Speech-Integrated Modeling for Behavioral Coding in Counseling

Do June Min and Rada Mihalcea
Department of CSE
University of Michigan
Ann Arbor, Michigan, USA
{dojmin,mihalcea}@umich.edu

Verónica Pérez-Rosas
Department of Computer Science
Texas State University
San Marcos, Texas, USA
vrncapr@txstate.edu

Kenneth Resnicow
School of Public Health
University of Minnesota
Minneapolis, Minnesota, USA
kresnic@umn.edu

Abstract

Counseling and psychotherapy are predominantly conducted in spoken form, yet computational models to analyze these conversations often use only text transcripts as a main data source. This dependency overlooks important vocal cues such as tone, pitch, and prosody, hence restricting the models' capacity to fully capture the dynamics of counselor-client interactions. We introduce Motivational Interviewing with Speech Cues (MISQ), a simple yet effective framework for enhancing the analysis of psychotherapy conversations that uses a large language model, a lightweight adapter, and a speech encoder to directly integrate speech cues. Our experiments show that MISQ consistently outperforms (∼5% relative improvement) approaches that omit or use speech indirectly, highlighting the essential role of speech in accurately capturing counselor and client behaviors.

1 Introduction

Vocal cues such as tone, pitch, and prosody are crucial in counseling conversations, as they convey emotions and therapeutic intent (Miner et al., 2022; Wampold, 2012). However, many current computational psychotherapy systems are based solely on text transcriptions, often derived from automatic speech recognition (ASR) models. Consequently, these models ignore important paralinguistic information (Cao et al., 2019; Wu et al., 2022), limiting their ability to fully capture the complexity of therapeutic interactions.

The process of converting speech to text inherently leads to the loss of important vocal information (Cui et al., 2024) since most ASR systems prioritize linguistic content over non-verbal speech cues. As shown in Figure 1, even a correctly transcribed utterance (e.g. 0% Word Error Rate (WER)) cannot fully capture the speaker's tone or emotional delivery, both of which are essential to understand the therapy-client interaction.

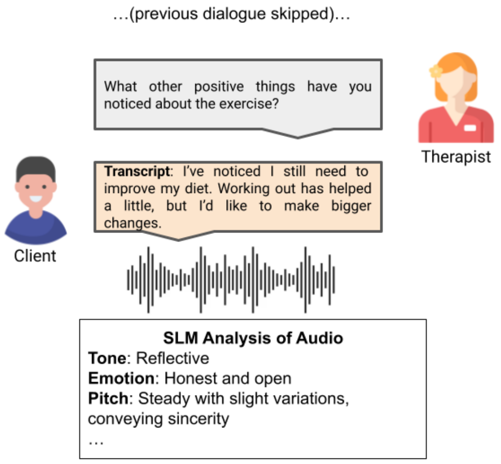

…(previous dialogue skipped)…
What other positive things have you noticed about the exercise?
Therapist
Transcript: I've noticed I still need to improve my diet. Working out has helped a little, but I'd like to make bigger changes.
Client
SLM Analysis of Audio
Tone: Reflective
Emotion: Honest and open
Pitch: Steady with slight variations, conveying sincerity
…

Figure 1: Example of therapist-client interaction in a counseling session with Speech-Language Model (SLM) analysis.

Recent work in automated MI analysis has predominantly relied on text transcriptions from ASR systems, but growing evidence highlights the value of multimodal signals, such as acoustic cues and prosodic features, in capturing nuanced client behaviors (Flemotomos et al., 2021; Nakano et al., 2022; Caponnetto et al., 2019).

In this work, we seek to bridge this gap by exploring how speech features from pre-trained speech encoders can be leveraged to improve the modeling of counseling conversations. We focus on behavioral coding and forecasting for counseling conversations conducted using Motivational Interviewing (MI), a counseling strategy that facilitates behavior change by resolving ambivalence through reflective listening and collaborative dialogue (Miller and Rose, 2009).

We introduce the Multimodal Integrated Model for Interactions in Counseling (MISQ), to combine a speech encoder with a large language model (LLM). We show that incorporating speech features significantly improves behavioral coding per-

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 152–158
Aug 25–27, 2025. ©2025 SIGDIAL

formance over text-only models, and that directly integrating raw speech consistently outperforms indirect methods, emphasizing its value in modeling counselor-client interactions.

2 Related Work

Speech Feature Fusion. Tavabi et al (Tavabi et al., 2020) employed pretrained VGGish embeddings to encode prosodic features, while Galland et al (Galland et al., 2023) leveraged raw speech signals via self-supervised encoders (e.g., HuBERT) to preserve paralinguistic information. Hossain et al (Hossain et al., 2024) uses Audio Spectrogram Transformer (AST) to extract audio embeddings (Gong et al., 2021). Notably, these works integrate speech through feature fusion, such as sequential or attention-guided fusion.

Dialogue Act Modeling and Behavioral Coding. Dialogue act modeling originates from the philosophical work of Austin on performative utterances and Searle's theory of illocutionary acts, which provided a framework for classifying speaker intentions into discrete types (Austin, 1962; Searle, 1976). This theoretical foundation enabled the development of computational systems capable of interpreting conversational intent. Behavioral coding in counseling is a specialized application of this paradigm, where dialogue acts are defined to capture specific therapeutic behaviors and client responses. For instance, well-established coding systems like the Motivational Interviewing Treatment Integrity (MITI) and the Motivational Interviewing Skill Code (MISC) provide structured taxonomies for these behaviors, enabling the systematic analysis of therapist proficiency and client engagement (Miller et al., 2003; Moyers et al., 2016).

Paralinguistic Analysis Prior work in paralinguistic analysis has explored how vocal features beyond lexical content contribute to understanding speaker affect and intent. Early studies examined prosodic cues such as pitch, energy, and speaking rate for emotion recognition in spontaneous speech (Devillers et al., 2003; Schuller et al., 2013). In clinical psychotherapy, variations in vocal prosody such as pitch, energy, and speech rate correlate with both therapeutic alliance and client engagement, and multimodal models have improved automated detection of counselor and client behavior (Bayerl et al., 2022; Flemotomos et al., 2021).

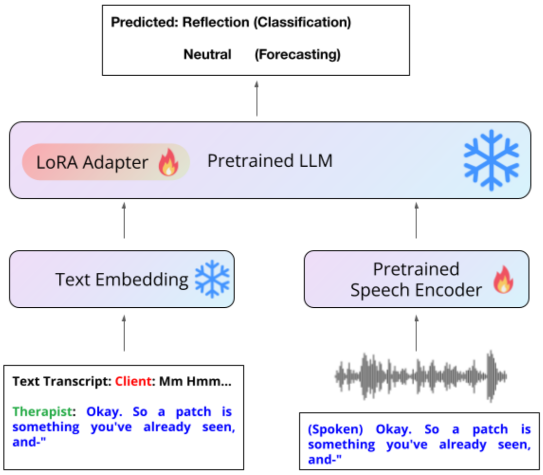

Predicted: Reflection (Classification)
Neutral (Forecasting)
LoRA Adapter
Pretrained LLM
Text Embedding
Pretrained Speech Encoder
Text Transcript: Client: Mm Hmm...
Therapist: Okay. So a patch is something you've already seen, and-"
(Spoken) Okay. So a patch is something you've already seen, and-"

Figure 2: **MISQ** consists of a trainable speech encoder integrated with a frozen text LLM, enhanced by a small, trainable LoRA adapter. Additionally, the original text embedding module is retained to process the text portion of the input.

3 Motivational Interviewing with Speech Cues (MISQ)

We build a multimodal language model, MISQ, by integrating speech information directly into our model using a speech encoder by processing text and speech features with the same underlying language model. This integration preserves the performance of LLM in text-based tasks (Kang et al., 2024; Chu et al., 2024) and allows the model to fully utilize the rich vocal information embedded in speech, such as tone and prosody, often lost in text-mediated approaches, such as paralinguistic captioning.

3.1 Model Overview

As shown in Figure 2, we combine a fine-tuned speech adapter, which consists of a LoRA adapter and a speech encoder, with a frozen text-based language model (LM). The speech encoder projects downsampled speech representations into the text embedding space, enabling seamless integration with the text model without the need to train a cross-modal language model from scratch, which would require significantly larger datasets. Our approach is similar to (Kang et al., 2024) but allows for a more efficient adaptation by incorporating parameter-efficient fine-tuning (PEFT).We make the code for our model public [1].

Speech Encoder. We extract an acoustic representation from raw audio using HuBERT (Hsu et al.,

[1] https://github.com/mindojune/speech_mi

2021), a pre-trained self-supervised speech encoder. We use a version fine-tuned on 960 hours of the LibriSpeech dataset (Panayotov et al., 2015), sampled at 16kHz to effectively capture diverse speech patterns. More specifically, we use the output from HuBERT's last hidden layer as input to the speech adapter. Since the sequence length of text tokens in a speech transcript is typically much shorter than the sequence length of its corresponding discretized speech, we apply an average pooling layer along the time dimension to downsample the speech representations. To align with the embedding dimension of the LLM, we further use a projection layer to upsample the downsampled speech embeddings (Kang and Roy, 2024).

Text Language Model & Adapter Module. For the text model backbone, we use `GeneZC/MiniChat-2-3B` (Zhang et al., 2023), a pre-trained language model with an embedding size of 3072. Rather than fully fine-tuning the model, we use PEFT to convert the frozen text-only model into a multimodal one, following strategies similar to (Lin et al., 2024). This approach ensures faster convergence and mitigates the risk of catastrophic forgetting and overfitting in the pre-trained LLM (Das et al., 2024). We integrate a LoRA lightweight adapter (Hu et al., 2022) to effectively inject speech representations into the model while preserving the language understanding capabilities of the original LLM.

Behavioral Modeling with Multimodal Embeddings. Our model (shown in Figure 2) starts by extracting and retaining the text embedding. and then processes speech through a speech adapter that extracts frame-level features via a pre-trained speech encoder. Next, it performs temporal downsampling to jointly process the text and speech embeddings. This process uses a LoRA adapter for speech-integrated behavior modeling, where the input to the LM is the concatenated sequence of text and speech features. The LM is trained to either predict or forecast the target behavior.

4 Experiments

4.1 Data

We use the publicly available AnnoMI dataset (Wu et al., 2023), which contains counseling conversations transcribed from 133 MI training demonstration videos. Since the conversations in the dataset depict effective and ineffective counseling skills, we focus on 110 conversations with effective counselors only.

All transcripts are gold-standard human transcripts (no ASR). Table 1 summarizes dataset composition and label distribution, and sample utterances are shown in 3.

Category	**Count**	**Percentage**
High-quality dialogues	110	–
High-quality utterances	8839	–
Therapist: Reflection	–	28%
Therapist: Question	–	28%
Therapist: Informational input	–	11%
Therapist: Other	–	33%
Client: Change talk	–	25%
Client: Neutral talk	–	64%
Client: Sustain talk	–	11%

Table 1: AnnoMI dataset composition and label distribution (from high-quality sessions) (Wu et al., 2023).

Ten MI practitioners annotated utterances. Annotation reliability metrics are in Table 2.

(Main) Therapist Behaviour	
Input	0.975
Reflection	0.991
Question	0.997
Other	0.996
Client Talk Type	
Change	0.916
Neutral	0.986
Sustain	0.890

Table 2: Inter-annotator agreement as intraclass correlation (Wu et al., 2023).

Data Processing. MISQ uses text and speech as input to model counselor-client interactions. Due to the considerable length of counseling sessions (often 10–20 minutes) and the inability of state-of-the-art models to handle such extended audio contexts effectively, we limit speech input to the target utterance only, rather than the full dialogue history (Wang et al., 2024). Note, however, that the full dialogue context is still provided to the model in text form to preserve conversational flow while leveraging the most relevant speech features to the target utterance.

4.2 Tasks & Evaluations

Behavioral coding of MI involves labeling conversational utterances with the corresponding behavior of the therapist and the client during counseling conversations. Labels function similarly to dialogue acts, capturing intent and communication strategies (Cao et al., 2019). We perform evaluations on two behavioral coding tasks: **categorization**, which predicts the behavioral code of the current utterance, and **forecasting**, which predicts the behavioral code of the next utterance. In our experiments, we predict and forecast seven therapist-client behaviors, including therapist behaviors: *reflection*, *question*, *input*, or *other*; and client change talk behaviors: *neutral* (unrelated to change), *sustain talk* (resistance to change), or *change talk* (motivation for change). Table 3 shows sample utterances labeled for therapist and client behavior.

Categorization seeks to recognize speakers' behaviors, while forecasting focuses on anticipating the next conversational behavior. Together, these tasks measure the model's understanding and prediction of counselor-client interactions. We measure model performance using accuracy and the F1 score to capture both overall correctness and the balance between precision and recall.

Prompt Excerpt	Label	Spkr
So, you've got this dream... to have a partner, a house, and a nice car.	Reflection	Therapist
Is this dream starting to– I am getting a picture of?	Question	Therapist
Yeah, it'd be good.	Neutral	Client
How offending fits in with that dream?	Question	Therapist
I guess it doesn't.	Change	Client

Table 3: Sample utterances from the AnnoMI Dataset

4.3 Baselines

During our experiments, we consider three main baselines.

Text Baseline. This model uses transcription only as input to the model.

Empty Sound. This ablated baseline retains the trained MISQ model, but receives a zero masked input for the speech modality. The model is used to evaluate the relevance of speech features in model performance.

Paralinguistic Captioning Model. Paralinguistic Captioning has been found to be effective in improving text models, as shown in DeSTA (Lu et al., 2024). We use a state-of-the-art paralinguistic captioning model (SLM), to extract paralinguistic features such as tone, pitch, and prosody from conversations and convert them into textual descriptions (e.g., "client pauses, then speaks with rising pitch"). These textual descriptions are then concatenated with the ASR transcripts and provided as input to the model.

We omit a full speech-language model baseline to focus on a lightweight integration for low-resource counseling settings. Paralinguistic captioning, which uses a full speech-language model to generate textual descriptions of vocal cues, captures key speech features indirectly. Our parameter-efficient approach yields greater behavioral-coding improvements while demanding significantly less computational power and data.

4.4 Training Setup

Our training objective is next-token prediction, optimized using standard cross-entropy loss. During training, we use the AdamW optimizer with betas=(0.9, 0.999). We set the learning rate at 1e-4, with batch size of 32. We train for a maximum of 10 epochs, with an early stopping criterion set to stop training if the validation loss does not improve for three consecutive epochs.

5 Results & Analyses

The results shown in Table 4 reveal that the models that incorporate speech features outperform the text-only baseline for categorization and forecasting tasks. MISQ consistently achieves the best results. This highlights the importance of directly incorporating vocal cues in modeling counselor-client interactions.

Categorization. For categorization, MISQ shows higher general accuracy than both the text baseline and the paralinguistic description models (71.67% vs 68.51% and 69.31%). Client and therapist accuracies also improved, with notable gains in therapist performance (77.88%). Macro F1 scores further confirm the MISQ's balanced performance across classes.

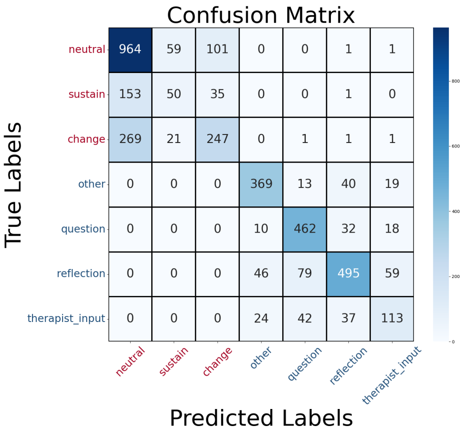
In addition, we observe that the performance of the MISQ with empty sound inputs drops sharply. This significant decline emphasizes that for MISQ, speech information is crucial to understanding and predicting participant behavior, and text alone does not provide adequate information for behavioral modeling. However, in Figure 3, we also note that some confusion remains between MISQ's prediction of `neutral`, `sustain talk`, and `change talk` possibly due to inherent ambiguities in client speech.

Forecasting. In forecasting, MISQ again leads in performance (52.94% accuracy), surpassing the

	Categorization				Forecasting			
Metrics	Text Baseline	Paralinguistic Captioning	Empty Sound	MISQ	Text Baseline	Paralinguistic Captioning	Empty Sound	MISQ
Overall Acc.	0.6851	0.6931	0.6878	**0.7167**	0.5018	0.5071	0.4949	**0.5294**
Client Acc.	0.6268	0.6315	0.6314	**0.6562**	0.6042	0.6084	0.6064	**0.6405**
Therapist Acc.	0.7449	0.7562	0.7440	**0.7788**	0.4000	0.4063	0.3824	**0.4188**
Macro F1	0.5930	0.6246	0.4849	**0.6473**	0.3132	**0.3781**	0.2936	0.3076

Table 4: Performance comparison between Text Baseline, Paralinguistic Captioning, Empty Sound, and MISQ models on Categorization and Forecasting tasks. The Empty Sound model is identical to MISQ but receives no audio input, isolating the impact of the speech modality. Bolded values indicate the best-performing model for each metric.

Text Baseline and Paralinguistic Captioning models. Client's accuracy increased to 64.05%, and therapist's accuracy improved modestly. Despite the inherent difficulty of forecasting, direct speech integration clearly provides helpful cues to anticipate conversational flow.



	neutral	sustain	change	other	question	reflection	therapist_input
neutral	964	59	101	0	0	1	1
sustain	153	50	35	0	0	1	0
change	269	21	247	0	1	1	1
other	0	0	0	369	13	40	19
question	0	0	0	10	462	32	18
reflection	0	0	0	46	79	495	59
therapist_input	0	0	0	24	42	37	113

Figure 3: Confusion matrix illustrating model's predictions against true labels for the **categorization** task.

Robustness to Noise. We also evaluate our model in challenging acoustic environments to assess whether it is capable of handling degraded audio signals. We simulate noisy acoustic environments by adding Gaussian noise at varying signal-to-noise ratio (SNR) levels to the raw speech input. Our experiments show a strong robustness to the noise injected. As illustrated in Figure 4, MISQ shows minimal performance degradation in the categorization task at different noise levels. In contrast, the model experiences a more pronounced performance decline for the forecasting task in the presence of noise. This degradation is most severe at the highest noise level, whereas at lower levels, the impact on performance is relatively minor. We hypothesize that paralinguistic features in the speech signal may be more resistant to noise compared to purely linguistic and semantic cues.

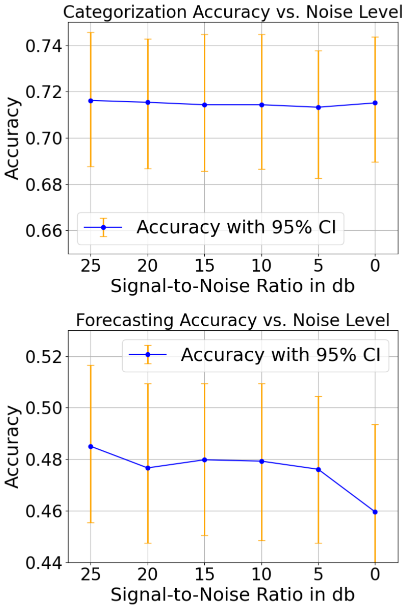


Figure 4: Model performance versus noise level for categorization (**top**) and forecasting (**bottom**) tasks. Dots show mean accuracy, and error bars represent 95% bootstrap confidence intervals (n=1000).

6 Conclusion

In this work, we presented a simple yet effective framework that integrates speech features to a pretrained text LM in order to improve the modeling of MI counseling sessions. By combining an LLM with a lightweight adapter and a speech encoder, our speech integration approach, MISQ, achieves a general accuracy improvement of up to 5.5% over baselines of text alone (text baseline) and conditioned text (paralinguistic captioning), demonstrating its effectiveness in behavioral coding tasks. These results underscore the critical role of raw speech features in capturing the nuanced dynamics of counselor-client interactions. Furthermore, our study highlights the importance of using direct speech integration for more accurate and insightful computational models in psychotherapy.

References

John Langshaw Austin. 1962. *How to do things with words*. William James Lectures. Oxford University Press.

Sebastian Bayerl, Gabriel Roccabruna, Shammur Chowdhury, Tommaso Ciulli, Morena Danieli, Korbinian Riedhammer, and Giuseppe Riccardi. 2022. What can speech and language tell us about the working alliance in psychotherapy. pages 2443–2447.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Pasquale Caponnetto, Jennifer DiPiazza, Giorgio Cappello, Shirin Demma, Marilena Maglia, and Riccardo Polosa. 2019. Multimodal smoking cessation in a real-life setting: Combining motivational interviewing with official therapy and reduced risk products. *Tobacco Use Insights*, 12:1179173X1987843.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *ArXiv*, abs/2410.03751.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2024. Speechverse: A large-scale generalizable audio language model. *Preprint*, arXiv:2405.08295.

Laurence Devillers, Ioana Vasilescu, and Catherine Mathon. 2003. Prosodic cues for perceptual emotion detection in task-oriented human–human corpus. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pages 1505–1508, Barcelona, Spain. Universitat Autònoma de Barcelona.

Nikolaos Flemotomos, Victor Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek Caperton, James Gibson, Michael Tanana, Panayiotis Georgiou, Jake Van Epps, Sarah Lord, Tad Hirsch, Zac Imel, David Atkins, and Shrikanth Narayanan. 2021. Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, 54.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2023. Seeing and hearing what has not been said: A multimodal client behavior classifier in motivational interviewing with interpretable fusion. *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.

Sayed Muddashir Hossain, Jan Alexandersson, and Philipp Müller. 2024. M3TCM: Multi-modal multi-task context model for utterance classification in motivational interviews. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10872–10879, Torino, Italia. ELRA and ICCL.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and Ozlem Kalinli. 2024. Frozen large language models can perceive paralinguistic aspects of speech. *Preprint*, arXiv:2410.01162.

Wonjune Kang and Deb Roy. 2024. Prompting large language models with audio for general-purpose speech summarization. *Preprint*, arXiv:2406.05968.

Tzu-Han Lin, How-Shing Wang, Hao-Yung Weng, Kuang-Chen Peng, Zih-Ching Chen, and Hung yi Lee. 2024. Peft for speech: Unveiling optimal placement, merging strategies, and ensemble techniques. *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 705–709.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. 2024. Desta: Enhancing speech language models through descriptive speech-text alignment. *Preprint*, arXiv:2406.18871.

William R. Miller, Theresa B. Moyers, Denise Ernst, and Paul Amrhein. 2003. *Manual for the Motivational Interviewing Skill Code (MISC)*. Center on Alcoholism, Substance Abuse and Addictions, The University of New Mexico, Albuquerque, NM. Published November 6, 2003.

William R Miller and Gary S Rose. 2009. Toward a theory of motivational interviewing. *American Psychologist*, 64(6):527–537.

Adam S Miner, Scott L Fleming, Albert Haque, Jason A Fries, Tim Althoff, Denise E Wilfley, W. Stewart Agras, Arnold Milstein, Jeff Hancock, Steven M Ash, Shannon Wiltsey Stirman, Bruce A. Arnow, and Nigam H. Shah. 2022. Uncovering the linguistic characteristics of psychotherapy: a computational approach to measure therapist language timing, responsiveness, and consistency. *medRxiv*.

Theresa Moyers, Lauren Rowell, Jennifer Manuel, Denise Ernst, and Jon Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65.

Yukiko I. Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting change talk in motivational interviewing using verbal and facial information. *Proceedings of the 2022 International Conference on Multimodal Interaction*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech Language*, 27(1):4–39. Special issue on Paralinguistics in Naturalistic Speech and Language.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):1–23.

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, page 406–413, New York, NY, USA. Association for Computing Machinery.

Bruce Wampold. 2012. Humanism as a common factor in psychotherapy. *Psychotherapy (Chicago, Ill.)*, 49:445–9.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2024. Retrieval augmented end-to-end spoken dialog models. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12056–12060.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Zixiu "Alex" Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Towards automated counselling decision-making: Remarks on therapist action forecasting on the annomi dataset. In *Interspeech*.

Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. 2023. Towards the law of capacity gap in distilling language models.