Comparison of Models for Detecting Off-Putting Speaking Styles

Diego Aguirre[1], *Nigel G. Ward*[1], *Jonathan E. Avila*[1], *Heike Lehnert-LeHouillier*[2]

[1]University of Texas at El Paso, Computer Science Department, USA
[2]New Mexico State University, Department of Communication Disorders, USA

daguirre6@utep.edu, nigelward@acm.org, jonathan.edav@gmail.com, hlehnert@nmsu.edu

Abstract

In human-human interaction, speaking styles variation is pervasive. Modeling such variation has seen increasing interest, but there has been relatively little work on how best to discriminate among styles, and apparently none on how to exploit pretrained models for this. Moreover, little computational work has addressed questions of how styles are perceived, although this is often the most important aspect in terms of social and interpersonal relevance. Here we develop models of whether an utterance is likely to be perceived as off-putting. We explore different ways to leverage state-of-the-art pretrained representations, namely those for TRILL, COLA, and TRILLsson. We obtain reasonably good performance in detecting off-putting styles, and find that architectures and learned representations designed to capture multi-second temporal information perform better.

Index Terms: pretrained models, prosody, autistic speech, temporal information

1. Introduction

Speaking style variation is important, for example in conversation, where participants may adjust their style to accommodate to the interlocutor, to produce speech suitable for the context, and to serve communicative goals [1, 2, 3, 4]. Style is complicated, not least because it relates to emotion, stance, personality, social identity and other factors. In the computational modeling space, much work on styles has been inspired by the potential to support the creation of virtual assistants and dialog systems that better serve the needs of diverse users, by adjusting to suit each individual [5, 6, 7]. Models of style may in addition ultimately help individual human speakers learn how to deploy styles more effectively for better social outcomes. Significant recent advances in in modeling speaking styles have been obtained in the context of speech synthesis [8, 9], but the focus of this paper is discriminative modeling.

Style detection has many applications, including for example better modeling of user and agent behavior for better call center analytics. In this paper we explore a new approach to style detection: exploiting pretrained models. In this we are inspired by recent work on related tasks — such as identifying disordered speech [10], synthetic speech [11], distorted speech [12], dysarthria [13], and emotion recognition — all of which have seen significant progress from the use of self-supervised approaches. These approaches commonly involve training an encoder on a large, unlabeled dataset to produce a general-purpose embedding representation of speech [14, 15]. These representations are used (and optionally fine-tuned) to build a final model for the target downstream task. Encoders based on different architectures (e.g., LSTM, CNN, Conformers) and different self-supervised learning strategies have been proposed, compared, and evaluated on a diverse set of downstream tasks.

However, to the best of our knowledge, the effectiveness of pretrained models for style detection has not previously been explored. In this work, we compare and evaluate components of these approaches on a socially-important aspect of speaking styles — predicting whether an utterance may be perceived as off-putting. Specifically, we train and evaluate these models on a corpus of dialogues manually labeled for off-puttingness at the utterance level. Our contributions include:

1. The presentation and public release of an initial model for detecting off-puttingness [16].
2. A demonstration of the value for this task of representations and models designed to capture multi-second temporal information.

2. Task and Data

Our task is that of detecting whether a given utterance is likely to be perceived as off-putting. We choose this for two reasons.

First, we believe this task is in some ways representative of other dimensions of style. We also suspected that it would have characteristics not commonly seen in other non-semantic speech tasks. Such tasks, for example those included in the Non-Semantic Speech Benchmark (NOSS) [17], primarily involve detecting speech patterns that are present in many short segments of the input signal. For example, speaker, emotion, language, dementia, and disordered speech detection are tasks where averaging embeddings extracted from 960ms segments is sufficient to build high-performing linear models [17, 18, 19]. We suspected that many of the speech patterns characteristic of styles occur at longer timeframes, and thus will need more than just features computed independently over only short segments of the input.

Second, we believe this task is very socially relevant. People who frequently produce utterances in a style that is perceived as off-putting are unlikely to have much social success. This may be the case for many people with autism. While many aspects of autistic speech have been intensively studied, most work has focused on low-level features, not patterns, and on simply computable properties, not how the speech is perceived by others [20, 21, 22, 23, 24, 25, 26]. We believe that a more complete understanding of the patterns of autistic speech and how it is perceived is needed to support the development of truly effective interventions [27].

Our dataset is derived from a corpus of dialogues in American English, each between an autistic child and a neurotypical confederate. We use the 12 described in [28] plus two more. These dialogues include a good number and good variety of utterances which could be perceived as off-putting. Each conversation is 7-10 minutes long, resulting in a total corpus duration of about 2 hours.

The judgments of whether each utterance was or was not off-putting were done independently by each of the first three authors, listening to all the autistic children's utterances, one-

Table 1: *Distribution of Utterance Annotations*

Duration (secs)	Utterances	
	Total	**Off-putting**
All	788	338
(0-1]	317	105
(1-2]	162	79
(2-3]	90	38
(3-4]	48	24
(4-5]	48	21
(5-6]	29	16
>6	94	55

by-one, in isolation, in random order. (While inter-speaker effects and other context are important for fully modeling autistic behavior [29], this task is more tractable.) The Fleiss' Kappa agreement score among the three annotators was 0.324, commonly interpreted as fair agreement. To combine annotations, we chose to consider an utterance as off-putting if it was so classified by at least one annotator. The final dataset consists of 338 (43%) off-putting utterances and 450 (57%) not off-putting ones. Table 1 presents the number of utterances classified as off-putting grouped by utterance duration.

As a side effort, we followed up with two categorizations of factors that may make utterances off-putting, both using a combination of unstructured observation and inductive analysis. First, we considered how these utterances may be perceived. They may give the impression that the speaker is angry, annoyed, aggressive or condescending; or being pedantic, forceful, unforgiving, demanding or complaining; or feeling bored, tired, uninterested, disengaged, or disconnected. Subsequent listening also suggested that off-putting impressions could arise if the speaker's utterances were poorly timed, either interrupting or coming too late, or if utterances were inappropriately short or long, or if the speaker's style choices were inappropriate for the context, in particular for the interlocutor's previous utterance. Other cases involved lack of clarity in turn-taking, turn-holding, or turn-yielding intentions.

Second, we considered acoustic and prosodic features that likely contribute to these perceptions. Off-putting utterances are sometimes over-articulated with frequent or strongly stressed syllables, or with high intensity or sharp initial intensity rise, or in breathy voice, or with unusually lengthened words, or unintelligible. Other issues included echoing the interlocutor's words and difficulties in starting or continuing utterances. Many of these properties have previously been noted in the literature. Clearly there is a wide diversity in both the acoustic properties of off-putting utterances and in the ways they are perceived.

3. Models

Our aim is to produce a useful model able to, given an utterance, determine if the style may be perceived as off-putting, but our primary objectives in this study were to address two research questions:

1. What speech representations and model architectures are most effective for building an off-putting style detector?
2. How important is temporal information across multi-second time frames when determining if an utterance is likely to be perceived as off-putting?

Accordingly, we train and evaluate a selection of models chosen to represent the state of the art and to support determination of which modeling methods are most effective.

To address the first research question, we try representations of four kinds: raw features (energy, F_0, and mel-frequency cepstral coefficients), representations produced by the CNN-based encoders used in TRILL [17] and COLA [18], and representations produced by TRILLsson [30]. COLA and TRILL were chosen since their representations perform well on a wide set of non-semantic speech tasks, such as those in NOSS (Non-Semantic Speech Benchmark) [17], and TRILLsson because the representations of its parent model, the Conformer-based model CAP12 [19], consistently outperform competing approaches, such as Wav2Vec2.0, on many paralinguistic tasks.

To address the second question, for each representation we train both a simple independent-slice model, where the features from each slice are aggregated by average pooling, and a time-sensitive model, using a recurrent network.

3.1. Low-Level Feature Representations

Our first approach consists of training classifiers in a fully supervised manner using low-level feature representations. Our intention was for these to be models simple enough to properly train on small data, and reasonable as a baseline. For each utterance in our dataset, base features are computed every 10ms, namely energy, F_0, and 13 mel-frequency cepstral coefficients, using the Midlevel Prosodic Features Toolkit [31]. Energy and F_0 are speaker-normalized to compensate for inter-speaker differences while keeping key prosodic information. From these representations we train two models. The first computes the mean and standard deviation of each base feature across all frames in the utterance, giving a fixed-size representation. This is then fed to a fully-connected network composed of two dense layers of 8 ReLU units each, followed by a sigmoid unit. The second model is designed to capture long-term dependencies. Each utterance is represented as a sequence of the low-level feature vectors and these are fed to a recurrent network composed of two GRU layers (32 units each), one dense layer (16 ReLU units), and one sigmoid unit. We use GRU units rather than LSTM units because they performed better in preliminary experiments. Otherwise we did not optimize the architecture.

3.2. CNN-based Representations (TRILL and COLA)

Our second set of models are built on the representations learned by TRILL [17] and COLA [18]. Both of these train CNN-based encoders that map 960ms audio segments to an embedding space. The self-supervised approach used by TRILL is based on a triplet-loss objective that steers the encoder, ResNetish [32], to produce similar representations for audio segments that are close in time. This is done by first creating a set of audio triplets where two audio segments (x_i, and x_j), sampled from the same audio clip, serve as positive examples and a third one (x_k), sampled from a different audio clip, serves as a negative example. The encoder g is then trained to produce representations where $||g(x_i) - g(x_j)|| \leq ||g(x_i) - g(x_k)||$. TRILL was trained on the subset of clips in AudioSet [33] that contain speech and it was shown to outperform competing approaches at the time it was released. Like TRILL, COLA trains an encoder (EfficientNet-B0 [34]) to output similar representations for pairs of audio segments extracted from the same clip. What differentiates COLA is that it uses all segments in the training mini-batch that come from other audio clips to build a large number of negative examples for each positive one.

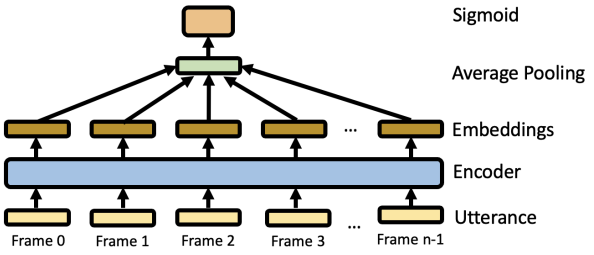

Figure 1: *Linear Model Architecture*

COLA was trained on the complete AudioSet corpus and it was shown to perform almost as well as TRILL when compared on the non-semantic tasks in NOSS.

For our TRILL-based models, we use the embeddings produced by the 19th layer of the ResNetish encoder (commonly referred to as TRILL-19) as they generally perform best. We used the pretrained models that were released through Tensorflow Hub. We extract embeddings from each non-overlapping 960ms segment in a given utterance, zero padding when necessary, to represent it as a sequence of 12288-d embeddings. Similarly, for our COLA-based models, we use the embeddings produced by the last global max-pooling layer of the EfficientNet-B0 encoder as done in [18]. Similar to our TRILL-based representations, we extract embeddings from non-overlapping 960ms segments in an utterance to represent it as a sequence of 1280-d embeddings.

For COLA, we did not find an available pretrained model, so we trained one using the publicly-released code. For this, we used the CREMA-D dataset [35], which consists of 7,442 sentences spoken by 91 actors across some basic emotional states (happy, sad, anger, fear, disgust, and neutral).

For both the TRILL- and COLA-based representations, we train both a linear and a recurrent model. For the former, as seen in Figure 1, we aggregate embeddings over time using average pooling and use the resulting fixed-size utterance representations to train a logistic regression classifier. Our recurrent model, presented in Figure 2, uses 2 stacked GRU layers (32 units each), one dense layer (16 ReLU units), and one sigmoid unit.

3.3. Conformer-based Representations (TRILLsson)

Our last set of models are built on the representations learned by the best performing TRILLsson [30] encoder. In [19], Shor *et al.* introduce a Conformer-based speech representation learning approach that outperformed competing approaches in many paralinguistic tasks. The model applies a convolutional feature encoder to the input signal and feeds the output to a speech encoder composed of blocks of Conformers (convolution-augmented Transformers) to generate embedding representations. Conformers, like CNNs, effectively extract useful local features from the input while also capturing global feature interactions, like Transformers. The model is trained using a modified Wav2Vec2.0 self-supervised strategy based on contrastive learning. The best performing representations are produced by the 12th layer of the 600M parameter variation of the model, referred to as CAP12, trained on 900k hours of unlabeled, YouTube audio data. CAP12 produces 1024-d embeddings for audio files of any length and it outperformed or matched competing approaches, including TRILL, in all NOSS tasks. This model, however, is very large and not publicly available. To address this, Shor *et al.* distill CAP12 into a series of smaller models, referred to as TRILLsson, using only public data [30]. The largest distilled model, TRILLsson5, based on the Audio Spectrogram Transformer architecture, achieves over 96% the accuracy of CAP12 on most NOSS tasks, and is publicly available. We used this pretrained version, as released through Tensorflow Hub.

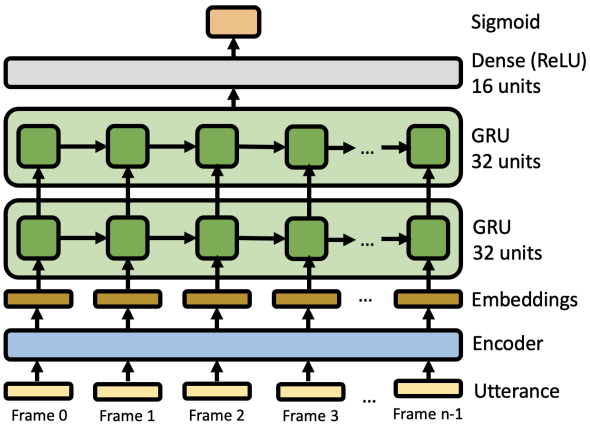

Figure 2: *Recurrent Model Architecture*

We use TRILLsson5 to train a linear model and a recurrent model, using the same embedding extraction and modeling strategy as for TRILL- and COLA-based models. In addition, we train a full-utterance model by feeding utterances (without segmenting them) to TRILLsson5 and using the resulting 1024-d embedding representations to train a logistic regression classifier.

4. Experiments and Results

We evaluate all models using 20-fold cross-validation, splitting the dataset into training (85%), validation (10%), and test (5%) sets in every fold. All models were built and evaluated using the same splits. We used a mini-batch size of 64 utterances, the Adam optimizer with a learning rate of 1e-3, and 500 epochs. After every training epoch, we evaluate the performance of the model on the validation set and keep track of the best performing model. After training, we choose the model that performed best on the validation set and evaluate its performance on the test set. To mitigate overfitting, we augmented our dataset by applying Gaussian noise to the input representations. For the recurrent models, we used dropout with a drop rate of 0.25 after the two GRU layers. We also applied L2 regularization with a 0.01 factor to the parameters of the sigmoid output of all models.

Table 2 shows the accuracy and F1-score of each model. These results indicate:

1. The embedding representations produced by TRILLsson5 generally outperform all other approaches. This is consistent with the performance gains observed in other non-semantic speech tasks.

2. For most learned representations, classification using recurrent models outperforms the linear model counterparts, suggesting that the speech patterns that emerge at time frames larger than 960ms are indeed important for detecting off-putting speaking styles. The advantage was however much less for

Table 2: *Performance results of models and representations on utterances of different durations*

	Decision Model	# of Params	Features/Pretrained Representation	Utterance Duration (secs) All	(0 - 1]	(1 - 2]	(2 - 3]	(3 - 4]	(4 - 5]	(5 - 6]	>6
Acc.	MLP	0.3K	Low-Level	0.623	0.647	0.562	0.644	0.625	0.646	0.690	0.596
	RNN	11.6K	Low-Level	0.610	0.634	0.636	0.467	0.625	0.583	0.586	0.638
	LR	12.3K	TRILL19	0.681	0.700	0.611	0.722	0.667	0.771	0.621	0.681
	RNN	1.2M	TRILL19	0.720	0.757	0.673	**0.778**	**0.729**	0.625	0.552	0.713
	LR	1.3K	COLA	0.600	0.634	0.611	0.589	0.500	0.521	0.552	0.585
	RNN	133K	COLA	0.631	0.653	0.617	0.633	0.604	0.667	0.483	0.617
	LR	1K	TRILLsson5 (0.96s)	0.726	0.760	0.642	0.756	0.708	0.750	0.690	**0.734**
	RNN	108.5K	TRILLsson5 (0.96s)	**0.744**	**0.770**	**0.679**	0.744	**0.729**	**0.812**	0.759	**0.734**
	LR	1K	TRILLsson5 (full)	0.712	0.748	0.667	0.700	0.688	0.688	**0.793**	0.681
F1	MLP	0.3K	Low-Level	0.494	0.423	0.458	0.543	0.500	0.564	0.667	0.578
	RNN	11.6K	Low-Level	0.507	0.326	0.614	0.400	0.609	0.524	0.571	0.667
	LR	12.3K	TRILL19	0.636	0.541	0.613	0.684	0.692	0.776	0.621	0.727
	RNN	1.2M	TRILL19	0.654	0.560	**0.679**	**0.762**	0.711	0.591	0.581	0.716
	LR	1.3K	COLA	0.521	0.496	0.577	0.507	0.478	0.410	0.480	0.581
	RNN	133K	COLA	0.503	0.389	0.508	0.507	0.578	0.600	0.516	0.625
	LR	1K	TRILLsson5 (0.96s)	**0.676**	**0.624**	0.623	0.725	0.696	0.750	0.727	**0.757**
	RNN	108.5K	TRILLsson5 (0.96s)	0.673	0.568	0.658	0.701	**0.745**	**0.791**	0.759	0.742
	LR	1K	TRILLsson5 (full)	0.652	0.565	0.654	0.649	0.717	0.681	**0.812**	0.712

TRILLsson5, suggesting that it already captures some aspects of temporal configurations.

3. Models trained on low-level representations performed poorly regardless of the model architecture. However they did outperform random guessing. This was true even for the independent-slice (MLP) model. This likely reflects the fact that averaged representations using low-level features extracted from very short frames (10ms) are enough to detect certain off-putting patterns, such as speech that is inappropriately loud or quiet.

4. COLA performance was lower than expected, and not generally better than using the low-level features directly, with no use of pretraining. This may be because the pretraining here was done on only a small dataset (CREMA-D).

We did not test statistical significance, but we note that these tendencies generally held for both F1 and accuracy metrics, and for utterances of different lengths.

We also computed the Fleiss' Kappa scores of the best performing models with respect to the targets: these included 0.462 for TRILLsson5-0.96s–RNN, 0.438 for TRILLsson5-0.96s–LR, 0.418 for TRILL19–RNN, and 0.406 for TRILLsson5-Full–LR. Although the comparison is not exact, it is encouraging that these are above the Fleiss' Kappa scores for the human annotators, which averaged 0.324.

5. Conclusion and Future Work

In this work, we evaluated and compared different modeling approaches and speech representations on a socially-relevant task: predicting if an utterance is likely to be perceived as off-putting.

The performance overall was modest, with the best accuracy only 0.744. We attribute this largely to the difficulty of the task, due in part to the heterogeneity among off-putting utterances, which is also reflected in the only "fair" agreement among human annotators. Another factor may be the difference in the speech used for pre-training (from non-autistic adults) and for the downstream task (from autistic children).

We show that TRILLsson produces better-performing representations than TRILL and COLA. We also show that models trained on sequences of embeddings outperform linear models trained on average-pooled embedding representations, suggesting that capturing patterns that emerge at time frames larger than 960ms is beneficial for speaking style modeling.

For future directions, we would like to perform similar experiments on a larger set of speaking styles, to evaluate other learned representations, such as [36], and to fine-tune encoders on our dataset. We would ultimately like to improve and extend this work to languages other than English and to younger children. In general, we see it as a priority for the field to explore the utility of models pretrained on adult data for tasks involving children.

6. References

[1] D. Tannen, "Conversational style," *Psycholinguistic models of production*, pp. 251–267, 1987.

[2] D. Biber, "Conversation text types: A multi-dimensional analysis," in *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*. Presses universitaires de Louvain Louvain, 2004, pp. 15–34.

[3] T. Prsir, J.-P. Goldman, and A. Auchlin, "Prosodic features of situational variation across nine speaking styles in French," *Journal of Speech Sciences*, vol. 4, no. 1, pp. 41–60, 2014.

[4] N. G. Ward and J. E. Avila, "A dimensional model of interaction style variation in spoken dialog," *Speech Communication, submitted*, 2022.

[5] M. Eskenazi and T. Zhao, "Report from the NSF future directions workshop, toward user-oriented agents: Research directions and challenges," 2020, arXiv preprint arXiv:2006.06026.

[6] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé, D. Dey, M. Harper, T. Howard, C. Kennington, I. Kruijff-Korbayová, D. Manocha, C. Matuszek, R. Mead, R. Mooney, R. K. Moore, M. Ostendorf, H. Pon-Barry, A. I. Rudnicky, M. Scheutz, R. S. Amant, T. Sun, S. Tellex, D. Traum, and Z. Yu, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 2022.

[7] N. Ward, J. E. Avila, and A. M. Alarcon, "Towards continuous estimation of dissatisfaction in spoken dialog," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 13–20.

[8] H. Cheng, H. Fang, and M. Ostendorf, "A dynamic speaker model for conversational interactions," 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. 2772–2785.

[9] C. Gordon, K. Georgila, V. Yanov, and D. Traum, "Towards personalization of spoken dialogue system communication strategies," in *Conversational Dialogue Systems for the Next Decade*. Springer, 2021, pp. 145–160.

[10] S. Venugopalan, J. Shor, M. Plakal, J. Tobin, K. Tomanek, J. R. Green, and M. P. Brenner, "Comparing supervised models and learned speech representations for classifying intelligibility of disordered speech on selected phrases," 2021, arXiv 2107.03985.

[11] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," 2019, arXiv 1904.05441.

[12] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The Interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Interspeech*, 2020.

[13] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," 2020, arXiv preprint arXiv:2001.00378.

[15] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," 2022, arXiv 2205.10643.

[16] D. Aguirre, "Off-putting speaking styles detection code," 2022, https://github.com/aguirrediego/off-putting-speaking-styles-detection.

[17] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *Interspeech 2020*. ISCA, 2020.

[18] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[19] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," 2021, arXiv preprint arXiv:2110.04621.

[20] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, 2015.

[21] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis," *Autism Research*, vol. 10, no. 3, pp. 384–407, 2017.

[22] S. Dahlgren, A. D. Sandberg, S. Strömbergsson, L. Wenhov, M. Råstam, and U. Nettelbladt, "Prosodic traits in speech produced by children with autism spectrum disorders: Perceptual and acoustic measurements," *Autism & Developmental Language Impairments*, vol. 3, pp. 1–10, 2018.

[23] S. P. Patel, K. Nayar, G. E. Martin, K. Franich, S. Crawford, J. J. Diehl, and M. Losh, "An acoustic characterization of prosodic differences in autism spectrum disorder and first-degree relatives," *Journal of Autism and Developmental Disorders*, vol. 50, pp. 3032 – 3045, 2020.

[24] S. Z. Asghari, S. Farashi, S. Bashirian, and E. Jenabi, "Distinctive prosodic features of people with autism spectrum disorder: a systematic review and meta-analysis study," *Scientific reports*, vol. 11, no. 1, pp. 1–17, 2021.

[25] S. Wehrle, "A multi-dimensional analysis of conversation and intonation in autism spectrum disorder," Ph.D. dissertation, University of Cologne, 2021.

[26] M. Grice, M. Krüger, Wehrle, F. Cangemi, and K. Vogeley, "Linguistic prosody in autism spectrum disorder – an overview," 2022, manuscript.

[27] S. Holbrook and M. Israelsen, "Speech prosody interventions for persons with autism spectrum disorders: A systematic review," *American Journal of Speech-Language Pathology*, vol. 29, pp. 2189–2205, 2020.

[28] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic entrainment in conversations of verbal children and teens on the autism spectrum," *Frontiers in Psychology*, p. 2718, 2020.

[29] K. Ochi, N. Ono, K. Owada, M. Kuroda, S. Sagayama, and H. Yamasue, "Entrainment analysis for assessment of autistic speech prosody using bottleneck features of deep neural network," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8492–8496.

[30] J. Shor and S. Venugopalan, "TRILLsson: Distilling universal paralinguistic speech representations," 2022, arXiv preprint arXiv:2203.00236.

[31] N. G. Ward, "Midlevel prosodic features toolkit," 2015-2022, https://github.com/nigelgward/midlevel.

[32] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[33] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[35] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[36] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," 2022.