Interspeech 2024
1-5 September 2024, Kos, Greece

This Paper Had the Smartest Reviewers – Flattery Detection Utilising an Audio-Textual Transformer-Based Approach

Lukas Christ[1], *Shahin Amiriparian*[2], *Friederike Hawighorst*[3], *Ann-Kathrin Schill*[3], *Angelo Boutalikakis*[3], *Lorenz Graf-Vlachy*[4], *Andreas König*[3], *Björn W Schuller*[1,2,5]

[1]University of Augsburg, Germany [2]CHI, TU Munich, Germany [3]University of Passau, Germany
[4]TU Dortmund, Germany [5]GLAM, Imperial College London, UK

lukas1.christ@uni-a.de

Abstract

Flattery is an important aspect of human communication that facilitates social bonding, shapes perceptions, and influences behaviour through strategic compliments and praise, leveraging the power of speech to build rapport effectively. Its automatic detection can thus enhance the naturalness of human-AI interactions. To meet this need, we present a novel audio textual dataset comprising 20 hours of speech and train machine learning models for automatic flattery detection. In particular, we employ pretrained AST, Wav2Vec2, and Whisper models for the speech modality, and Whisper TTS models combined with a RoBERTa text classifier for the textual modality. Subsequently, we build a multimodal classifier by combining text and audio representations. Evaluation on unseen test data demonstrates promising results, with Unweighted Average Recall scores reaching 82.46% in audio-only experiments, $85.97\,\%$ in text-only experiments, and $87.16\,\%$ using a multimodal approach.

Index Terms: flattery, speech classification, human-AI interaction, computational paralinguistics, Transformers

1. Introduction

Flattery is a pervasive social influencing behaviour in which one individual (i.e., the flatterer) provides deliberate appreciation toward another individual or group (i.e., the flattered) [1]. Specifically, flattery accentuates the positive qualities of the flattered with the objective of interpersonal attractiveness and ultimately winning favour [2, 3]. While flattery is rather difficult to detect for the flattered, it is easier to be detected by bystanders [4]. As such, flattery is one of the most commonplace social influencing behaviours that many individuals employ, receive, or witness on a daily basis [5, 4]. Research across various disciplines has extensively explored the potency of flattery, particularly within organisational contexts. Kumar and Beyerlein [2], for instance, devised the Measure of Ingratiatory Behaviors in Organizational Settings (MIBOS) scale to unveil how employees adeptly employ flattery as an upward influencing technique towards immediate superiors [2]. Likewise, [6] demonstrated that job interviewees may employ flattery toward their job interviewers to enhance these job interviewers' evaluations. In the broader management context, it was shown that journalists exhibit more positive coverage of a firm and its CEO when flattered by the CEO and capital markets analysts also respond to CEO flattery by issuing more positive firm ratings [7, 8].

Motivated by the relevance of flattery in interpersonal communication, we investigate its automatic detection via Machine Learning (ML) methods, utilizing speech data. Such a methodology may be applicable in, e. g., human-computer interaction, communication training [9], and computational psychometrics [10]. Automatic analysis of speech for understanding different facets of human communication is an active research field, with Speech Emotion Recognition (SER) arguably being the most prominent task (e. g., [11, 12, 13]). Other examples include the detection of humour (e. g., [14, 15, 16]), sarcasm [17], interpersonal adaptation [18], and defensive communication [19]. Oftentimes, it is beneficial to also consider linguistic information, i. e., transcripts of the speech samples, when addressing such tasks [20, 21, 22]. To the best of our knowledge, our work is the first to consider the task of detecting flattery from speech. Our contributions are as follows: (i) we introduce a novel dataset for flattery detection from speech; (ii) we provide ML approaches for automatic flattery detection.

2. Dataset

Considering the large body of literature on flattery in the management context, we analyse flattery as applied by business analysts. We obtain a dataset of 2159 dyads of analyst questions and CEO answers within earnings calls of large, publicly listed U. S. hard- and software as well as pharmaceutical firms from 2013 to 2018, which were transcribed and published by Seeking Alpha. Drawing from prior research [23, 24], we systematically develop a reliable, context-sensitive, content-analytical measure for detecting and assessing analyst flattery. A detailed account of our annotation guidelines is provided in [25]. A team of three expert human annotators label the dataset for instances of flattery on a span-level, i. e., flattery is identified in subsentential word sequences. An example is considered flattery if and only if all three annotators agree on it.

In our machine learning experiments, the problem of flattery detection is framed as a sentence-level binary classification task here. The transcripts are split into sentences via the pySBD [26] tool. We then project the subsentential annotations to the sentence level by treating a sentence as a positive example of flattery if it contains a passage identified as flattery by the annotators. To build the audio dataset, we first reconstruct word-level timestamps from the speech data using the Montreal Forced Aligner (MFA) [27] library. Based on these timestamps, the speech recordings are cut such that each resulting audio sample corresponds to one sentence. In total, $10\,903$ such samples uttered by 255 different speakers are obtained, of which 752 ($6.90\,\%$) are considered flattery. Overall, the dataset consists of almost 20 hours of speech, with an average sample length of 6.59 seconds.

A speaker-independent splitting into a training, development, and test set is created, where the training partition comprises $70\,\%$ of the speakers (178/255) while $15\,\%$ of the speakers are assigned to the development and test partition each. We ensure that the fraction of positive examples as well as the mean duration of samples in each partition is comparable. A detailed overview of the resulting dataset is provided in Table 1.

10.21437/Interspeech.2024-87

Table 1: *Dataset statistics.*

	train	development	test	total
# speakers (m, f)	178 (162, 16)	39 (35, 4)	38 (35, 3)	255 (232, 23)
# samples (flattery)	7167 (6.7%)	1878 (7.4%)	1858 (7.2%)	10903 (6.9%)
mean sample dur. (std) [s]	6.6 (±5.6)	6.6 (±5.3)	6.5 (±5.4)	6.6 (±5.5)
total dur. [s]	13:09:29	3:25:55	3:21:56	19:57:22

3. Methods

With both audio samples and their transcripts at our disposal, we conduct three types of experiments. First, we train text-based classifiers, for which we not only use the manual gold standard transcripts but also explore the outputs of various ASR systems (Section 3.1). Second, we aim to predict flattery based on the speech samples only, utilising a range of pretrained audio foundation models (Section 3.2). Third, we seek to combine the merits of text-based and audio-based approaches in one model (Section 3.3). Considering the class imbalance (cf. Table 1), we choose Unweighted Average Recall (UAR), also known as balanced accuracy as our evaluation metric in all experiments. All models are taken from the huggingface hub[1].

3.1. Text-based Classification

We build a text classifier utilising a pretrained RoBERTa [28] model in its *base* variant, i. e., a 12-layer Transformer encoder with about 110M parameters. We add a classification head added after the final layer's encoding and fine-tune all weights of the model utilising the training partition. The training process runs for at most 7 epochs but is aborted earlier if no improvement on the development set is observed for two epochs. The learning rate is set to 10^{-5} after initial experiments with the values $\{10^{-4}, 5 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}\}$. As the loss function, binary cross-entropy with positive samples weighted inversely to their frequency is utilised. We repeat the training process five times with different fixed random seeds.

Since in practice manual "gold standard" transcripts are typically not available, we also explore automatically generated transcripts obtained from different Automatic Speech Recognition (ASR) models. We consider 6 different pretrained models from the Whisper [29] family, ranging from the *tiny* variant with 39M parameters to *large* models comprising 1.5B parameters. We generate automatic transcripts using the models without any further adaptation. For each of the 6 ASR systems' outputs, we train instances of the RoBERTa classifier described above, applying the same procedure and hyperparameters as for the "gold standard" texts. Moreover, we compute the Word Error Rate (WER) on the dataset for every ASR system.

3.2. Audio-based Classification

We consider three different types of pretrained audio Transformers, namely variants of Audio Spectrogram Transformer (AST) [30], Wav2Vec 2.0 (W2V) [31], and Whisper [29]. AST [30] is a Transformer model with 12 layers that takes spectrograms as input. Specifically, we utilise AST trained on the Speech Commands V2 dataset [32], as this is the only speech-related model provided in [30]. W2V [31] is pretrained for reconstructing masked parts of speech signals. We employ both the *base* (12 Transformer layers) and *large* (24 Transformer layers) variant of W2V which were both pretrained and, subsequently, finetuned on the Librispeech [33] dataset containing 960 hours of speech. Moreover, as the task of SER is arguably related to our problem of flattery detection, we experiment with a W2V model finetuned on MSP-Podcast [34] for SER [13], denoted as W2V-MSP. As for Whisper [29], we make use of the *base*, *medium*, and *large* pretrained models.

Our choice of model families is motivated by the finding of [13] that models in the fashion of W2V acquire linguistic information when fine-tuned. Hence, we opt for a selection of W2V models fine-tuned for both ASR and SER, variants of Whisper, as a more recent ASR-based approach and AST that is solely trained on spectrograms and should thus not be equipped with linguistic knowledge. This, in combination with the text-based classification, allows us to reason whether flattery can mainly be recognised via prosody, text or both.

3.2.1. Layer-Wise Encodings

It has been shown that different layers of pretrained Transformer models for speech encode different acoustic and linguistic properties of an input signal [35]. Hence, for each model mentioned above, we investigate the aptitude of each of its layers for the flattery detection task.

First, for every model, we extract representations of the speech signal from each layer. For AST, we take the layer's embedding of the special `[CLS]` token as the representation. For Whisper and W2V models, representations are obtained by averaging over all of the respective layer's token representations. We then determine the most promising layer per model by training linear binary classifiers on each layer's representations. Given the large amount of trials, we choose Support Vector Machine (SVM) classifiers as a lightweight and deterministic option here. We optimise the regularisation constant C and the weighting of minority class examples in every experiment. In the second row of experiments, for every model, we only consider the layer with the best results in the first step and, in addition, the final layer. For both layers' representations, more extensive SVM hyperparameter searches are conducted that also optimise the kernel type (rbf, linear, Sigmoid, polynomial), the kernel coefficient γ, and, if applicable, the degree of the kernel function.

3.2.2. Audio Model Tuning

From each model family (AST, W2V, Whisper), we first select the variant performing best in the initial SVM experiments (cf. Table 3) and then finetune the pretrained model by training (i) the full model, and (ii) a version pruned to the layer that performed best among all layers in the SVM experiments. To do so, we add a linear classification head on top of each pretrained model, similar to the fine-tuning of RoBERTa (cf. Section 3.1). Analogously to the feature extraction process, we feed the final layers' `[CLS]` encoding for AST and the mean over the final layers' token representations for W2V and Whisper into the classification head. We determine a suitable learning rate for each model by training its pruned version for one epoch with different learning rates ($\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$) and three fixed random seeds. Binary Cross Entropy is employed as the loss function. We apply random oversampling to tackle the imbalanced label distribution here. Experiments with a weighted loss function did not yield promising results. The final models are trained with five different fixed random seeds.

3.3. Text + Audio Fusion

The fusion of the speech and text modality makes use of the models trained on speech only and text only, respectively. We consider the text-based models trained on i) the gold standard

[1] https://huggingface.co/models

transcripts, ii) the weakest ASR system's (Whisper-tiny) outputs, and iii) the best ASR system's (Whisper-large) outputs (cf. Table 2). For the sake of uniformity, we utilise the best finetuned audio model, i. e., W2V-MSP (cf. Table 3) for the speech modality. We apply a weighted **late fusion** on the respective predictions, where the weights for both models are chosen according to their respective performance on the development set. Furthermore, we experiment with **early fusion**. Specifically, for each pair of audio and text models, we extract their final layers' representations and concatenate them. Then, SVM classifiers are trained on these features, analogously to the process described in Section 3.2.1. Both the late and early fusion methods are deterministic, however, the models to be fused are all trained for the same five fixed random seeds. Thus, we can report means and standard deviations across these seeds by always fusing the models trained with the same seed.

4. Results

In the following, we report the results of the text-based (Section 4.1), speech-based (Section 4.2) and the fusion of text and speech (Section 4.3).

4.1. Text-based Classification Results

Table 2 presents the results of training the RoBERTa classifier on different transcripts. In addition, the WERs of the different ASR models are given.

Table 2: *ASR + Text pipeline results. We report mean UAR values and standard deviations across five fixed seeds.* Gold standard *refers to the transcriptions generated by humans.*

Transcriptions	# params (ASR)	% WER	RoBERTa [% UAR] dev	test
Whisper-tiny	39M	26.60	78.79 (±1.05)	80.96 (±0.98)
Whisper-base	74M	20.90	81.15 (±1.44)	80.23 (±1.41)
Whisper-small	244M	16.43	80.51 (±2.05)	83.49 (±1.14)
Whisper-medium	769M	14.94	81.26 (±1.39)	83.47 (±1.35)
Whisper-large	1.5B	14.68	**81.68** (±1.88)	83.71 (±1.68)
Whisper-large-v2	1.5B	14.80	79.50 (±1.65)	82.71 (±1.77)
gold standard	-	-	**82.67** (±1.69)	85.97 (±1.94)

It can be observed that the larger Whisper ASR models perform better than the smaller ones regarding their WER, with the *tiny* model producing texts with a WER of 26.60 % while the WERs of the *medium*, and *large* variants are around 15.00 %. Regarding the flattery classification, the best average result of 82.67 % UAR on the development and 85.97 % UAR on the test set is achieved when training with the gold standard transcripts. All results prove to be stable across seeds, as no standard deviation exceeds 2 % on the test set. While the gold standard transcriptions model outperforms the best ASR transcriptions model, i. e., *Whisper-large*, by more than 2 percentage points on the test data, all ASR transcript-based models still achieve over 80 % mean UAR on the test set. This indicates that for the task of textual flattery detection, even relatively high WERs such as 26.60 % for *Whisper-tiny* are not too detrimental to the text classifier's performance. One explanation for this is that high WERs in this particular data set are mainly due to highly domain-specific terms that carry no information related to flattery and are thus less relevant for the classification. Nevertheless, there is a connection between WERs and the corresponding classification results, with *Whisper-tiny* being responsible for the worst result on the development set (78.79 % UAR) while the best ASR-based classification result on the development set (81.68 % UAR) is achieved with the transcripts of *Whisper-large* that have the lowest WER among all the ASR models (14.68 %).

4.2. Audio-based Classification Results

The results for the audio-based flattery detection with both SVMs and finetuning are given in Table 3.

Table 3: *Results of the audio-based experiments. For the finetuning experiments, mean UAR values and standard deviations across five fixed seeds are given. The best SVM result per model family on the development set is underlined, the best development results overall are boldfaced for both SVMs and fine-tuned models.*

Model	Layer	SVM [UAR] dev	test	Finetuning [UAR] dev	test
AST	4	57.49	51.34	56.32 (±.1.46)	51.99 (±1.70)
AST	12	55.85	54.46	52.41 (±.0.60)	53.44 (±0.42)
W2V-base	7	75.36	72.94	-	-
W2V-base	12	66.84	62.63	-	-
W2V-large	11	78.45	75.60	-	-
W2V-large	24	73.70	69.17	-	-
W2V-MSP	11	79.70	82.23	-	-
W2V-MSP	12	**79.71**	82.46	**78.94** (±0.64)	80.60 (±0.58)
Whisper-base	5	69.27	.69.13	-	-
Whisper-base	6	70.04	66.62	-	-
Whisper-medium	23	79.46	76.31	72.32 (±6.44)	74.52 (±6.35)
Whisper-medium	24	79.37	75.52	76.94 (±2.83)	78.91 (±2.26)
Whisper-large	29	78.54	72.61	-	-
Whisper-large	32	77.05	76.28	-	-

The AST experiments yield considerably worse results than those based on the different W2V and Whisper variants. While most results of W2V and Whisper exceed 70 % UAR, all AST-based experiments only slightly surpass the chance level of 50 % UAR. Considering the UAR values of over 80 % observed in the text-based experiments, we assume that this performance gap is partially due to W2V and Whisper encoding linguistic information, which is not the case for AST. Consequently, the low UAR values for AST suggest that flattery can rarely be detected via prosodic information only. Another aspect that may contribute to AST's rather poor performance is that the SpeechCommand data it is initially trained on differs from our data in that all its speech samples are only one second long. Lastly, as our data is obtained from calls, the audio quality may be impaired, suppressing prosodic attributes of the speech samples that might prove beneficial for audio-based classification.

The layer-wise results confirm that different layers of pretrained models are of different suitability for the flattery detection task. This is particularly prominent for the W2V variants, where layer 7 clearly outperforms the final layer (12) in the base model and layer 11 leads to a considerably better result (75.60 %) on the test set than layer 24 (62.63 %) in the large model. As for the Whisper models, the best layers are always close, but never identical, to the ultimate layer. All W2V and Whisper variants yield results better than 75 % UAR on the development set in their best layer in the SVM results, with W2V-MSP achieving the best UAR values overall on both the development (79.71 %) and test (82.46 %) set.

Finetuning, on average, does not improve upon the SVM results. The standard deviation of 6.44 for Whisper-medium, however, shows that, depending on the random seed, results over 80 % UAR on the test set are possible. Overall, the best audio-based classifiers perform slightly worse than the best text-based

classifiers that achieve over 83 % mean UAR on the test set, cf. Table 2.

4.3. Text + Audio Fusion Results

We report the multimodal results in Table 4.

Table 4: *Results for the experiments fusing audio (A) and textual (T) information. We provide means and standard deviations across five fixed seeds, where the best result on development per transcription method is underlined, while the best overall is boldfaced.* T only *refers to the textual experiments reported in Table 2 for reference.*

Transcriptions	Method	[UAR] dev	[UAR] test
Whisper-tiny	T only	78.79 (±1.05)	80.96 (±0.98)
	Late Fusion A+T	79.72 (±1.50)	82.12 (±1.70)
	Early Fusion A+T	<u>81.85</u> (±2.04)	83.69 (±1.86)
Whisper-large	T only	81.68 (±1.88)	83.71 (±1.68)
	Late Fusion A+T	82.02 (±1.90)	83.94 (±1.39)
	Early Fusion A+T	<u>83.62</u> (±1.56)	84.71 (±1.01)
gold standard	T only	82.67 (±1.69)	85.97 (±1.94)
	Late Fusion A+T	83.02 (±1.56)	86.41 (±1.86)
	Early Fusion A+T	**<u>84.80</u>** (±1.33)	87.16 (±1.33)

It is evident that for all transcripts considered, a combination with the speech modality improves upon the text-only approach. Hence, it can be assumed that our speech-based models, though arguably also making use of linguistic information, encode information that complements the text-only representations to a degree. A comparison of late and early fusion shows that in all cases, the early fusion approach outperforms the late fusion method. The comparably weak performance of the latter may be attributed to the poor calibration we observe in our fine-tuned Transformers' predictions.

Among the different transcripts, the largest improvement over the purely textual model can be observed for those generated with Whisper-tiny, i. e., the worst performing ASR system (cf. Table 2). Specifically, its mean early fusion UAR value on the development set (81.85 %) exceeds its mean text-only UAR result (78.79 %) by 3.88 %. The relative improvement is lower for the transcripts obtained via Whisper-large and the gold standard, namely 2.38 % and 2.58 %, respectively. This suggests that the audio modality can complement text-only approaches to flattery detection especially when the ASR system's WER is relatively high. A closer manual inspection of data points for which audio and text classifiers disagree reveals another class of instances that benefit from speech-based classification: certain phrases, such as variants of *Great!* and *Good morning*, are sometimes labelled as flattery and sometimes not, depending on their context. Thus, as our simple text-based models do not consider the surrounding sentences, audio-based classifiers prove helpful in correctly predicting flattery in such utterances.

4.4. Generalisation to Female Speakers

Given that women are considerably underrepresented in our dataset (cf. Table 1), we investigate our models' generalisability for female speakers. In Table 5, the results of the best fine-tuned Transformer models are broken down into female and male speakers. While for both the text and the audio approach, the UAR values for women are lower than those for men, they are typically close. The largest gap is observed for the text-based predictions on the development set, with the UAR for females (75.99) being about 7 percentage points lower than that for males (83.27) on average. It is also evident that the results for the comparatively few female data points tend to vary more depending on the random seed. An explanation for the rather small gap in performance for female and male speakers may be that, at least in the business context, there may not be many gender-based differences when it comes to using flattering phrases. As the W2V models arguably also draw heavily on linguistic information, this reasoning would apply to them as well.

Table 5: *Results of RoBERTa finetuned on the gold standard transcripts (cf. Table 2) and finetuned W2V-MSP (cf. Table 3) for female (F) and, respectively, male (M) speakers only. We report means and standard deviations across* 5 *fixed seeds.*

Approach	Subset	[UAR] dev	[UAR] test
RoBERTa	F only	75.99 (±4.75)	83.76 (±4.68)
	M only	83.27 (±1.47)	86.21 (±1.91)
W2V-MSP (finetuned)	F only	77.99 (±3.59)	84.21 (±2.59)
	M only	78.91 (±0.65)	80.21 (±0.83)

5. Discussion

We observe that the textual modality, i. e., *what* is said, is crucial for predicting flattery. Second, the speech signal, while yielding less promising results on its own, still encodes valuable information that complements and thus improves text-based classification – especially in cases where the automatic transcription of utterances performs comparably poorly. Besides, the speech-based experiments with AST, Whisper, and W2V again demonstrate that fine-tuned ASR-based audio foundation models encode both linguistic and prosodic information.

Potential limitations to the generalisability of our models are induced by the nature of the data set. As the data is sourced from business analyst calls in US companies, it is arguably highly context-specific and not representative of the general population with respect to demographic aspects such as educational background or age.

6. Conclusion

We introduced the problem of flattery detection from speech alongside a novel data set. Furthermore, we trained an extensive set of ML approaches based on speech, text, and the combination of both modalities, thus providing insights into the nature of this novel task. Future work may include extending the database to cover broader demographics. Regarding the methodology, considering larger textual units in order to capture the sentences' contexts better is a promising avenue. Moreover, more refined fusion methods than those utilised in Section 3.3 can be devised. While we cannot publish our raw data due to copyright restrictions, we make our code, extracted features, and the best-performing models available[2].

7. Acknowledgements

This work was supported by MDSI – Munich Data Science Institute as well as MCML – Munich Center of Machine Learning. Björn W. Schuller is also with the Konrad Zuse School of Excellence in Reliable AI (relAI), Munich, Germany.

[2]https://github.com/lc0197/flattery_from_speech

8. References

[1] E. E. Jones, *Ingratiation: A social psychological analysis*. S.I.: Irvington Publishers, 1975.

[2] K. Kumar and M. Beyerlein, "Construction and validation of an instrument for measuring ingratiatory behaviors in organizational settings." *Journal of applied psychology*, vol. 76, no. 5, pp. 619–627, 1991.

[3] J. D. Westphal, "Board games: How ceos adapt to increases in structural board independence from management," *Administrative science quarterly*, pp. 511–537, 1998.

[4] R. Vonk, "Ingratiation," in *Encyclopedia of social psychology*, R. F. Baumeister, Ed. Sage, 2007, pp. 481–483.

[5] R. A. Gordon, "Impact of ingratiation on judgments and evaluations: A meta-analytic investigation." *Journal of personality and social psychology*, vol. 71, no. 1, p. 54, 1996.

[6] A. P. Ellis, B. J. West, A. M. Ryan, and R. P. DeShon, "The use of impression management tactics in structured interviews: A function of question type?" *Journal of applied psychology*, vol. 87, no. 6, p. 1200, 2002.

[7] J. D. Westphal and M. B. Clement, "Sociopolitical dynamics in relations between top managers and security analysts: Favor rendering, reciprocity, and analyst stock recommendations," *Academy of Management Journal*, vol. 51, no. 5, pp. 873–897, 2008.

[8] J. D. Westphal and D. L. Deephouse, "Avoiding bad press: Interpersonal influence in relations between ceos and journalists and the consequences for press reporting about firms and their leadership," *Organization Science*, vol. 22, no. 4, pp. 1061–1086, 2011.

[9] T. McEwen, "Communication training in corporate settings: Lessons and opportunities for the academe," *American Journal of Business*, vol. 12, no. 1, pp. 49–58, 1997.

[10] P. Cipresso and J. C. Immekus, "Back to the future of quantitative psychology and measurement: psychometrics in the twenty-first century," p. 2099, 2017.

[11] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: a transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[12] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. INTERSPEECH*, 2021, pp. 3400–3404.

[13] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 09, pp. 10 745–10 759, 2023.

[14] L. Christ, S. Amiriparian, A. Kathan, N. Müller, A. König, and B. W. Schuller, "Towards multimodal prediction of spontaneous humour: A novel dataset and first results," *arXiv preprint arXiv:2209.14272*, 2023.

[15] S. Amiriparian, L. Christ, A. König, E.-M. Meßner, A. Cowen, E. Cambria, and B. W. Schuller, "Muse 2022 challenge: Multimodal humour, emotional reactions, and stress," in *Proc. ACM Multimedia*, 2022, pp. 7389–7391.

[16] M. K. Hasan, W. Rahman, A. Bagher Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, and M. E. Hoque, "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proc. EMNLP-IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2046–2056.

[17] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, "Multimodal sarcasm detection and humor classification in code-mixed conversations," *IEEE Transactions on Affective Computing*, 2021.

[18] S. Amiriparian, J. Han, M. Schmitt, A. Baird, A. Mallol-Ragolta, M. Milling, M. Gerczuk, and B. Schuller, "Synchronization in interpersonal speech," *Frontiers in Robotics and AI*, vol. 6, p. 116, 2019.

[19] S. Amiriparian, L. Christ, R. Kushtanova, M. Gerczuk, A. Teynor, and B. W. Schuller, "Speech-based classification of defensive communication: A novel dataset and results," in *Proc. INTERSPEECH*, 2023, pp. 2703 – 2707.

[20] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh *et al.*, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," *arXiv preprint arXiv:2104.10121*, 2021.

[21] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[22] L. Christ, S. Amiriparian, A. Baird, A. Kathan, N. Müller, S. Klug, C. Gagne, P. Tzirakis, L. Stappen, E.-M. Meßner *et al.*, "The muse 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation," in *Proc. MuSe*, 2023, pp. 1–10.

[23] A. König, J. Mammen, J. Luger, A. Fehn, and A. Enders, "Silver bullet or ricochet? ceos' use of metaphorical communication and infomediaries' evaluations," *Academy of Management Journal*, vol. 61, no. 4, pp. 1196–1230, 2018.

[24] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[25] A.-K. Schill, A. Boutalikakis, F. Hawighorst, L. Graf-Vlachy, and A. S. Konig, "Analyst flattery, ceo narcissism, and ceo communication specificity," in *Academy of Management Proceedings*, vol. 2022, no. 1. Academy of Management Briarcliff Manor, NY 10510, 2022, p. 10892.

[26] N. Sadvilkar and M. Neumann, "PySBD: Pragmatic sentence boundary disambiguation," in *Proc. NLP-OSS*, E. L. Park, M. Hagiwara, D. Milajevs, N. F. Liu, G. Chauhan, and L. Tan, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 110–114.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Proc. INTERSPEECH*, vol. 2017. Stockholm, Sweden: International Speech Communication Association (ISCA), 2017, pp. 498–502.

[28] 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.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[30] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. INTERSPEECH*, 2021, pp. 571–575.

[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[32] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[34] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[35] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.