Interspeech 2024
1-5 September 2024, Kos, Greece

Multimodal Belief Prediction

John Murzaku[*1,3], *Adil Soubki*[*1,3], *Owen Rambow*[2,3]

Stony Brook University, USA
[1]Department of Computer Science
[2]Department of Linguistics
[3]Institute for Advanced Computational Science
{jmurzaku,asoubki}@cs.stonybrook.edu, owen.rambow@stonybrook.edu

Abstract

Recognizing a speaker's level of commitment to a belief is a difficult task; humans do not only interpret the meaning of the words in context, but also understand cues from intonation and other aspects of the audio signal. Many papers and corpora in the NLP community have approached the belief prediction task using text-only approaches. We are the first to frame and present results on the multimodal belief prediction task. We use the CB-Prosody corpus (CBP), containing aligned text and audio with speaker belief annotations. We first report baselines and significant features using acoustic-prosodic features and traditional machine learning methods. We then present text and audio baselines for the CBP corpus fine-tuning on BERT and Whisper respectively. Finally, we present our multimodal architecture which fine-tunes on BERT and Whisper and uses multiple fusion methods, improving on both modalities alone.
Index Terms: multimodal belief prediction, speech belief prediction, computational paralinguistics

1. Introduction

A critical aspect of both written and spoken conversation is identifying the extent to which discourse participants commit themselves to propositions. However, as is frequently observed in sociolinguistics [1], humans rarely assert commitments exactly and employ all manner of tools to convey this tactfully. This makes the task of extracting and predicting how committed speakers are to propositions (i.e., their beliefs) more challenging than it might first seem, especially if we are given only written text. Previous work has shown that acoustic-prosodic speech features can distinguish deception (no commitment to a belief) and truthfulness (commitment to a belief), or trust and mistrust [2, 3, 4], which emphasizes the need to jointly model text and audio features.

Over the last 15 years, the belief prediction task (also called event factuality prediction) has attracted significant interest from the NLP community, resulting in several corpora [5, 6, 7, 8, 9, 10, 11]. However, most of this work has only focused on predicting belief from the text modality. A notable exception is the CommitmentBank (CB) corpus [10] which annotates both text and audio corpora, with 350 utterances coming from Switchboard [12]. However, the annotators only read the dialog transcription and audio was discarded. Expanding on CB, the authors present the CB-Prosody (CBP) corpus [13], which annotates speaker commitment/belief on the same 350 Switchboard examples, but instead annotators only hear the audio. To our knowledge, we are the first to present a multimodal belief prediction task and find that incorporating audio signals results in a 12.7% relative reduction in mean absolute error (MAE) and 6.4% relative increase in Pearson correlation when compared to text-only approaches.[1]

Our main contributions are summarized as follows:

1. We are the first to present and report results on the multimodal belief prediction task and do so using the CBP corpus. Furthermore, we contribute missing information to the CBP corpus (audio start and end times) and will release all data.
2. We perform an acoustic-prosodic analysis of CBP using openSMILE features [14], and find 25 significant features. We train the acoustic-prosodic features using a XGBoost-RF model and are the first to report audio-only results with these features.
3. We present audio-only and text-only baselines with pretrained models, specifically fine-tuning BERT [15] and Whisper [16]. We find that Whisper significantly outperforms our acoustic-prosodic XGBoost-RF baseline.
4. We present a multimodal belief architecture, jointly fine-tuning BERT and Whisper. We compare our multimodal results to our text-only and audio-only baselines, and ultimately find our multimodal architecture provides a significant improvement in belief detection. Furthermore, we investigate early and late fusion methods, finding that late fusion of features performs better than early fusion.

The paper is organized as follows. Section 2 describes our data, specifically the CBP corpus and how it frames the belief prediction task. In Section 3, we describe our acoustic-prosodic and neural approaches to the task. Results for all modalities and experimental settings are reported in Section 4. We conclude in Section 5 and present ideas for future work in multimodal belief detection.

2. CB-Prosody Corpus

We use the CB-Prosody corpus (CBP), which annotates on top of the manually transcribed English Switchboard corpus utterances. Annotators were given speaker utterances, specifically the content of the complement clause, and asked to evaluate the level of certainty the speaker appears to have regarding the truth of the proposition. Specifically, they were asked to rate the level of commitment on a 7-point Likert scale where explicit labels were displayed for three classes: the speaker is certain that it is true, the speaker is not certain whether it is true or false, and the speaker is certain that it is false. The intermediate values contained partial commitments (i.e. the speaker is weakly or strongly possibly certain/not certain). The average among at least eight annotators' judgements is then reported as the final annotation value.

[*]Denotes equal contribution to this paper.

[1]https://github.com/cogstates/multimodal-belief

10.21437/Interspeech.2024-2103

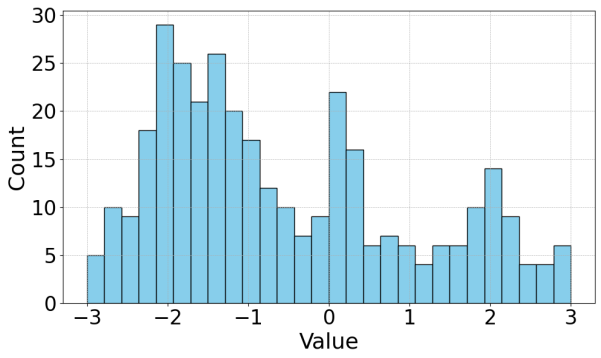

Figure 1: *A histogram of the average commitment values among annotators for all 338 utterances in CB Prosody.*

We manually extract the clip start and end times from Switchboard, convert the files to wav format, and then perform upsampling to 16kHz on the audio clips. We discard 12 examples that were either misaligned with the gold Switchboard files or had significant interruptions (not a continuous utterance from the speaker) and end up with 338 total examples.

Summary statistics for the audio segments and word counts in each example are shown in Table 1. All annotations are provided on a continuous scale. To report dataset statistics, we discretize the continuous annotations into 28 bins. Figure 1 shows our bins and the corresponding count for each bin. There is a clear imbalance in the label distribution with the majority of labels falling in between the bin values $[-2, -1]$. There are also two noticeable peaks around bin values 0 and 2. The fewest label counts are near bin values -3, 1, and 3.

Table 1: *CBP summary statistics*

	Min	Max	Mean	Median	Stdv
Duration (s)	1.01	16.44	4.61	4.04	2.73
Words	5.00	47.00	16.49	15.00	7.48

To illustrate the difference between the text-only speaker commitment annotations on CB and the audio speaker commitment annotations on CBP, we consider the following example from the corpus: *He didn't think it looked predawn to him.* In this example, annotators are asked to annotate how certain the speaker is that *it looked predawn to him*. We note that the speaker is not the referent of *he* and *him* – this corpus does not annotate the cognitive state of people mentioned in text, only of the speaker. Here, the text-only CB corpus annotates this as 0.55, suggesting a slight inclination towards non-commitment (i.e., 0) regarding the statement about it looking predawn. To an annotator who only sees the text, this leaning towards no commitment is due to the lexical content and syntactic structure of the sentence, where the negation (*He didn't think*) could introduce uncertainty about the speaker's commitment.

On the other hand, the CBP corpus annotates this as -2.16, indicating a strong commitment to the falseness of the statement (i.e., committing to it *not looking predawn*). In this example's audio, there is a heavy emphasis on *he didn't*, suggesting that the speaker trusts the "he", and this he believes that it doesn't look predawn to him, meaning a strong leaning towards committing to the falseness of the statement.

Table 2: *Statistically significant top 11 openSMILE IS09 features*

Rank	Feature	Avg r
1-5	MFCC	0.158
6-7	F0	0.124
8-10	MFCC	0.120
11	voiceProb	0.112

3. Our Approaches

We frame belief prediction as a regression task, following previous literature [9, 17, 18, 19]. Given an utterance and a sequence of corresponding features $S = [f_1, f_2, ..., f_n]$, we wish to produce a value $\hat{y} \in [-3, 3]$ that is as close to the annotation value y as possible.

3.1. Audio Feature Based

For our feature based approach, we extract features from our audio segments using openSMILE [14]. We hypothesize that there is an overlap in significant features between our task and deception detection. Hence, we specifically use the 384 feature Interspeech 2009 (IS09) emotion challenge set [20], which previous work on deception detection found to be useful [2, 21].

We also perform a simple feature selection approach and aim to answer the following question: **Are there OpenSmile features (from the IS09 set) that are correlated with the continuous CBP annotation values?** To answer this, we calculate the Pearson ρ among our features and the belief values and choose the statistically significant features (any correlation with p-value < 0.05). We find 25 statistically significant features and provide a summary of our top 11 features (measured by p-value) in Table 2. The most represented features are MFCC measures/coefficients, showing in 8 of our top 11 features. We also find F0 and voiceProb measures/coefficients as significant features. Interestingly, our significant features are directly in line with significant features found in deception detection such as MFCCs and voiceProb [22], which points to the similarities between our tasks.

3.2. Fine-tuning

The models we fine-tuned for each modality are as follows:
Text – We use the pre-trained BERT model [15] for our fine-tuning experiments, specifically *bert-base-uncased* which contains 110M parameters. We also performed fine-tuning experiments with RoBERTa [23] and got near identical or worse results which is in line with previous work on CB [18], so we therefore use BERT instead. We follow a standard fine-tuning approach and add a regression head to which we feed the final BERT representations to predict a continuous value.
Audio – We use the pre-trained Whisper model [16] for our fine-tuning experiments, specifically *whisper-base*, which contains 72.6M parameters. We note that we also tried experiments with Hubert [24] , but achieved better performance with Whisper. We try both mean and max pooling of our audio features after fine-tuning and before our regression head, finding that max pooling slightly outperforms mean pooling. We therefore perform all experiments with max pooling. Similarly to our text only experiments, we perform a standard fine-tuning approach, only adding a regression head to predict a continuous value, to which we feed the final Whisper representations to.
Multimodal – Our multimodal architecture fine-tunes on both

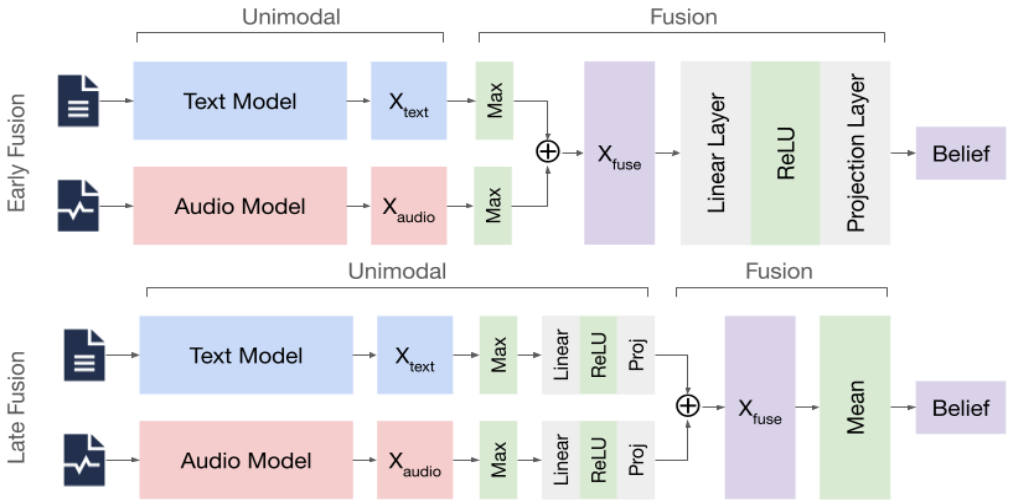

Figure 2: *Block diagram describing the early (above) and late (below) fusion models.*

BERT and Whisper. A natural research question that arises for our task is which multimodal feature fusion method performs best: early feature fusion or late feature fusion. In the multimodal emotion recognition task, which uses similar methods of jointly fine-tuning of text and audio pre-trained models, it has been shown that late feature fusion works better than early feature fusion [25]. A similar result was obtained for multimodal persuasion prediction [26]. We describe our feature fusion architectures in the next subsections.

All of our fine-tuning experiments are run on a single Tesla V100 or H100 GPU. Each fold took on average about 15 minutes to fine-tune, totaling about 75 minutes per experiment configuration.

3.3. Multimodal Early Fusion

Our early fusion model is shown at the top of Figure 2. We first jointly fine-tune BERT and Whisper on our text and audio inputs, respectively. We then pass the final hidden representations from each modality through a max pooling layer and concatenate them. Finally, we pass the concatenated pooled representations to our regression head consisting of a linear layer, ReLU, and output a final continuous value. When incorporating openSMILE features, they are concatenated after pooling.

3.4. Multimodal Late Fusion

Our late fusion model is shown at the bottom of Figure 2. We first fine-tune BERT and Whisper on our text and audio inputs, respectively. We do not concatenate them as we do in our early fusion model. Rather, we max pool both representations individually and then pass them through their individual regression heads. We then finally concatenate the two representations and do mean averaging to get our final continuous value. When using openSMILE features, they are passed through their own regression head.

4. Experiments and Discussion

4.1. Data

We use all 338 examples of the CBP corpus described in Section 2. To address the label imbalance shown in Figure 1, we perform a stratified five-fold cross validation using a fixed seed for all experiments. We do not hold out an established validation set because of the small size of the corpus, and therefore do not perform any hyperparameter tuning.

Table 3: *XGB-RF results with all IS09 features and correlated IS09 features.*

Features	MAE ↓	Pearson ↑
All IS09 Feats	$1.25_{\pm 0.09}$	$0.18_{\pm 0.10}$
Correlated Feats	$1.22_{\pm 0.08}$	$0.24_{\pm 0.04}$

4.2. Training and Evaluation Methods

For our tasks, we either directly perform regression or add a regression head to our pre-trained models. Our approaches are as follows:

XGB-RF For our audio only feature based approach with openSMILE, we train an XGBoost-Random Forest (XGB-RF) regression model. We perform z-score normalization on all of our features. We do not perform hyperparameter tuning and use the standard XGB-RF hyperparameters of scikit-learn [27].

Fine-tuning We fine-tune all models for a fixed 10 epochs and report the Pearson correlation and mean absolute error 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. We note that we tried the Huber loss function which was used in previous work on text-only belief prediction [18, 17], but ultimately found that this performed worse compared to MSE. For our BERT experiments, we set the maximum sequence length to 256. For our Whisper experiments, we pad all audio clips to 30 seconds and perform normalization.

Evaluation We evaluate on Pearson correlation (ρ) and mean absolute error (MAE). Previous work on text-only belief prediction found that Pearson ρ is useful in biased test sets since it assesses how well a model captures variability in the gold data and MAE captures absolute fit [19]. We report the average Pearson ρ and MAE across all five folds and their corresponding standard deviations. We perform significance testing on both metrics using paired bootstrap resampling, which has shown to be notably effective and robust for small test sets [28].

Table 4: *Mean MAE and Pearson ρ results for the 5-fold CV experiments comparing modality and fusion strategies. Significant ($p < 0.05$) improvements over the text-only BERT model are indicated by* †.

Text	Audio	Fusion	MAE ↓	Pearson↑
BERT	-	-	0.71 $\pm$0.08	0.78 $\pm$0.03
-	Whisper	-	1.02 $\pm$0.14	0.53 $\pm$0.13
-	OpenSMILE	-	1.22 $\pm$0.08	0.24 $\pm$0.04
-	Whisper + OS	Early	1.27 $\pm$0.07	0.14 $\pm$0.10
-	Whisper + OS	Late	1.10 $\pm$0.15	0.41 $\pm$0.19
BERT	Whisper	Early	0.66† $\pm$0.09	0.81† $\pm$0.04
BERT	Whisper	Late	0.62† $\pm$0.08	0.83† $\pm$0.04
BERT	OpenSMILE	Early	1.19 $\pm$0.06	0.37 $\pm$0.06
BERT	OpenSMILE	Late	0.68 $\pm$0.10	0.80 $\pm$0.04
BERT	Whisper + OS	Early	0.92 $\pm$0.05	0.62 $\pm$0.10
BERT	Whisper + OS	Late	0.66† $\pm$0.08	0.79 $\pm$0.03

4.3. Unimodal and Multimodal Models

We first train our XGB-RF regressor for our acoustic-prosodic feature baseline on all the IS09 features and then only the significant IS09 features described in Section 3. Results are shown in Table 3. We find that using correlated features results in a 2.4% decrease in MAE and a 33.3% increase in Pearson ρ.

The results for all model configurations are shown in Table 4. The best unimodal model is text-only BERT which achieves a mean MAE of 0.71 and Pearson ρ of 0.78. While the remaining four audio-only unimodal models tested consistently perform worse than the text-only model, the multimodal fusion models decidedly outperform unimodal approaches. Our best result comes from the late fusion of BERT and Whisper and significantly ($p < 0.05$) improves over unimodal BERT achieving a mean MAE of 0.62 and Pearson ρ of 0.83. These are 12.7% and 6.4% relative improvements in MAE and Pearson, respectively. In fact, our three best results come from multimodal fusion approaches and all result in significant ($p < 0.05$) improvements in MAE demonstrating the value of incorporating audio signals for belief prediction.

4.4. Impact of Fusion Strategy

Theoretical work comparing early and late fusion methods predicts that, with enough training data, early fusion will perform best but, in the absence of sufficient data, late fusion will do better [29]. Comparing our models by fusion strategy we see that, as theory predicts given our small dataset, the late fusion models tend to perform better regardless of the features being incorporated. While we do not find significant differences (MAE $p = 0.11$, Pearson $p = 0.17$) between early and late fusion of BERT and Whisper, we do find significant differences when we compare all early fusion results with all late fusion results. This pattern has also been observed for related tasks including persuasiveness prediction ($n \approx 205$) [26] and emotion recognition ($n \approx 4,400$) [25]. Our results support existing literature both theoretically and empirically.

4.5. Comparison of Whisper and openSMILE

It is not entirely clear that the pre-training objective used for Whisper will produce representations which capture information similar to the acoustic-prosodic features from openSMILE. It is possible that Whisper and openSMILE provide some orthogonal information. When comparing late fusion of BERT and Whisper with the late fusion of BERT and openSMILE we find that Whisper significantly ($p < 0.05$) outperforms openSMILE as a source of audio features by MAE (though not by Pearson). When comparing to the late fusion of all three feature sources, we do not find the increase compared to our BERT text-only model to be significant for Pearson ρ. However, the MAE of fusing all three features was lower and significant. This suggests that Whisper is the best source of audio features, though openSMILE performs competitively and may be preferable in applications with limited computational resources, or in situations in which we need to use interpretable features.

Incorporating openSMILE using early fusion performed much worse than by using late fusion. While this was expected, as discussed above, it was more substantial than observed with Whisper. We suspect it is possible to reduce this gap with additional changes to the architecture or normalization techniques but leave this to future work as our primary objective is to demonstrate the importance of audio features for this task.

4.6. Comparison to Other Work

While there are no comparable results on CBP, text-only belief prediction is a well established task in NLP. For the entire CB corpus, containing 556 utterances, state-of-the-art results achieve an MAE of 0.56. Given the disparity in training examples, our best performing model does well with an MAE of 0.62 (and 0.71 text-only). Performance for the text-only task varies among other belief corpora with state-of-the-art results ranging from 0.33 MAE on FactBank to 0.73 on UDS-IH2 [9, 17].

Though none of these make use of audio signals, our results are roughly in line with them which suggests that our approach achieves strong performance even in this broader framing.

5. Conclusion and Future Work

In this paper, we are the first to present audio-only and multimodal results on the belief detection task. We first analyze acoustic-prosodic features extracted from openSMILE, and provide an audio only baseline. We find similarities and overlaps with the deception detection task. Next, we provide unimodal and multimodal baselines fine-tuning pre-trained speech and text models. For our audio only baseline, we find that fine-tuning Whisper outperforms our acoustic-prosodic feature-based approach with openSMILE. We explore early and late fusion techniques, and find both early fusion and late fusion help compared to text-only. We achieve a new state-of-the-art result by fine-tuning BERT and Whisper, and using late fusion. Our state-of-the-art system achieves a 12.7% decrease in MAE and a 6.4% increase in Pearson ρ compared to our text-only baseline.

Our results provide a baseline for the multimodal belief prediction task and show that an audio signal does help text-only models. With these insights, in future work we intend to apply multimodal belief models on other corpora such as the Common Ground corpus [11] which annotates beliefs and common grounds on the CALLHOME corpus [30]. We intend to do multi-task learning experiments on belief and deception and see if they can help each other.

6. References

[1] P. Brown and S. C. Levinson, *Politeness: Some universals in language usage*. Cambridge university press, 1987, vol. 4.

[2] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues," in *Proc. Interspeech 2018*, 2018, pp. 416–420.

[3] S. I. Levitan and J. Hirschberg, "Believe it or not: Acoustic-prosodic cues to trust and mistrust in spoken dialogue," in *Speech Prosody*, vol. 2022, 2022, pp. 610–614.

[4] X. Chen, S. Ita Levitan, M. Levine, M. Mandic, and J. Hirschberg, "Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 199–214, 2020.

[5] R. Saurí and J. Pustejovsky, "Factbank: a corpus annotated with event factuality," *Language resources and evaluation*, vol. 43, pp. 227–268, 2009.

[6] V. Prabhakaran, T. By, J. Hirschberg, O. Rambow, S. Shaikh, T. Strzalkowski, J. Tracey, M. Arrigo, R. Basu, M. Clark, A. Dalton, M. Diab, L. Guthrie, A. Prokofieva, S. Strassel, G. Werner, Y. Wilks, and J. Wiebe, "A new dataset and evaluation for belief/factuality," in *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 2015, pp. 82–91.

[7] A.-L. Minard, M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, and C. van Son, "MEANTIME, the NewsReader multilingual event and time corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4417–4422.

[8] A. S. White, R. Rudinger, K. Rawlins, and B. Van Durme, "Lexicosyntactic inference in neural models," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4717–4724.

[9] R. Rudinger, A. S. White, and B. Van Durme, "Neural models of factuality," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 731–744.

[10] M.-C. De Marneffe, M. Simons, and J. Tonhauser, "The commitmentbank: Investigating projection in naturally occurring discourse," in *proceedings of Sinn und Bedeutung*, vol. 23, no. 2, 2019, pp. 107–124.

[11] M. Markowska, M. Taghizadeh, A. Soubki, S. Mirroshandel, and O. Rambow, "Finding common ground: Annotating and predicting common ground in spoken conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Dec. 2023, pp. 8221–8233.

[12] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–374, 2000.

[13] T. Mahler, M.-C. de Marneffe, and C. Lai, "The prosody of presupposition projection in naturally-occurring utterances," in *Proceedings of Sinn und Bedeutung*, vol. 24, no. 2, 2020, pp. 20–37.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10, 2010, p. 1459–1462.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[17] J. Murzaku, P. Zeng, M. Markowska, and O. Rambow, "Re-examining FactBank: Predicting the author's presentation of factuality," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 786–796.

[18] N. Jiang and M.-C. de Marneffe, "He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics," *Transactions of the Association for Computational Linguistics*, pp. 1081–1097, 2021.

[19] G. Stanovsky, J. Eckle-Kohler, Y. Puzikov, I. Dagan, and I. Gurevych, "Integrating deep linguistic features in factuality prediction over unified datasets," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2017, pp. 352–357.

[20] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech*, 2009.

[21] S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. Schuller, "Is Deception Emotional? An Emotion-Driven Predictive Approach," in *Proc. Interspeech 2016*, 2016, pp. 2011–2015.

[22] G. Mendels, S. I. Levitan, K.-Z. Lee, and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection," in *Interspeech*, 2017.

[23] 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.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021.

[25] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition," in *Proc. Interspeech 2022*, 2022, pp. 4725–4729.

[26] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.

[27] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[28] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 388–395.

[29] L. M. Pereira, A. Salazar, and L. Vergara, "On comparing early and late fusion methods," in *International Work-Conference on Artificial Neural Networks*. Springer, 2023, pp. 365–378.

[30] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech LDC97S42," 1997, Web Download. Philadelphia: Linguistic Data Consortium.

7. 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) and by funding from the Defense Advanced Research Projects Agency (DARPA) under the INCAS (HR01121C0186, No. HR001120C0037, and PR No. HR0011154158) and the CCU (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034) programs. 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 or DARPA.

We thank the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources needed for this work made possible by NSF grant No. 1919752 (Major Research Infrastructure program).