances and vocal fold vibration have recently been used to empirically document emotional cues conveyed in voice [16], [17]. Mel-Frequency Cepstral Coefficients (MFCC) are more important in paralinguistic voice analysis and to identify phonetic content than non-verbal voice attributes [18]. Formants have been found to be sensitive to a wide range of emotions and mental states, and they provide state-of-the-art cognitive load classification results [19].

Low *et.al* [20] has reported on psychiatric disorders using speech survey. This paper reported that source features, harmonic features, prosodic features, formants, spectral features, duration of speech, and speech rate are helpful in assessing the condition of client behavior. From the above discussion in choosing the set of features that could help in classifying clients' utterances in MI sessions, we decided to explore the OpenSmile emotion-based features to classify CT vs STFN. And also speech production perspective handcrafted features for analysing the client categories.

In this work, we have extracted some of the hand-crafted features as well as *OpenSmile* features for analyzing and classifying the client categories into CT or STFN. We have preprocessed the speech signal before we extract the features. We normalized each utterance and resampled it to 8KHz and later removed the non-speech segments from the utterances using *sox*. We extract MFCC features from 25ms window size with a shift of 10ms along with frame level log energy. We also computed fundamental frequency (F0) using wide band analysis of speech signal with temporal window size of 20ms with a shift of 10ms using PEFAC algorithm [21]. Along with this, we also computed first four formant frequencies

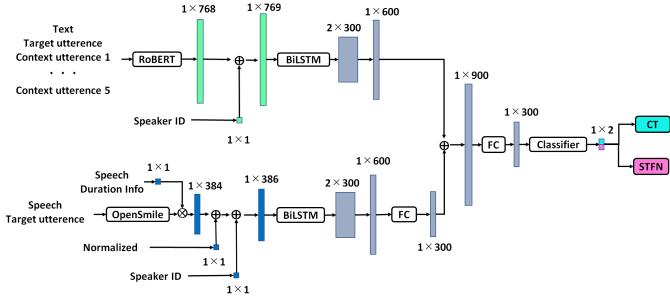


Fig. 1. BiLSTM RNN model for classifying Client's Change talk (CT) or non-change talk (STFN) using Speech and language information.

F1, F2, F3, F4 and bandwidth BW1, BW2, BW3, BW4 and amplitudes AMP1, AMP2, AMP3, AMP4 using wide band analysis by picking four peak locations from 12<sup>th</sup> order AR model. The excitation strength (es) features are computed at glottal closure instances, and the jitter, shimmer, and loudness of speech features are extracted from *OpenSmile* features. We computed the mean across all the frames for a given utterance for all these features. These feature values are used for performing the statistical test in detecting the CT and STFN.

Along with the above-handcrafted features, we have extracted 384-dimensional speech features from *OpenSmile* using IS09 config file [22]. This feature set consists of fundamental frequency (F0), voice probability, zero crossing rate, MFCC, and Root mean square energy statistics along with its deltas computed over an utterance. This feature set is used in training the BiLSTM model for classifying the CT and STFN.

#### IV. MODEL

In this section, we propose BiLSTM neural network models which use language and speech features for classifying client's CT and STFN. As detailed in Section III-A, we have computed the language feature embeddings for each utterance as an output from ROBERTa Japanese pre-trained model as an average of the last hidden state of all utterances. We computed the language embeddings for five preceding utterances as context to the target client utterance. Counselor utterances were included in the context utterances. We included a binary dimension at the beginning of every language embedding to indicate the speaker, which results in a 769-dimensional language embedding as an input to train the model. The BiLSTM model maintains long-range connections between context and the target utterance and produces a 300-dimensional vector from the hidden state. Being this process is bi-directional, we concatenated these two hidden state vectors into a single 600-dimensional tensor as language representation.

In contrast with the language contextual feature representation, we use only the target utterance for representing the speech feature for training the BiLSTM model. As detailed in Section III-B, we have concatenated the binary dimension at the beginning of every speech vector to indicate the speaker with 384-dimensional OpenSmile features along with the normalized speech length information. We have weighted each feature value with the original speech duration. More

Duration of Speech CT(1) and STFN(2)

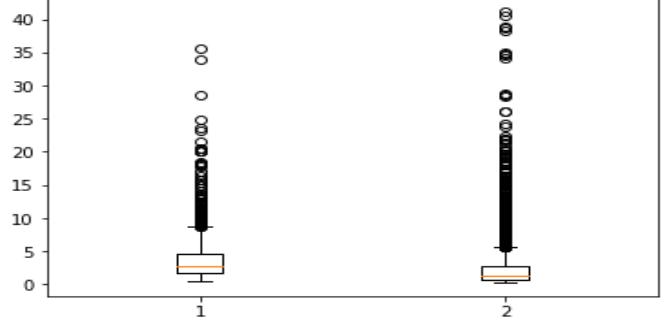


Fig. 2. Box plot for full dataset to show the significance of speech duration (y-label) in detecting the CT and STFN utterances

details regarding weighting the feature vector are detailed in the ablation study (please refer to section V-A). With the addition of these two vectors, the total dimension of the speech vector for training the BiLSTM model is 386. Similar to language representation, a 600-dimensional single tensor represent speech information as well. This speech feature tensor is fed to a fully connected (FC) layer to reduce the dimension to 300dimensions.

Both the language (600 dim) and speech (300 dim) representation are concatenated to obtain a single vector of 900 dimensions. This 900-dimensional multimodal tensor was fed to a fully connected layer to downsize the vector into a 300-dimensional tensor and it is applied to a classification layer. In order to build a two-class classification model (CT / STFN) cross-entropy loss is calculated over a softmax layer. The model is shown in Figure1. We hypothesized that language information is more beneficial for categorizing spoken utterances and that speech information complements language information, which is why we used more units for language representation than for speech representation [13].

#### V. EXPERIMENTS AND RESULTS

We conducted the experiments using the nutrition and fitness MI sessions dataset detailed in Section II. We use 81% of MI data for training (39 sessions), 10% for evaluating (5 sessions), and 8% of MI data for testing (4 sessions) the model. From TableI, we could observe the data for change talk (CT) is very less when compared with the non-change talk class (STFN). To address this data imbalance, we assign class weights for CT and STFN as 0.4 and 0.1 respectively to calculate the cross-entropy loss based on the ratio of CT to STFN. Adadelta has used an optimizer and the model is trained with 24 batch size for 300 epochs.

As detailed before in Section IV, we use two modalities of language and speech to classify the client's CT and STFN classes. We use five context utterances along with target utterances for language information whereas only target utterances for speech information. As it is clear from the dataset distribution that data is highly imbalanced between CT and STFN classes, we sub-sampled data for STFN class and tested the model. We also evaluated our models using all samples in

TABLE II  
ABLATION RESULTS IN CLASSIFYING CT VS STFN USING SPEECH FEATURES ALONG WITH LANGUAGE INFORMATION

Speech Model	CT			STFN			F1-Macro
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
Lang+Face [13]	<b>0.475(0.607)</b>	<b>0.804(0.804)</b>	<b>0.600(0.692)</b>	<b>0.950(0.913)</b>	<b>0.807(0.799)</b>	<b>0.873(0.852)</b>	<b>0.735(0.772)</b>
<i>Lang + S</i>	0.329(0.387)	0.882(0.882)	0.480(0.538)	0.960(0.943)	0.610(0.581)	0.746(0.719)	0.613(0.628)
<i>Lang + S<sub>wt</sub></i>	0.477(0.517)	0.680(0.712)	0.561(0.599)	0.923(0.906)	0.838(0.805)	0.879(0.853)	0.720(0.726)
<i>Lang + S<sub>dur</sub></i>	0.451(0.477)	0.758(0.758)	0.565(0.586)	0.938(0.916)	0.800(0.762)	0.864(0.832)	0.715(0.709)
<i>Lang + S<sub>wt+dur</sub></i>	<b>0.446(0.506)</b>	<b>0.804(0.804)</b>	<b>0.573(0.621)</b>	<b>0.948(0.776)</b>	<b>0.783(0.776)</b>	<b>0.858(0.847)</b>	<b>0.716(0.734)</b>

the test set without sub-sampling. The test set has only 16% CT utterances. Since our goal is to detect the CT from the client's utterances, we have used F1-score as our metric to evaluate the performance of our model.

**Lang+Speech Model:** We added speaker ID information to language representations before we fed them to the BiLSTM network. Likewise, we did the same for our speech representations in the initial approach using OpenSmile features without any weighting to speech features. *Lang+Speech* (*Lang+S*) model results in the F1-score of 0.48 and 0.746 for CT and STFN classes respectively for the full test set. And for the sub-sampled test set, it is 0.538 and 0.719 CT and STFN classes respectively. To investigate the impact of the duration of each utterance in detecting the CT, we conducted an ablation study on speech features. We performed the statistical significance test on the duration of speech on train data, test data and on entire full data set. We found that the duration of speech was highly significant at the 0.01 level ( $p < 0.01$ ) in t-test with high Cohen's D score for the three sets (train, test, full). For the full dataset, the *t-score* is 22.10, the *p-score* is  $4.2e^{-106}$  and Cohen's D score is 0.5643. The box plot for the full dataset speech duration is shown in Figure 2.

#### A. Ablation Study

TableI shows the percentage of client utterances by category, where FN is approximately 5 times greater than CT and 8 times greater than ST. This is because short acknowledging utterances such as "yes" are assigned to FN. This means that information with a short utterance duration is more likely to be FN.

From the statistical test, it is evident that the duration of speech is important in the speech representations, so we have weighted the speech features for the target utterance with the speech length. Since our goal is to detect CT, Tensors for utterances with short duration are weighted with smaller weight whereas the utterances with a long duration and a high probability of being CT are weighted heavily. We anticipated this weighting for speech features could improve the performance of the model.

Therefore, we created the following three types of speech features before feeding them to the BiLSTM model. (1) weighting the speech feature vector ( $S_{Wt}$ ) before input to BiLSTM by the utterance length, (2) adding the normalized (mean zero and unit variance) speech duration  $S_{dur}$  as a feature to the OpenSmile feature set, and (3) performing both of the above ( $S_{Wt}$  and  $S_{dur} \Rightarrow S_{wt+dur}$ ). We trained the classification model with language representations along with the above three sets speech representations. The results of ablation study are shown in Table II. The results indicate

TABLE III  
COMPARISON OF SPEECH FEATURES MEAN, T-STATISTICS, AND COHEN'S D-SCORE VALUES FOR CT VS STFN, AND ASTERISK (\*) DENOTES THE STATISTICAL SIGNIFICANCE WITH  $p < 0.05$

Features	CT mean	STFN mean	t - score	p-score	D - score
F0	173.25	171.28	1.70	8.95e-02	0.044
Loudness	1.858	2.128	-5.92	5.18e-08*	-1.209
Jitter	0.018	0.014	3.72	3.44e-04*	0.759
Shimmer	0.094	0.078	5.44	4.23e-07*	1.110
es	0.023	0.025	-5.85	5.18e-09*	-0.153
MFC0	-0.984	-1.841	5.13	2.94e-07*	0.134
F1	267.95	265.29	1.57	1.17e-01	0.041
F2	688.67	712.27	-7.23	5.16e-13*	-0.189
F3	1422.43	1440.43	-4.31	1.64e-05*	-0.113
F4	2176.53	2204.73	-6.55	6.13e-11*	-0.171
BW1	147.99	138.88	4.11	3.90e-05*	0.107
BW2	306.23	321.89	-5.33	9.95e-08*	-0.139
BW3	402.60	431.74	-7.23	5.22e-13*	-0.189
BW4	547.15	537.58	2.33	2.00e-02*	0.061
AMP1	34.31	34.88	-2.43	1.53e-02*	-0.063
AMP2	22.39	20.73	10.07	9.76e-24*	0.263
AMP3	9.96	8.64	10.17	3.57e-24*	0.265
AMP4	-0.07	-0.18	1.30	1.93e-01	0.034

that the *Lang + S<sub>wt+dur</sub>* model that consists of speech and language representation is the most effective. Numbers in parenthesis indicate the results with resampled test set. It is evident that weighting with speech duration and adding it as a feature is improving the performance of the model significantly, compared to *Lang + S* model. The performance (F1-score) of the proposed language and speech model (0.57) is comparable to the language and facial model (0.60) [13]. Besides the model building, we also performed statistical analysis for hand-crafted speech production perspective features to understand how the speech features are significant in the client's Change Talk.

#### B. Analysis of Audio Features

We perform statistical tests on our extracted handcrafted speech production perspective features using the complete dataset to test the potential of each feature for the classification of client categories. This analysis also aimed to gain insights about the speech features that could help in the client's change talk detection. We perform the statistical *t-test* on two independent samples CT and STFN with each feature set to find the statistical significance in classifying each class. We also computed Cohen's D score for validating the significance of each feature and these statistical scores are given in Table III. We observe that all the hand-crafted features have shown the statistical significance with  $p < 0.05$  in identifying the CT and STFN except for the features F0, F1, and amplitude 4.

From Table III, though F0 is not statistically significant for CT vs STFN, we observed that F0 is comparatively higher

when CT is compared with STFN. It is observed that F0 is higher for the highly motivated or taking steps to change the behavioral utterances of the client. Excitation strength(es) at glottal closure instances is the strength of the vocal folds closures when the speech is produced, which is higher for the speakers who are not willing to change their behaviors (STFN) when compared with CT speakers. This shows inverse relation with F0 – higher pitch lower the strength of excitation and vice-versa. Variations in the fundamental frequency are represented by jitter and shimmer. Jitter and shimmer are higher for CT compared with STFN utterances. Jitter and shimmer show a direct relation with F0 and are also statistically significant. MFCC0 has a higher absolute energy for STFN when compared with CT. Loudness of the speech is higher for STFN utterances than CT utterances. Absolute spectral magnitude energy is directly related to loudness. Absolute spectral magnitude energy is directly related to loudness, and these energy features are statistically significant. Formants and amplitudes have been attributed to vocal tract area shape and air pressure. Bandwidth can be attributed to vocal tract losses. Formant 1 has linguistic details where as higher formants are task-oriented. The formants and bandwidth frequencies are lower for CT utterances when compared with STFN, whereas, amplitudes are higher for CT when compared with STFN utterances. All four formants, bandwidth, and amplitude features show statistical significance except formant 1 and amplitude 4.

## VI. CONCLUSIONS

In this paper, we present a multimodal approach for detecting Change Talk or STFN using language and speech representations of client utterances from motivational interviewing sessions focused on nutritional diet and physical activities. We also presented the analysis of speech production-based features and their statistical significance in the client's Change Talk. We performed the ablation study to show the duration of speech is significant in detecting the CT and STFN utterances. The proposed setting using language and speech can classify CT and STFN utterances with an F1-score of 0.573 and 0.621 for CT for full test data and sub-sampled data respectively, which is par with the language and facial model.

As a future direction, the utterance classification task will be extended to include counselor utterances. If we can predict the categories of upcoming client and counselor utterances, it would be useful for providing guidance to MI trainees as well as for dialogue generation in MI.

## VII. ACKNOWLEDGMENTS

This work is carried out by first author while he was a PostDoc at IUI-Lab, Seikei University, Tokyo, Japan. Thanks to Prof. Yukiko Nakano for her valuable suggestions and discussions and Tomoya Tanaka for helping in some of the experiments.

## REFERENCES

- [1] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford press, 2012.
- [2] B. Lundahl and B. L. Burke, "The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses," *Journal of clinical psychology*, vol. 65, no. 11, pp. 1232–1245, 2009.
- [3] R. K. Martins and D. W. McNeil, "Review of motivational interviewing in promoting health behaviors," *Clinical psychology review*, vol. 29, no. 4, pp. 283–293, 2009.
- [4] C. Aswamenakul, L. Liu, K. B. Carey, J. Woolley, S. Scherer, and B. Borsari, "Multimodal analysis of client behavioral change coding in motivational interviewing," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 356–360.
- [5] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.
- [6] D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [7] J. Cao, M. Tanana, Z. E. Imel, E. Poitras, D. C. Atkins, and V. Srikanth, "Observing dialogue in therapy: Categorizing and forecasting behavioral codes," *arXiv preprint arXiv:1907.00326*, 2019.
- [8] L. Tavabi, T. Tran, K. Stefanov, B. Borsari, J. D. Woolley, S. Scherer, and M. Soleymani, "Analysis of behavior classification in motivational interviewing," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2021. NIH Public Access, 2021, p. 110.
- [9] B. Xiao, D. Can, J. Gibson, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Behavioral coding of therapist language in addiction counseling using recurrent neural networks," in *Interspeech*, 2016, pp. 908–912.
- [10] J. Gibson, D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "Attention networks for modeling behaviors in addiction counseling," in *Interspeech*, 2017, pp. 3251–3255.
- [11] M. Tanana, K. Hallgren, Z. Imel, D. Atkins, P. Smyth, and V. Srikanth, "Recursive neural networks for coding therapist and patient behavior in motivational interviewing," in *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 2015, pp. 71–79.
- [12] L. Tavabi, K. Stefanov, L. Zhang, B. Borsari, J. D. Woolley, S. Scherer, and M. Soleymani, "Multimodal automatic coding of client behavior in motivational interviewing," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 406–413.
- [13] Y. I. Nakano, E. Hirose, T. Sakato, S. Okada, and J.-C. Martin, "Detecting change talk in motivational interviewing using verbal and facial information," in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 2022, pp. 5–14.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [16] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.
- [17] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.
- [18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [19] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. Choi, "Formant frequencies under cognitive load: Effects and classification," *EURASIP journal on advances in signal processing*, vol. 2011, pp. 1–11, 2011.
- [20] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.
- [21] S. Gonzalez and M. Brookes, "Pefac-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.
- [22] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," 2009.