SMM23, Workshop on Speech, Music and Mind 2023
18 August 2023, Trinity College Dublin, Ireland

Apparent personality prediction from speech using expert features and wav2vec 2.0

R. Barchi[1], *L. Pepino*[1,2], *L. Gauder*[1,2], *L. Estienne*[1,2], *M. Meza*[1], *P. Riera*[1,2], *L. Ferrer*[1]

[1] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[2] Departamento de Computación, FCEN, Universidad de Buenos Aires (UBA), Argentina

[gbarchi, lpepino, mgauder, lestienne, mmeza, priera, lferrer]@dc.uba.ar

Abstract

Studies have shown that virtual assistants that adapt to the personality of the speaker are more effective and improve the overall experience of the user. For this reason, automatic detection of a user's personality has recently become a task of interest. In this work, we explore the task of detecting a person's personality using their speech. To this end, we use the "First impressions Dataset" consisting of videos annotated with apparent personality labels. We train various systems using different modeling techniques and features extracted from the speech recordings including expert features commonly used for emotion recognition, and self-supervised representations given by wav2vec 2.0. We analyze the importance of each of these feature sets and relevant subsets for predicting the "Big-five" personality traits. Our results show that wav2vec 2.0 features are the most useful ones, and that their combination with expert features can result in additional gains.

Index Terms: apparent personality prediction, human-computer interaction, eGeMAPS, wav2vec.

1. Introduction

Personality traits have been extensively studied in relation to several aspects of human behavior such as emotional states [1], interests [2] and decision-making styles [3]. In the past few years, the growing popularity and demand of virtual assistants has led to an increased interest for automatically gathering this type of user information. This is due to the fact that interactive systems which are able to match users' personality have proved to improve overall experience [4] and communication effectiveness [5]. Further, several studies have shown that users tend to attribute personalities to virtual interfaces, even when a small number of cues is present, responding to them in a similar way than in human-human interactions [6].

Even though describing the personality of any individual is a complex task, a simplified model has been widely accepted, called the Five-Factor Model of Personality [7], consisting of ratings for five independent traits called the "Big-five": openness, conscientiousness, extraversion, agreeableness, and neuroticism. These ratings can be self-reported by the subject themselves through questionnaires or annotated by external listeners. Unfortunately, no dataset with speech or video recordings and self-reported personality is currently publicly available. For this reason, in this work, as in most similar works in the literature, prediction of personality traits is done using a dataset with ratings provided by listeners. This task is commonly called apparent personality prediction (APP).

The literature on APP has studied the use of video and audio signals and their combination for the prediction of personality traits [8, 9, 10]. In this work, we focus on paralinguistic features extracted from speech signals, which have been shown to be useful for predicting both self-reported [11, 12] and apparent personality [13, 14]. Among the speech-based features, pitch was shown to correlate positively with extraversion, and negatively with conscientiousness and agreeableness [15]. Other parameters that have been linked to personality are energy-related ones, loudness and sharpness, which have also shown to correlate positively with extraversion [16] and negatively with agreeableness [15]. An additional parameter that has been widely cited in the literature is speech rate, which was found to correlate with conscientiousness and agreeableness [17], as well as with extraversion [16]. The works above do not attempt to do prediction, focusing only on correlations between variables. A recent work that studies automatic prediction of personality traits using only the speech signal [18] shows that a model that used an extensive set of acoustic descriptors to predict self-reported personality ratings explained only 16% of the variance (unfortunately, the dataset used for this work is not publicly available). This demonstrated the difficulty of the task and the need for more research in this area.

In this work, we aim to provide a baseline for the performance of apparent personality prediction using only the speech signal. Such results, to our knowledge, have never been reported in the literature. To this end, we explore the use of expert features commonly used in emotion detection (eGeMAPS and speech ratio), and wav2vec 2.0 embeddings, modeling them with random forest regressors and a deep neural networks. We use the "First Impressions" dataset, the biggest publicly-available APP dataset to date, which consists of 10,000 short duration audios with Big-five personality ratings annotated by external listeners. We performed manual and automatic annotations on the presence of music in the recordings and used them to study the effect of music in the prediction. We highlight a problem with the official splits where chunks from the same video appear in training and test data resulting in optimistic results. Further, we study the impact of audio duration on performance. Finally, we include an analysis of feature importance by personality trait. The manually annotated music labels, the new data splits, and the scripts used for this work are available freely for research purpose upon request.

2. Methods

2.1. The "First Impressions" Dataset

The "First impressions" dataset [19], is a large publicly-available corpus labelled with personality ratings. It consists of 10,000 15-second videos extracted from youtube video blogs. It was first released for a personality prediction challenge (ChaLearn LAP 2016). Each video is rated with 5 labels corresponding to the Big-five dimensions, which were collected

10.21437/SMM.2023-5

via Amazon Mechanical Turk (AMT) and then converted to normalized ratings, following the procedure described in [20]. These ratings are used as target values for our work.

The corpus consists of English speakers (mostly native) of different ethnic groups and gender. Prior work showed a significant bias in the ratings given to males and females in this dataset [21], with higher average ratings for all 5 traits for women than for men. Regarding ethnicity, lower ratings are given to African-americans for all traits. In terms of data balance, male and female subsets are almost evenly distributed (M:45.4%, F:54.6%), while a severe imbalance is found across ethnic groups (Asian:3.3%, Caucasian:86.0%, African-American:10.7%). Given these observations, in our experiments, the data is split stratifying by gender, ethnicity, and average rating, ensuring that all splits contains the same proportion of females and males, of speakers of each ethnic group, and of speakers with similar average rating across all 5 traits.

In the official splits for this corpus, regions from the same video sometimes appear across training and test splits. In Section 3 we show that this leads to unrealistically optimistic results since models see some of the test speakers under the same acoustic conditions than in the training data. Hence, in this work, we split the data by video identifier.

In our initial data exploration we found samples containing background music, either edited or *in-situ*. To allow for an analysis of the impact of background music in system performance, we annotated samples containing music using two different approaches. The first one used Yamnet [22], an acoustic-event detection model trained on the Audioset Youtube corpus [23] to classify sounds into 521 event classes. We labelled a sample as having music if the probability of this event was larger than 0.1 according to Yamnet. This resulted in a subset of 1665 samples, which will be referred to as "Y-music", with an average SNR of 9.3 dB (std=5.8). After some analysis we found that Yamnet failed to detect low-volume music. Because of this, manual annotations were performed, by listening to 3 overall seconds of every audio, corresponding to initial, middle and final segments, and labelling as containing music any sample for which any level of music was heard by the annotator. This resulted in a subset of 3332 samples labelled as containing music, which we will call "M-music", with an average SNR of 12.5 dB (std=7.5) and a subset called "M-no_music", with an average SNR of 17.2 dB (std=9.5); which should not contain any music.

2.2. Feature extraction

Speech characteristics such as speech rate, intonation and energy, that are known to be affected by the speaker's emotion have also been shown to be affected by the speaker's personality [11]. For this reason, we explore the use of features that have been shown useful for the task of automatic emotion detection. The first set of features is the Minimalistic Acoustic Parameter Set (eGeMAPS) [24] which was shown useful for emotion detection in a number of previous studies [25, 26]. The features were computed using the openSMILE python library [27], in which 25 low-level descriptors (LLDs) including MFCC, F0, jitter, shimmer, spectral flux, spectral slope are extracted per frame (window size=30 ms, stride=10 ms) for the whole audio, and then summarized into 88 statistical descriptors, including mean, standard deviation, and percentiles. In addition to eGeMAPS, we compute the speech ratio for every sample, defined as the ratio between speech duration and overall duration. Speech timestamps were extracted using Silero [28], a neural network-based voice activity detector.

The second set of features explored in this work are wav2vec 2.0 (w2v2) embeddings [29], which were also found useful for emotion detection [30]. These embeddings are extracted from the output of the CNN encoder and 12 transformer blocks of the w2v2 base model implemented in S3PRL[1] [31], leading to 13 768-dimensional time series. Using all intermediate layers was shown in [32] to lead to better results than using only the last layer. To accelerate training, a local mean pooling with a window of 4 frames was applied to each time series, resulting in 12.5 embeddings per second.

To analyze whether background sounds may be degrading system performance, we explored extracting features over concatenated speech regions ("speech-only") and compared it with those extracted over the full signal ("full wav"). While background sounds might still affect the speech regions, we hypothesised that their impact could be mitigated by discarding the non-speech regions, as shown for emotion detection in [33]. For this analysis, we run Silero speech activity detector to determine the regions containing speech.

2.3. Model training and testing

We use two modeling approaches to predict the five personality ratings. The first approach is a multivariate random forest regressor (RFR) trained to predict all 5 ratings simultaneously. The model was implemented in the scikit-learn v1.0.2 python library [34]. Its parameters were set to default values: 100 trees, no maximum depth and mean squared error averaged over all 5 traits as error computation criterion. The input features are given by the concatenation of all or some of the following sets: eGeMAPS, the speech ratio, and a summarized version of the w2v2 sequence of embeddings. The summary w2v2 vector is obtained by concatenating the embeddings extracted from all 13 layers averaged over time. This leads to a 9984-dimensional vector for which PCA is applied, reducing the dimensionality to 256. The PCA transform is obtained using the training data.

The second modeling approach is a deep neural network (DNN) trained using the 13 w2v2 time series. Following the model proposed in [30], weights are learned to combine the 13 embeddings at each frame, resulting in a single 768-dimensional vector per frame. These vectors are processed with a shallow frame-wise neural network and, finally, mean-pooled over time obtaining a single 768-dimensional vector per sample. Finally, an output layer takes this vector and predicts values for the 5 personality traits. The dense layers before mean pooling allow the model to potentially learn a new set of features useful for solving the APP task before losing the temporal information. This gives the DNN model an advantage over the RFR model, since in that case mean pooling is applied over the raw features after which PCA is performed, a process that does not take the APP objective into consideration. Mean squared error is used as training loss and the models are trained for 20 epochs. Details on the architecture, training procedure and hyperparameters can be found in [30].

For both model types, the training process consisted of 5-fold cross-validation, with folds divided by video identifier and stratified by gender, ethnicity and average rating. In order to obtain confidence intervals for every regression result, cross-validation splits were repeated 10 times, varying the random seed in the fold-determination algorithm. Further, in both cases, bootstrapping was performed 100 times for each cross-validation iteration to assess the variability due to the test data.

[1] https://github.com/s3prl/s3prl

We use the coefficient of determination, R^2, as performance metric, computed on the concatenated predictions for all folds for each trait. Further, for some results we report the average per-trait R^2, R^2_{ave}.

3. Results and discussion

3.1. Background music effect

As described in section 2.1, the "First impressions" dataset was found to contain a significant amount of audio samples with background music. In order to quantify the impact of this type of noise on model performance, we compared the R^2_{ave} for the RFR trained and tested on the Y-music, M-music, and M-no_music subsets. For this experiment, we first sampled 1000 audios from each subset, since larger subsets would tend to give better results due to having more available training data. This experiment showed a significantly worse performance for the "M-music" (R^2_{ave} = 0.129, std = 0.056) and "Y-music" (R^2_{ave} = 0.111, std = 0.054) subsets compared to the performance on the "M-no_music" subset (R^2_{ave} = 0.149, std = 0.061). Since the main goal of our work is to explore the use of speech-based features, we use the "M-no_music" subset for the rest of the experiments in this paper, which contains 6668 audio samples. The manual annotations on the presence of music will be used in future work for further studies on the impact of music in this task.

3.2. Global feature-type importance

Figure 1 displays R^2_{ave} scores, corresponding to different combinations of expert features (eGeMAPS and speech ratio) and embeddings obtained from w2v2 as inputs to a RFR, as well as the results for the DNN trained on w2v2 embeddings. We compare the performance achieved by using the full audio signal (full wav) and only the concatenated speech regions (speech-only) as input for feature extraction. From the RFR results, we can see thatusing speech ratio in combination with egemaps features improves the model performance significantly in both conditions "full-wav" and "speech-only". This suggests that: 1) the speech ratio information is important for this task, and 2) egemaps features are unable to capture this information in its whole. When both sets are used, the results on speech-only and full wav are similar, suggesting that, for this dataset, it is not necessary to filter out the non-speech regions before feature extraction. In fact, the best RFR results are obtained when using the full waveform for feature extraction. Figure 1 shows that systems that include w2v2 embeddings outperform those that use only expert features. This suggests that w2v2 embeddings contain relevant information for the APP task which is not present in the expert features. The w2v2 embeddings are excellent features for speech recognition, outperforming those traditionally used for this purpose [29]. Hence, we hypothesize that these features may be allowing the system to model, among other things, word choices, something not readily available in the expert features. Further, we can see that adding the expert features to the w2v2 embeddings leads to small improvements for the RFR models when the full waveform is used. This suggests that some of the information in these features is not well-represented in the w2v2 embeddings.

Figure 1 shows that the performance of the DNN model trained on w2v2 embeddings is significantly better than that of any of the RFR models. This gain is likely due to the fact that the DNN model has access to the full sequence of w2v2 embeddings, while the RFR model only sees a PCA-transformed average of those embeddings. Unfortunately, unlike for the RFR model, for the DNN model we were not able to obtain further gains from the combination of eGeMAPS LLDs and w2v2 embeddings (results not shown to avoid cluttering the figure). This could be due to the need for hyperparameter tuning as DNNs can be very sensitive to the learning rate, number of training epochs, layer or batch normalization, and dropout, among other hyperparameters. We leave this exploration, which will require holding out test data so that hyperparameter tunning can be done without leading to optimistic results, for future work.

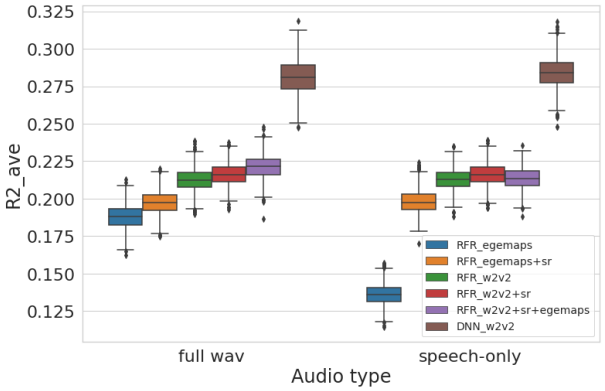

Figure 1: R^2_{ave} *scores for different combinations of features and audio inputs for RFR and DNN models.*

3.3. Per-trait results and effect of splitting strategy

As mentioned in Section 2.1, the official splits of the First Impressions dataset had sections of the same video in training, development and/or test splits, and were not stratified. Table 1 compares R^2_{ave} and individual per-trait R^2 scores, achieved by the best RF and the best DNN model using the official and our splitting strategy which divides the data by video and stratifies by gender, ethnic group and average rating. We can see that the global results observed in Table 1 carry over to individual traits, with the DNN model outperforming the RFR model for every personality trait. Further, we can see that "agreeableness" was the hardest trait to predict, as also found by the top performing teams of the Chalearn personality challenge [8, 9, 10]. We hypothesize that this could be due to an inherent difficulty of the task due to a lower agreement across annotators for this particular trait. Unfortunately, the annotator agreement for this dataset is not available. Finally, comparing results for both types of splits, we can see that the official splits led to optimistic results, likely due to sharing parts of the same videos in the test, development, and training splits. We encourage future users of this dataset to split the data according to video identifier. We can now compare our results with those obtained for the first impressions challenge. Our best result using the official splits has an R^2_{ave} of 0.33. In contrast, the best results reported after the challenge were around 0.50 [8]. Those systems, though, use both the video and the audio signal in the models while our systems use only the audio signal, since our goal is to analyze how much information about a person's personality can be extracted exclusively from their speech. Unfortunately, we were not able to find in those papers a report of audio-only results to make a direct comparison with our results.

3.4. Per-trait expert feature importance

In order to identify which expert features are the most important for predicting the rating of each personality trait, we divided

Table 1: *R^2_{ave} and individual R^2 scores, using the official and our proposed folds, corresponding to the best models: RFR with w2v2+eGeMAPS+sr features and DNN with w2v2 features.*

model	splits	O	C	E	A	N	R^2_{ave}
RFR best	official	0.30	0.28	0.30	0.20	0.32	0.28
	proposed	0.23	0.22	0.23	0.15	0.26	**0.22**
DNN best	official	0.34	0.36	0.35	0.22	0.36	0.33
	proposed	0.29	0.32	0.31	0.17	0.32	**0.28**

them into 7 subgroups (defined in the caption of Table 2). Then, a RFR was run for each subset and combination of 2 and 3 subsets using the procedure described in Section 2.3 except that, for these experiments, separate RFR were run to predict each trait to allow each model to freely choose the most useful features for each trait. Table 2 displays per-trait scores corresponding to the top performing combinations of groups of features (nº groups=1, 2 and 3), together with the best possible combination of features. We can see that the most important features are energy, pitch and spectrum cues, depending on the trait. The results regarding energy agree with the evidence found in literature in which it was shown to correlate with agreeableness [15] and neuroticism [35]. In fact, we found significant correlations ranging from 0.20 to 0.29 between all personality ratings and two eGeMAPS features ("loudness_sma3_amean", and "equivalentSoundLeveldB") which capture absolute energy-related characteristics. The importance of this acoustic cue was also confirmed by another experiment (not included in the table): a significant degradation in R^2_{ave} was observed when doing per-sample normalization of the energy signal compared to using the unnormalized energy as in the table, indicating that the absolute value of the energy is important for this task. Regarding pitch, we found an association with extraversion, as also found in [35, 11], and conscientiousness. Moreover, spectral information was found relevant to predict openness, agreeing with the results in [35]. When analysing 2- and 3-way combinations, the speech ratio is consistently found as a useful addition to the top feature. This finding is consistent with the observations in [36] where sound-silence ratio was found to correlate positively with extraversion. Another feature related to speech ratio, the speaking rate, was found to correlate with conscientiousness and agreeableness [17], as well as extraversion [16]. Further, the marginal performance gains which are observed when combining more than three groups of features suggest that the remaining features are not useful for prediction.

3.5. Duration analysis

Figure 2 shows the impact of audio duration on the R^2_{ave} performance when restricting the input samples to have fixed durations between 1 and 15 seconds (the duration of the original samples). The results are obtained using an RFR with eGeMAPS features on the full audio signal, and shorter samples obtained by cutting the original ones at a random location. One shorter cut is created for each original sample for each marker in the plot. We can see that the performance degrades drastically for shorter audio samples. In fact, at 15 seconds the performance curve has not converged yet, suggesting that the system could still benefit from larger samples.

4. Conclusions

In this work, we study the feasibility of predicting apparent personality from short speech recordings of subjects. We show that expert features commonly used in emotion detection (eGeMAPS and speech ratio), as well as self-supervised speech representations (w2v2), contain relevant information for this task. Further, we show that the combination of both types of features leads to gains with respect to the best individual system when modeled with a random forest regressor. We further improved the performance by using a DNN taking the sequence of w2v2 embeddings as input. Combining w2v2 with eGeMAPS LLDs did not result in further improvements and more work is needed in this direction. We hope that the results presented in this study, together with data splitting guidelines, code (which is available upon request), feature extraction techniques, and dataset annotations, prove useful for future research in this task.

Table 2: *Results for the best 1-, 2-, and 3-way combination of expert features for each trait. sr:speech ratio, eg:energy, spec:spectral, spec_v:spectral voiced, uv:unvoiced, $F0$:fundamental frequency. The list of features within each subset can be found in [24].*

Personality trait	Feature groups		R^2
Openness	top (n=1)	spec	0.120
	top (n=2)	sr+spec	0.190
	top (n=3)	sr+$F0$+spec_v	0.210
	best	sr+$F0$+eg+spec+spec_v+temp	**0.220**
Conscientiousness	top (n=1)	$F0$	0.070
	top (n=2)	sr+$F0$	0.180
	top (n=3)	sr+$F0$+eg	0.190
	best	sr+$F0$+eg+spec_v+temp	**0.200**
Extraversion	top (n=1)	$F0$	0.130
	top (n=2)	sr+$F0$	0.180
	top (n=3)	sr+$F0$+spec_v	0.220
	best	sr+$F0$+spec+spec_v+uv+temp	**0.220**
Agreeableness	top (n=1)	eg	0.020
	top (n=2)	sr+spec	0.100
	top (n=3)	sr+$F0$+spec_v	0.120
	best	sr+$F0$+eg+spec+spec_v+temp	**0.120**
Neuroticism	top (n=1)	eg	0.100
	top (n=2)	sr+spec_v	0.200
	top (n=3)	sr+$F0$+spec_v	0.230
	best	sr+$F0$+spec+spec_v+uv+temp	**0.240**

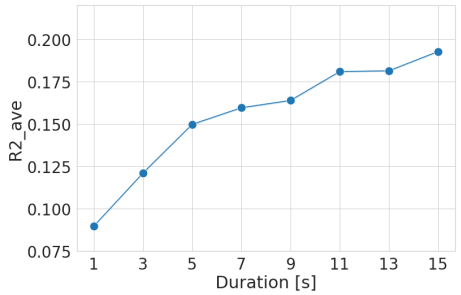

Figure 2: *RFR performance using the eGeMAPS set as a function of audio duration.*

5. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award no. FA9550-18-1-0026.

6. References

[1] K. M. DeNeve and H. Cooper, "The happy personality: a meta-analysis of 137 personality traits and subjective well-being." *Psy-*

chological bulletin, vol. 124, no. 2, p. 197, 1998.

[2] "The roles of personality traits and vocational interests in explaining what people want out of life," *Journal of Research in Personality*, vol. 86, p. 103939, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S009265661830254X

[3] R. E. Othman, R. E. Othman, R. Hallit, S. Obeid, and S. Hallit, "Personality traits, emotional intelligence and decision-making styles in lebanese universities medical students," *BMC Psychology*, vol. 8, 2020.

[4] P. Zhang, "Review of wired for speech: How voice activates and advances the human-computer relationship by c. nass, s. brave, the mit press, 2005," *Information Processing Management*, vol. 42, p. 1397–1399, 09 2006.

[5] J. Oberlander and A. J. Gill, "Individual differences and implicit language: personality, parts-of-speech and pervasiveness," 2004.

[6] C. Nass, Y. Moon, B. Fogg, B. Reeves, and D. Dryer, "Can computer personalities be human personalities?" *International Journal of Human-Computer Studies*, vol. 43, no. 2, pp. 223–239, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581985710427

[7] L. R. Goldberg, "The structure of phenotypic personality traits." *The American psychologist*, vol. 48 1, pp. 26–34, 1993.

[8] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, *Deep Bimodal Regression for Apparent Personality Analysis*, 11 2016, pp. 311–324.

[9] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *ECCV Workshops*, 2016.

[10] Y. Güçlütürk, U. Güçlü, M. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *ECCV Workshops*, 2016.

[11] F. Mairesse, M. Walker, M. Mehl, and R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res. (JAIR)*, vol. 30, pp. 457–500, 09 2007.

[12] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," *Proceedings of the 2nd international workshop on Social signal processing. ACM*, 10 2010.

[13] R. Solera-Ureña, H. Moniz, F. Batista, R. Astudillo, J. Campos, A. Paiva, and I. Trancoso, "Acoustic-prosodic automatic personality trait assessment for adults and children," 11 2016, pp. 192–201.

[14] T. Polzehl, K. Schoenenberg, S. Möller, F. Metze, G. Mohammadi, and A. Vinciarelli, "On speaker-independent personality perception and prediction from speech," *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, vol. 1, 01 2012.

[15] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420080405

[16] J. Sherzer, "K. r. scherer h. giles, eds., social markers in speech," *Homme*, vol. 20, no. 3, pp. 169–170, 1980. [Online]. Available: https://www.persee.fr/doc/hom_0439-4216_1980_num_20_3_368124

[17] B. Smith, B. Brown, W. Strong, and A. Rencher, "Effects of speech rate on personality perception," *Language and speech*, vol. 18, pp. 145–52, 04 1975.

[18] Z. N. Marrero, S. D. Gosling, J. W. Pennebaker, and G. M. Harari, "Evaluating voice samples as a potential source of information about personality," *Acta Psychologica*, vol. 230, p. 103740, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001691822002554

[19] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions -dataset and results," *European Conference on Computer Vision*, 10 2016.

[20] B. Chen, S. Escalera, I. R. Subramanian, V. Ponce-López, N. B. Shah, and M. O. Simon, "Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits," 2016.

[21] J. C. S. J. Junior, A. Lapedriza, C. Palmero, X. Baró, and S. Escalera, "Person perception biases exposed: Revisiting the first impressions dataset," pp. 13–21, 2021.

[22] M. P. . D. Ellis., "Yamnet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, 2013.

[23] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[24] F. Eyben and Scherer, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[25] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *Computer Speech Language*, vol. 65, p. 101119, 2021.

[26] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 478–484.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile – the munich versatile and fast open-source audio feature extractor," 01 2010, pp. 1459–1462.

[28] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[30] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[31] S. wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[32] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," pp. 914–921, 2021.

[33] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 519–523.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] T. Polzehl, S. Möller, and F. Metze, "Automatically assessing personality from speech," 10 2010, pp. 134 – 140.

[36] C. D. Aronovitch, "The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker," *The Journal of Social Psychology*, vol. 99, no. 2, pp. 207–220, 1976, pMID: 979189. [Online]. Available: https://doi.org/10.1080/00224545.1976.9924774