Adil Soubki, John Murzaku, Peter Zeng, and Owen Rambow. 2025. Synthetic Audio Helps for Cognitive State Tasks. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 1701–1708, Albuquerque, New Mexico. Association for Computational Linguistics.

Synthetic Audio Helps for Cognitive State Tasks

Adil Soubki[*♦♠], **John Murzaku**[*♦♠], **Peter Zeng**[♦♠], **Owen Rambow**[♣♠]
[♦]Department of Computer Science [♣]Department of Linguistics
[♠]Institute for Advanced Computational Science
Stony Brook University
{asoubki,jmurzaku}@cs.stonybrook.edu

Abstract

Automatically recognizing a human's complete cognitive state from text is a difficult task; from text, a model has to recognize a combination of concepts including belief, emotion, common ground, sentiment, and intention. Humans do not only track and update cognitive state from the meaning of words and sentences, but also from paralinguistic cues such as prosody. The NLP community has broadly focused on text-only approaches to cognitive state tasks, but audio can provide vital missing information. We posit that text-to-speech (TTS) models learn to track aspects of cognitive state in order to produce naturalistic audio, and that the signal audio models implicitly identify is orthogonal to the information that language models exploit. We present Synthetic Audio Data fine-tuning (SAD), a framework where we show that seven tasks related to cognitive state modeling benefit from multimodal training on both text and zero-shot synthetic audio data from an off-the-shelf TTS system. We show an improvement over the text-only modality when adding synthetic audio data to text-only corpora. Furthermore, on tasks and corpora that do contain gold audio, we show our SAD framework achieves competitive performance using text and synthetic audio compared to text and gold audio.

1 Introduction

A significant amount of work in NLP focuses on tasks that involve extracting information about the cognitive states of human entities from text. This includes predicting beliefs (or "event factuality") (Saurí and Pustejovsky, 2009), recognizing emotions (Canales and Martínez-Barco, 2014), recognizing sentiment (Wiebe, 1990), tracking common ground (Markowska et al., 2023), predicting conversation success (Zhang et al., 2018), identifying intentions (Colombo et al., 2020), among others.

Prior work has shown that audio signals, when available, improve performance for a number of tasks involving cognitive states (Murzaku et al., 2024; Zhao et al., 2022; Nojavanasghari et al., 2016). Meanwhile, text-to-speech (TTS) systems have improved rapidly over the past several years – particularly when it comes to synthesizing more naturalistic audio. Part of what models must learn in order to generate realistic speech involves saying words in a way that matches the cognitive states those words reveal. In this paper we investigate two research questions related to these observations.

RQ1: In the event that one has a task with human audio, how does using synthetic (TTS generated) audio compare to the gold-standard (human) audio? **RQ2:** Can synthetic audio help even for tasks which never had human audio to begin with?

We hypothesize that synthetic audio will perform worse than human audio (**RQ1**) but better than text-only (no audio). Furthermore, the aspects of cognitive state that TTS models learn to predict in order to produce naturalistic speech will provide orthogonal signal to the patterns text-based language models pick up on and improve performance, even on datasets for which human audio is not available (**RQ2**).

Our main contribution is to present SAD, a multimodal synthetic audio data framework that boosts performance on cognitive state tasks that do not contain audio, or offers competitive performance on tasks that do.

The paper is organized as follows. A survey of previous work is provided in Section 2. We summarize the SAD framework in Section 3 and present our experiments on SAD in Section 4. We conclude and provide a discussion and implications of our novel framework in Section 5.

We emphasize that this paper does not introduce new machine learning architectures; instead, we show that synthetic audio data through our SAD framework can lead to improved performance with-

[*]Denotes equal contribution

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 1701–1708

out the necessity of introducing new, more complex architectures. We view our framework as a generalizable solution; it can profit and adapt from advances in language models, TTS models, and multimodal models. We release our framework and models on GitHub[1].

2 Related Work

To the best of our knowledge, we are the first to present a multimodal text and audio framework with synthetic audio data from TTS systems for cognitive state tasks. However, regarding experiments with audio signal, there has been previous work on multimodal (text and audio) and unimodal (audio only) models for corpora in emotion, belief, deception, and sentiment.

Multimodal There has been some work in fusing text and audio features for cognitive state tasks, specifically in emotion and belief. In emotion, Zhao et al. (2024) present a novel architecture containing a refined attention mechanism, a novel perception unit aligning the emotion frame to the global audio context, and a new convolution procedure to effectively fuse audio and text features. Kyung et al. (2024) fine-tune BERT (Devlin et al., 2019) and fuse with ASR features derived from speech, achieving state-of-the-art results on the multimodal emotion corpus IEMOCAP (Busso et al., 2008), which we also test our SAD framework on. In the multimodal belief prediction task, Murzaku et al. (2024) were the first to show that fusing text with audio features helps, achieving state-of-the-art results on the CB-Prosody corpus (Mahler et al., 2020).

Regarding deception, there has been previous work on acoustic and lexical approaches. Testing on the CXD corpus (Levitan et al., 2015), Mendels et al. (2017) show that a multimodal architecture boosts performance compared to a unimodal text-only approach.

Audio Only There has also been work focusing on the audio-only modality for cognitive state tasks, but considerably less than multimodal. For the deception detection task, (Levitan et al., 2018; Chen et al., 2020b; Levitan and Hirschberg, 2022) focus on training classical machine learning methods with acoustic and prosodic features. Regarding emotion detection, Pepino et al. (2021) were the first to fine-tune a pre-trained speech model for emotion detection, specifically Wav2Vec2.0 (Baevski et al., 2020).

[1] https://github.com/adil-soubki/sad-training

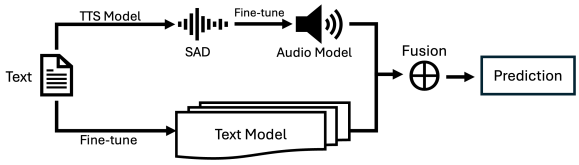

Figure 1: Overview of the SAD framework, beginning with a text input. We then perform zero-shot TTS on the text to get audio and then fine-tune an audio model. In parallel, we fine-tune a text model. We then fuse the features from both modalities to get a final prediction.

3 SAD Overview

Text We perform all fine-tuning experiments with BERT, specifically bert-base-uncased. We closely mirror the experimental setup of previous work on text and audio multimodal cognitive state architectures which find BERT to be the top performing text encoder (Zhao et al., 2022; Murzaku et al., 2024).

TTS: OpenAI We generate synthetic audio data using the OpenAI TTS API (OpenAI, 2023). We specifically generate all of our data using the Alloy voice and the tts-1-hd model which is optimized for audio quality. While we also performed experiments with different voices, we fix to one voice (Alloy) due to API costs.

TTS: Open Source To emphasize the generalizability and viability of our SAD framework, we also perform experiments with an open source TTS model, specifically MatchaTTS (Mehta et al., 2024). We choose MatchaTTS due to its minimal memory requirements, competitive speed on long utterances, and top performance in terms of mean opinion score in human evaluations. We also experimented with other TTS APIs such as (ElevenLabs, 2023), but chose to stick with OpenAI API and Matcha as they were the most cost effective and yielded the highest quality audio.

Audio We follow Murzaku et al. (2024) and use the pre-trained Whisper model (Radford et al., 2023) as our audio encoder, specifically whisper-base. For all Whisper experiments, we pad all audio clips to the maximum 30 second limit.

Multimodal We combine all the previously listed individual components into a unified multimodal architecture. We show this architecture in

Figure 1. We perform both early fusion and late fusion experiments, similar to (Murzaku et al., 2024; Zhao et al., 2022; Nojavanasghari et al., 2016) : in our late fusion model, BERT and Whisper are fine-tuned separately with their representations max pooled and concatenated after passing through individual regression heads. In the early fusion model, the final hidden representations are max pooled and concatenated before being jointly fine-tuned and passed through a shared regression head. We emphasize the generalizability of the SAD framework: each individual component can be replaced with fine-tuned or task-specific models.

4 Experiments

4.1 Tasks

A broad overview of the tasks we test SAD on is shown in Table 1. We specifically test on four types of tasks: control tasks which are not about the writer's cognitive state, and for which we hypothesize that synthetic audio data will not improve compared to text only; sentiment; belief; and emotion. We also show the included modalities from each corpus; only three corpora contain both text and audio (SWBD-S, CB-Prosody, and IEMOCAP). We describe each task and its corpora in detail.

Control Tasks Our control tasks include three tasks chosen from SuperGLUE (Wang et al., 2019): BoolQ (Clark et al., 2019) (question answering), WiC (Pilehvar and Camacho-Collados, 2019) (word sense disambiguation), and WSC (Levesque et al., 2012) (pronoun resolution and common sense reasoning). We choose these three tasks as they do not explicitly model cognitive state, expecting SAD to not improve performance.

Sentiment We test on two corpora that annotate for sentiment. The Switchboard Sentiment (SWBD-S) corpus (Chen et al., 2020a) annotates segments of Switchboard (Godfrey et al., 1992) for sentiment averaged among three annotators, resulting in a continuous sentiment value from [-1, 1]. This corpus contains gold audio; we therefore compare our proposed SAD framework to multimodal experiments with the gold audio. We also test on the IMBD corpus (Maas et al., 2011) which is a standard benchmark for author sentiment in NLP.

Belief The term "belief" refers to how committed is the author or speaker to the truth of a proposition. Murzaku et al. (2024) were the first to show that multimodal architectures with specifically text and speech signal boost performance on belief tasks compared to standard text only approaches. We therefore test whether SAD can help on three belief corpora: CB-Prosody (Mahler et al., 2020) which Murzaku et al. (2024) tested on, CB (De Marneffe et al., 2019), and FactBank (Saurí and Pustejovsky, 2009). In CB, annotators were given transcripts of speaker utterances or written text and asked to evaluate the level of certainty (or belief) the speaker appears to have regarding the truth of the proposition, specifically the content of the complement clause. Expanding on CB, CB-Prosody, which contains gold audio data, annotates 350 Switchboard examples present in CB, but annotators *only* heard the audio clips rather than reading transcripts. Both of these corpora have continuous annotations; specifically, belief values fall from a continuous range of [-3,3].

We also choose FactBank (Saurí and Pustejovsky, 2009) as a benchmark corpus for SAD, as it is one of the first carefully constructed datasets for belief prediction. We specifically use the author only examples and the split from Murzaku et al. (2022). FactBank uses categorial labels.

Emotion We include two corpora for emotion recognition, since emotion is frequently expressed in audio. In order to synthesize audio which is naturalistic, TTS systems must learn to recognize utterances which reveal emotions, especially those of higher intensity, and vary their output accordingly. As such we expect there to be some signal from TTS for emotion recognition both when the dataset contains gold audio (IEMOCAP, (Busso et al., 2008)) and when it does not (GoEmotions, (Demszky et al., 2020)).

4.2 Experimental Setup

Training For all experiments, we add either a classification or regression head depending on the task. All experiments are trained for 10 epochs; no hyperparameter tuning is performed. We use a learning rate of 2e-5 and batch size of one.

Data To save API costs, we randomly downsample datasets until the cost of generating audio using OpenAI's TTS system costs $10 USD per corpus. We use the same subcorpus when doing comparison experiments with Matcha-TTS. We mark the datasets that were downsampled in Table 1. We provide more details on the data and experiments in Appendix B.

Type	Task	Gold		Size	Metric	Text	Audio			Multimodal		
		Text	Audio				Gold	Matcha	OpenAI	Gold†	Matcha	OpenAI†
Control	BoolQ 🌱✂	✓	✗	509⌄⌄	Acc ↑	60.0	-	**69.7**	69.4	-	65.3	67.1
	WIC 🌱	✓	✗	6,066	Acc ↑	**57.3**	-	49.1	47.4	-	59.6	57.1
	WSC 🌱	✓	✗	658	Acc ↑	**63.5**	-	**63.5**	60.6	-	**63.5**	**63.5**
Sentiment	SWBD-S 🗁	✓	✓	2,856⌄⌄	MAE ↓	0.339	0.465	0.461	0.462	**0.334**	0.341	**0.334**
	IMDB 🌱✂	✓	✗	372⌄⌄	Acc ↑	89.5	-	63.5	58.5	-	88.3	**89.7**
Belief	CB-Prosody 🗁	✓	✓	334	MAE ↓	0.693	1.083	0.931	0.906	**0.665**	0.699	0.668
	CB 🌱	✓	✗	500	MAE ↓	0.785	-	1.189	1.154	-	0.756	**0.741**
	FactBank 🌱	✓	✗	7,540	F1 ↑	74.9	-	68.7	66.3	-	**76.0**	**76.0**
Emotion	IEMOCAP 🌱	✓	✓	7,529	F1 ↑	56.6	55.5	51.7	52.2	**63.4**	57.6	59.3
	GoEmotions 🌱	✓	✗	4,753⌄⌄	F1 ↑	51.4	-	38.3	36.2	-	52.7	**53.1**

Table 1: Overview of cognitive state tasks, including gold data availability, corpus size, evaluation metrics, and results for various modalities. Size represents the number of samples in the dataset (⌄⌄ indicates the dataset was down-sampled due to cost). We report Acc ↑ (Accuracy ↑), F1 ↑ (F1 ↑ score), and MAE ↓ (Mean Absolute Error) as metrics depending on the corpus. If the dataset did not have a canonical split, metrics are averaged over five folds (indicated by 🗁). Otherwise the metrics are averaged over three seeds (indicated by 🌱). If the audio data exceeded Whisper's 30 second context, it was truncated when training (datasets significantly affected are indicated by ✂). A binomial test ($\pi_0 = 0.5$) is used to determine if the frequency models outperform the text-only baseline on non-control tasks is significant († indicates $p < 0.05$).

We note that we experimented with better or larger text models, specifically RoBERTa (Liu, 2019) and Flan-T5 (Chung et al., 2024), but did not notice an improvement in our multimodal results over BERT as a text encoder (which is in line with findings by Murzaku et al. (2024)).

4.3 Evaluation

Table 1 shows the metrics we evaluate each corpus on. We perform regression on three corpora, since they contain continuous values as annotations: SWBD-S, CB-Prosody, and CB. We therefore evaluate these three corpora on mean absolute error (MAE). For the rest of the corpora, which contain categorical labels, we perform an accuracy or F-measure evaluation. We indicate the metric for each corpus in the table, and whether more or less is better. If the TTS audio improves model performance, we would expect models which use that signal to outperform those which do not more frequently than random across the non-control datasets. We therefore use a binomial test ($\pi_0 = 0.5$) to determine the significance of this frequency. Where relevant, we also note differences within tasks using independent paired t-tests.

4.4 Results

We first examine each set of experiments by task category and then conclude with a discussion of our research questions.

Control Since we hypothesize that the additional signal that TTS models provide comes from their ability to represent cognitive states through the way they synthesize audio, the control tasks were chosen as tasks which do not reflect the cognitive state of the speaker. We expect multimodal training to not affect the results here and this is indeed the result for WIC and WSC. However, BoolQ performs best when given *only* the audio data, which is surprising. This is likely due to BERT's poor handling of such long sequences of context before the question. Since the audio model can only handle 30 seconds of context, we truncate the remaining audio which seems to have helped for some reason. We expect that further tuning of the input representations would bring this result in line with the other control tasks.

Sentiment Since sentiment analysis requires predicting the author's feelings (either in general or towards a particular aspect), we expect audio to improve our results. After averaging across all folds and seeds we do see this but the effect size is somewhat smaller than expected. For SWBD-S, the multimodal models just barely edge out the text-only variant, with gold audio and OpenAI synthetic audio both achieving the best MAE of 0.334. The SAD model using OpenAI TTS similarly performs best by a narrow margin on IMDB, scoring an accuracy of 89.7%. As the gold audio in the SWBD-S corpus doesn't help much for sentiment,

it makes sense that synthetic audio does not help either: apparently sentiment is mainly conveyed lexically in our sentiment corpora.

Belief The belief expressed by an utterance can entirely change based on the way the utterance is said (e.g., **John* said it's true* vs. *John said, it's *true**). It is possible that TTS systems have useful priors in such cases. Across all three belief datasets, after averaging metrics over all folds and seeds, the best performing models are the multimodal variants. For CB-Prosody, the gold audio results in a 4% decrease in MAE and the OpenAI TTS model trails closely behind with a 3.6% reduction in MAE. For the text-only CB, both Matcha and OpenAI synthetic audio models improve over the text-only baseline to achieve a 3.7% and 5.6% decrease in MAE, respectively. FactBank also sees a 4.4% F1 error decrease over the text-only model for both Matcha and OpenAI SAD models.

Emotion In order to synthesize audio which is naturalistic, TTS systems must learn to recognize utterances which reveal emotions, especially those of higher intensity, and vary their output accordingly. As such we expect there to be some signal from TTS for emotion recognition in both the condition where the dataset contains gold audio (IEMOCAP) and when it does not (GoEmotions). For IEMOCAP we see a 6.8 point improvement in F1 over the text-only model when using gold audio. The synthetic audio models also outperform text-only with OpenAI and Matcha seeing a 1.0 and 1.7 point improvement, respectively. For GoEmotions, improvements of similar scale over the text-only are observed with Matcha showing a 1.3 point boost and OpenAI showing a 1.7 point boost. These differences are significant ($p < 0.05$) using independent two-sample t-tests.

Matcha vs. OpenAI Though the audio-only models typically performed far worse than the multimodal variants, one might expect the TTS system which performed the best on audio-only fine-tuning (typically, Matcha) to perform the best in SAD fine-tuning. Our results do not support such a trend. TTS generations from OpenAI matched or outperformed Matcha in multimodal fine-tuning for all seven non-control tasks, though often by a small margin. In both cases early fusion tended to perform best.

RQ1: How does SAD compare to using gold audio? Generally, our experiments found SAD to perform worse than using gold audio but better than no audio at all. For SWBD-S and CB-Prosody, SAD (using OpenAI TTS) matched or nearly matched the performance of gold audio; for IEMOCAP, there was a sizable degradation (4.1 points F1) between the best SAD model and gold performance as discussed above. If we consider just the noncontrol datasets with gold audio, a binomial test (using all experiments with either multiple seeds or folds, as shown in Table 1) indicates that the frequency with which the OpenAI multimodal models outperform their text-only counterparts is significant ($p < 0.05$). This is not the case for Matcha. The same analysis across gold and SAD models also indicates gold models perform better than SAD models significantly ($p < 0.05$) often. When comparing performance of fusion models which receive gold audio against those that receive synthetic audio, IEMOCAP is the only case where a significant difference ($p < 0.05$) is observed using an independent two-sample t-test. This is also the only case where gold audio significantly ($p < 0.05$) improves the unimodal models. In other words, emotion tasks were particularly sensitive to audio quality.

RQ2: Does SAD help data with no gold audio? Yes. Our experiments show that datasets that never had audio to begin with also see an improvement in performance. If we consider just the non-control datasets, binomial testing finds OpenAI (but not Matcha) audio to improve performance significantly ($p < 0.05$) often, similar to the case for datasets with gold audio. Performing the same analysis for the control datasets, as expected, does not find significance. The absence of (unexplained) improvements in the control tasks, suggests that TTS models contain latent signals for cognitive states.

5 Conclusion

We have introduced a new approach to prediction tasks about the cognitive state of a speaker or writer. We show that using TTS to create synthetic audio helps across seven tasks when used in conjunction with text, compared to using only text. While the effect sizes are currently small, performance gains will likely grow as TTS systems improve over the coming years. Our research suggests that exploiting additional modalities, even when synthetic, may be a useful strategy in NLP tasks if we have reason to believe that the additional modality may carry orthogonal signal for the task.

Limitations

Cognitive State Focus Our work focuses on a targeted subset of tasks within cognitive state modeling. We understand that, while our SAD framework supports our hypothesis that TTS models capture cognitive state features, the idea may not be generalizable to broader NLP tasks. We leave this to future work and intend to explore broader tasks.

TTS Model Choice A large portion of SAD primarily focuses on using closed API TTS models. We therefore understand that some areas may lack model details and implementation details. We however will release all scripts and details for generating our data through the API.

Ethical Considerations

As with other work on cognitive states, we risk the misinterpretation that AI models may be anthromorphized as having near-human level cognition. We stress that our work shows that our framework can help text-only models when presented with natural sounding audio, but does not give AI models full cognitive state understanding or cognition.

We note that our paper is foundational research and we are not tied to any direct applications.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions); by funding from the Defense Advanced Research Projects Agency (DARPA) under the CCU project (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034); as well as by the Intelligence Advanced Research Projects Activity (IARPA) under the HIATUS program (contract 2022-22072200005). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, DARPA, or IARPA.

We thank both the Institute for Advanced Computational Science (IACS) and the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources and API fees needed for this work. These resources were made possible by NSF grant No. 1531492 (SeaWulf HPC cluster maintained by Research Computing and Cyberinfrastructure) and NSF grant No. 1919752 (Major Research Infrastructure program), respectively.

We thank our ARR reviewers, whose comments have contributed to improving the paper.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador. Association for Computational Linguistics.

Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. 2020a. A large scale speech sentiment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6549–6555, Marseille, France. European Language Resources Association.

Xi Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic, and Julia Hirschberg. 2020b. Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, 8:199–214.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7594–7601.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

ElevenLabs. 2023. Elevenlabs.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jehyun Kyung, Serin Heo, and Joon-Hyuk Chang. 2024. Enhancing multimodal emotion recognition through asr error compensation and llm fine-tuning. In *Interspeech 2024*, pages 4683–4687.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 1–8.

Sarah Ita Levitan and Julia Hirschberg. 2022. Believe it or not: Acoustic-prosodic cues to trust and mistrust in spoken dialogue. In *Speech Prosody*, volume 2022, pages 610–614.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues. In *Proc. Interspeech 2018*, pages 416–420.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Taylor Mahler, Marie-Catherine de Marneffe, and Catherine Lai. 2020. The prosody of presupposition projection in naturally-occurring utterances. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 20–37.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*.

Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg. 2017. Hybrid acoustic-lexical deep learning approach for deception detection. In *Interspeech*.

John Murzaku, Adil Soubki, and Owen Rambow. 2024. Multimodal belief prediction. In *Interspeech 2024*, pages 1075–1079.

John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. Re-examining FactBank: Predicting the author's presentation of factuality. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.

OpenAI. 2023. Text-to-speech.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Janyce M. Wiebe. 1990. Identifying subjective characters in narrative. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Zihan Zhao, Yanfeng Wang, and Yu Wang. 2022. Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In *Interspeech 2022*, pages 4725–4729.

Ziping Zhao, Tian Gao, Haishuai Wang, and Björn Schuller. 2024. Mfdr: Multiple-stage fusion and dynamically refined network for multimodal emotion recognition. In *Interspeech 2024*, pages 3719–3723.

A Experiment Details

All experiments besides our OpenAI experiments used our employer's GPU cluster. We performed experiments on a Tesla V100-SXM2 GPU. Compute jobs typically ranged from 5 minutes for zero-shot TTS generation to 6 hours for multimodal fine-tuning. The text model used was `google-bert/bert-base-uncased` (110M params) and the audio model used was `openai/whisper-base` (72.6M params). We fine-tune all models for a fixed 10 epochs and report the relevant metrics at the last epoch. All experiments use a learning rate of 2e-5 and a batch size of 1. We do not perform any hyperparameter tuning or hyperparameter searches. We use the mean squared error (MSE) loss function for regression tasks and cross-entropy loss for classification tasks. We pad text to BERT's maximum sequence length of 512 and audio clips to Whisper's maximum sequence length of 30 seconds. We checked training loss curves to ensure that models were converging.

We report the average over three seeds (42, 0, 1) for corpora with an established train/test/dev split. For other corpora, we perform five-fold cross-validation and report the average over all five folds.

B Data Processing

Synthetic Audio Data Prior to generating synthetic audio, for text with context longer than the target text for which a task is to be performed on, we shorten the text to just the target span. This is because prosody of generated audio tends to degrade as the lengths of the generations increase. To save costs, we down-sample datasets such that the cost of generating audio using OpenAI's TTS system costs $10 USD, and match that data when generating audios with Matcha-TTS.

FactBank For datasets such as FactBank (Saurí and Pustejovsky, 2009), which annotate single event tokens, we extract the syntactic span using spaCy (Honnibal and Montani, 2017). We create a custom `head2span` module which takes the event head word and returns the syntactic span.

CB The CB dataset (De Marneffe et al., 2019) contains three sentences: two previous sentences of context, and the target sentence where the matrix clause is annotated. For all experiments, we use only the last, or the target sentence.

SWBD-S The SWBD-S corpus contains three annotations for sentiment averaged among three annotators resulting in a continuous sentiment value from [-1, 1]. We manually process this using our own scripts.