Towards understanding social interactions through audio signals

F.R. VOSSEBELD, University of Twente, The Netherlands

Everyone has social interactions in their lives. The way these social interactions occur might depend on the kind of person with whom is interacted. Different social relationships can be defined for this. These social relationships can be very different. Think of how the interaction with a stranger is different than with a friend. This research focuses on conversations between different relationships. Audio signal analysis is performed to determine whether the social relationship can be predicted purely on audio features. After this, different audio feature groups are compared to see whether one is more important for the prediction than the other. Then, the importance of the separate acoustic features is looked into. In order to do this, audio features are extracted from labelled conversations for which the social relationship is known. Random Forest performs the best on the data, achieving a balanced accuracy of 0.54 ± 0.06.

Additional Key Words and Phrases: social interaction, relationships, audio signal analysis, phonetics, emotions, conversation, audio features

1 INTRODUCTION

Everywhere in the world, people are constantly interacting with each other. Be it a mother that tries to teach her child something or the same mother that is firing an employee from her company. These are two very different social interactions. Also two very different social relationships. The mother speaks differently to her child than to her employees. This difference in speech between different social relationships is the core of this research. In what way do we speak differently to different persons? The goal of this research is to find out what audio features are key predictors for social interactions and then find a way to classify audio fragments to the corresponding social relationship between the people in the fragment. The final product should be able to predict whether the conversation is between e.g. strangers, a couple or different social relationships.

Social interaction is a concept that exists for tens of millions of years. Already in the Age of Dinosaurs, forms of social behaviour have been captured [36]. These social interactions come with different social relationships. The social relationship a son has with his mother is a different one than the one he has with e.g. his teacher or his brother. Many aspects of their preferred interaction will be different between the different relationships. An important way of social interaction is conversation. As you can imagine, a conversation with the cashier (a stranger) might sound very different from a conversation with your partner. The goal of this research is to find out what this audible difference is in different social interactions.

When speaking, there are different things that we (automatically) alter depending on the social relationship between you and the other person. For example, when you want to speak with confidence, pitch and speech rate are important properties that determine how this confidence is perceived [15]. In this research, properties such as mean pitch, mean loudness or mean alpha ratio will be referred to as features.

For this goal, different social relationships must be established and defined. Eight different labels are defined. This was done based on research by Liu et al. [18]. Two labels are left out because they could not be inferred confidently from the videos, namely *Sibling* and *Opponent*. This leaves six labels that will be used in this research, namely *Friend*, *Stranger*, *Service*, *Colleague*, *Parent-Offspring* and *Couple*.

OpenSmile [9] is an open-source software package that can extract all different kinds of audio features. The Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [10] is used to define the different features that will be extracted. This set of parameters describes Low Level Descriptors (LLD's) of the audio such as pitch or loudness. A total of 25 LLD's are described in the eGeMAPS [10]. These result in 88 different acoustic features.

1.1 Aim

Although research has been done in the field of social interactions and audio signal analysis, there is little research on the combination of the two. The role of audio and speech in particular for social relationships is to be investigated. The paper will analyze conversations from the Ego4D data set to find out whether certain audio features can help in distinguishing social relationships. The following research question (**RQ**) is raised.

RQ. How can we predict a social relationship based on audio signals?

We try to answer this question using the following sub-questions (**SQ**):

SQ 1 Are we able to accurately predict social relationships based on audio?

SQ 2 What (groups of) audio features are important for predicting social relationships?

1.2 Contribution

By answering the research question that was proposed above, we hope to find out how the audio signals of a conversation are influenced by the social relationship between people. With this paper, we try to provide more knowledge about the audio features that are important when distinguishing social interactions. The paper should give more insight in classifying social interactions without using visuals.

1.3 Organization of the paper

In this paper, first, the related works are discussed. In this section, past research about audio signal analysis and social interaction will be analyzed. The next section provides the proposed methodology. The steps for finding answers to the research questions are described here. In the Experiments section, the data set is presented with the validation part where the metrics will be presented to test the performance of the models. The results are also in this section. The

TScIT 37, July 8, 2022, Enschede, The Netherlands

results will be discussed in the following section; the Discussion section. Then, conclusions will be drawn with how they contribute to the field. Finally, some ideas are presented on how to continue working on this topic.

2 RELATED WORK

Audio signal analysis is not something new. Speech recognition, which is a theme within the audio signal analysis, is implemented in all kinds of new equipment nowadays. With the rise of Deep Learning, advanced models are available that can extract useful features from audio signals. Frameworks such as Wav2Vec (1.0 or 2.0) [25, 4] can give useful vector representations of audio files. Software packages such as OpenSMILE [9] and Praat [24] have proven to be very versatile. OpenSMILE offers an open-source audio feature extractor with a huge range of different extractable features. This is used a lot for emotion recognition but also for other things like sentiment extraction [23] or autism analysis [27].

Social interactions have been researched a lot in the field of sociology and psychology. In the last years, social interactions have also started to become a more interesting topic to data scientists and more research started happening in this field. However, most of this research is on the visuals of these interactions and not on the audio. These models often use Deep Neural Networks to classify the visual interactions [29, 2]. Even research has been done on social interaction from an egocentric view [12] which is interesting since the data set used in this research is also from an egocentric view [14].

The eGeMAPS feature set that is used in this research is developed to include features that can indicate emotional changes in voice production, features that have proven their values in other studies and features that have shown theoretical significance. After being created in 2016, many researches have used this feature set for emotion recognition. Not only is there a lot of emotion recognition research done using this feature set but [11, 33, 38], the set is also used for other purposes. The feature set has proven to be effective in the assessment of Parkinson's disease [35] but also in assessing psychiatric disorders [19]. This range of applications indicates the versatility of the set.

Phonetics in different social contexts have been researched in the past. For example, research has been conducted to find out the role of acoustics/audio features in communicating politeness [6, 17]. Also, the acoustic differences between acted and authentic emotional vocalizations have been researched. [3] The features that are used in these researches are similar to the features that are used in this research.

Something very applicable to this research and to which a lot of research has been done is the field of emotion. For the last 2 decades, emotion prediction through audio signals has been attempted many times. Some of them are on both the text and the speech, but some of them are also only on the speech/audio. An important returning conference in the field of this research is the INTERSPEECH conference [27, 26]. This yearly conference comes up with a challenge every year. The challenge in 2009 aimed to do emotion recognition from audio signals [26]. These challenges have resulted in different feature sets that can extract more information from the audio. In the last decades, much growth is seen in audio signal analysis. Where there was only binary arousal classification a decade ago, researchers have built models that can distinguish 5+ emotions with an accuracy of up to 90 per cent [20].

3 OUR PROPOSED METHOD

3.1 Extracting features from audio signals

As mentioned earlier, in this research audio features are first extracted using OpenSMILE [9]. The eGeMAPS [10] is used to describe which features should be extracted. Eyben et al. describe 88 features in this set and divided them into 4 different parameter groups. These parameter groups are as follows:

(1) Frequency-related parameters (24 features)
(2) Energy-related parameters (14 features)
(3) Spectral parameters (43 features)
(4) Temporal parameters (7 features)

A Python script is used to extract a total of 88 features per data point.

3.2 Finding the right model

To find the right model, 10-fold cross-validation is used to compare different algorithms. For the cause of using machine learning models, the Python library scikit-learn [22] was utilized. First, the data was standardized. For this, a StandardScaler is used which takes the standard deviation σ and mean μ of a feature and calculates a score per data point per feature according to Equation 1.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Both normalized input and non-normalized input are tried on the models. These sets are stratified to ensure that the training and test set have the same class ratio. Since the data set contains multiple sub-videos that are from the same video, Group k-fold testing should be used. This ensures that the train and test set never contain the same video. This is done in order to protect the model against data leakage. This way, the model is protected against overfitting to the video itself. This combination of grouping and stratifying is referred to as Stratified Group k-Fold cross-validation.

3.3 Testing models

The different algorithms that are used are the *Random Forest Algorithm (RF)* [16], the *Support Vector Machine (SVM)* [8] and the *k-nearest neighbours algorithm (KNN)* [1]. Hyper-parameter tuning was used to find better hyper-parameters for the different models. To do this, Bayesian Search [31] was used. This way, a big variety of hyper-parameters are tested. The different hyper-parameters that are tuned are in Table 1. The goal of the tuning is to find the model that has the highest average balanced accuracy in the 10-fold cross-validation. After choosing the classification algorithm that performs best, the data set can be split up into a training set (80 per cent) and a test set (20 per cent). The next step is training the best performing algorithm on the training set.

Table 1. Hyperparameter space of three different classification algorithms + found values

Algorithm	Hyperparameter	Range	Found Value
SVM	C	(1e-1 , 1e+2)	*51*
	gamma	(1e-4, 1)	*0.001*
	kernel	'rbf', 'poly', 'sigmoid'	*'rbf'*
KNN	leaf size	(1, 50)	*11*
	no. neighbours	(1, 30)	*9*
	p	(1, 2)	*1*
Random Forest	no. estimators	(200, 2000)	*1551*
	max features	'auto', 'sqrt'	*'sqrt'*
	max depth	(10, 110)	*81*
	min samples split	(2, 10)	*5*
	min samples leaf	(1, 4)	*1*
	bootstrap	(0, 1)	*0*

3.4 Finding the importance of feature groups

To find out whether there are feature groups that are more important than others for this research, Random Forest models are trained on a subset of features where one feature group (see par. 3.1) is excluded. Again, Stratified Group Cross-Validation is used to check the performance. However, this time 5-fold cross-validation was performed.

3.5 Finding the most important features

To answer RQ2, the importance of the different features has to be found. To do this, feature importance methods are used from the scikit learn library [22]. In this research, two ways of extracting feature importance are compared. Firstly, there is the feature importance based on the mean decrease in impurity (MDI) [5]. The second way to get feature importance is permutation feature importance [13]. These feature importance methods are compared by sorting the features on their importance and then building the model with an increasing number of top features.

By ordering the features on the extracted importance and training models on only a specific number of important features, it will become clear which features are (not) necessary for classification. Since there are two different orders of feature importance (permutation vs decrease in impurity), the performance of the models with a limited number of features will tell us which way of extracting feature importance works better. The scoring of these models will be evaluated again with Stratified Group 10-fold Cross-Validation.

4 EXPERIMENTS

4.1 Data set

The data set that was used in this research is the Ego4D data set [14]. This data set consists of 3,670 hours of video, all filmed from an egocentric perspective. The material consists of all kinds of different scenarios, including social interactions. The videos are filmed by 931 unique camera wearers from 74 worldwide locations. Since the focus lies on finding out more about social interactions and relationships, only a subset of the videos is used. This subset is found by using the transcripts provided by Ego4D. The algorithm that was used to split the videos into pieces with social interactions considered the following things.

- Conversation is longer than 10 seconds
- Both persons say more than 25 characters

Table 2. Example distribution in a fold of 10-fold cross-validation

Label	Train	Test	Total
Friend	436	48	484
Stranger	99	10	109
Service	121	14	135
Colleague	93	10	103
Parent-Offs	49	5	54
Couple	95	11	106

An algorithm that takes these rules into consideration results in 996 separate videos that contain conversation. These (sub)videos are parts of 255 unique original videos. A summation of all the video lengths results in 30,767 seconds / 513 minutes of conversation material. This material has been manually labelled by watching the videos. Since the visual material is not needed within the scope of this research, the audio has been extracted. This is done using the FFmpeg [32] library in Python [34]. In Table 2 the distribution of labels in the sub-set is displayed. From the figure, it is clear that this is a very imbalanced data set.

4.2 Validation

Since the data set is very imbalanced, with a lot of data on one label (Friend) and five smaller/outvoted labels, good metrics are important to test the performance.

$$Accuracy = \frac{\text{Correct predictions}}{\text{Total predictions}} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

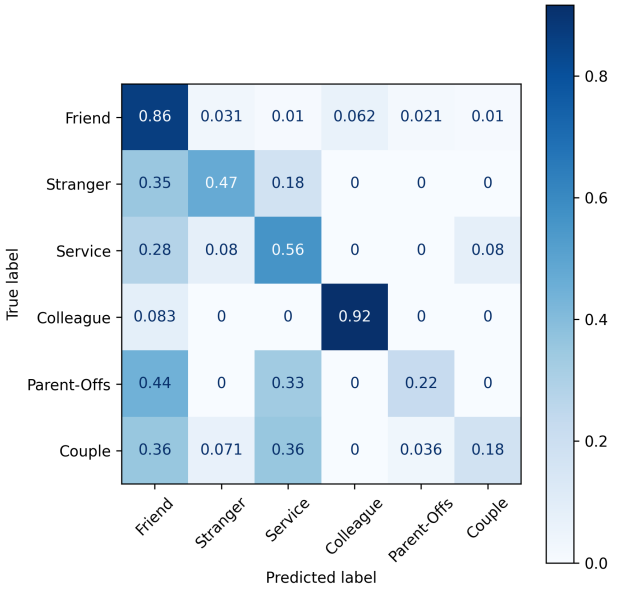


Fig. 1. Normalized Confusion Matrix of Random Forest

The metrics in Equation 2, 3 and 4 are the base metrics on which the model is tested. However, since the models are multi class classification ones, an average metric score per label should be taken to test the performance of the overall model. As seen in Table 3, three different metrics are used. Namely, *Accuracy*, *Weighted F1-score* and *Balanced Accuracy* were chosen. Accuracy is calculated according to Equation 2. The other two scores are however a bit more complicated because they combine binary classification metrics and take a calculated average of these per label. In the case of *Balanced Accuracy*, this is calculated by taking the average of recall (as seen in Eq. 3) that was obtained on each class [22]. The other one, the *weighted F1-score* is calculated by obtaining the F1-score (as seen in Eq. 4) per label and then finding their average weighted by the number of true instances per label [22]. By using Stratified Grouped 10-fold cross-validation, the variation of the scores on the whole data set can be tested and this makes sure that the model generalizes well.

4.3 Results

4.3.1 Extracted audio features. Five data points resulted in a few NaN values for some features. These five were excluded from the research, leaving 991 data points. From the 991 remaining sub videos, there were 88 audio features per piece. PCA and t-SNE are used to present the different features in a figure [21]. First, the features were scaled using a Standard Scaler (see Eq. 1). Then the dimension reduction was performed with PCA and t-SNE. Figure 2a shows a three-dimensional presentation of the extracted features per label with PCA. Figure 2b also shows a three-dimensional presentation but with t-SNE.

4.3.2 Comparison of classification algorithms. In Table 3 the results of the 10-fold Stratified Group Cross-Validation are shown. The

Table 3. Stratified Group 10-fold Cross-Validation scores per classification algorithm (normalized vs. non-normalized)

	accuracy	weighted f1	balanced accuracy
RF	**0.67 ± 0.05**	**0.64 ± 0.05**	**0.54 ± 0.06**
RF (norm.)	**0.67 ± 0.05**	**0.64 ± 0.05**	**0.54 ± 0.06**
SVM	0.48 ± 0.01	0.32 ± 0.01	0.16 ± 0.01
SVM (norm.)	0.65 ± 0.03	0.63 ± 0.03	0.52 ± 0.06
KNN	0.55 ± 0.02	0.49 ± 0.02	0.36 ± 0.03
KNN (norm.)	0.65 ± 0.03	0.62 ± 0.03	0.51 ± 0.03

10-fold cross-validation delivers 10 scores per metric. The results are the mean scores with their standard deviations. The normalized input is compared to the non-normalized input. The normalized input seems to outperform the non-normalized input, except for the Random Forest for which the normalization does not matter.

4.3.3 Random Forest results. The Random Forest turned out to be the best working model. This can be found in Table 3. The tuned Random Forest was then trained on the training set. After making predictions and comparing them to the true labels, a confusion matrix was created. This confusion matrix is normalized since this is easier to read because the data set is very imbalanced. The normalized confusion matrix can be found in Figure 1. With a balanced accuracy of approximately 0.54, the model fits the data and does not randomly predict.

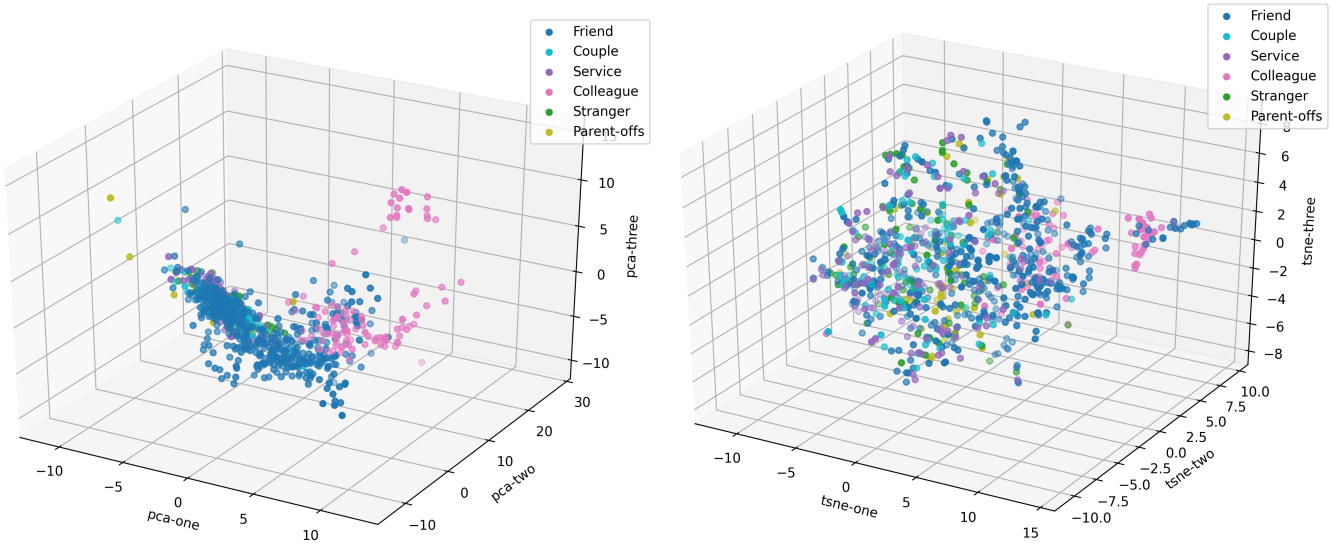
4.3.4 Important features. This subset of features was chosen by selecting the first N features, ordered by feature importance. Starting from the first 3 features, in steps of 3 till N is equal to the total number of features. The performances of these models were evaluated using 10-fold cross-validation. The results are in Figure 3. The glow in the figure is the range of the scores. The dark line is the mean. After comparing the two feature importance methods, the feature importance on a decrease of impurity seems to converge faster. This can be inferred from the fact that the values in Figure 3a converge faster to a higher score than in Figure 3b. This is supported by the fact that the area under the line (e.g. balanced accuracy) is bigger for the mean decrease in impurity feature importance (*40.89*) than for the permutation importance (*37.38*).

4.3.5 Comparing the feature groups. To find out whether some feature groups are more important than others, a 5-fold cross-validation was performed with a subset of features. Also, the score is included when all groups are included. There are no significant differences between the groups. The significance level was measured using a Mann-Whitney U test. The results are in Figure 4.

4.3.6 Most important features. In Table 4 the top 15 features are presented. More information on these features can be found in the research by Eyben et al. [10].

5 DISCUSSION

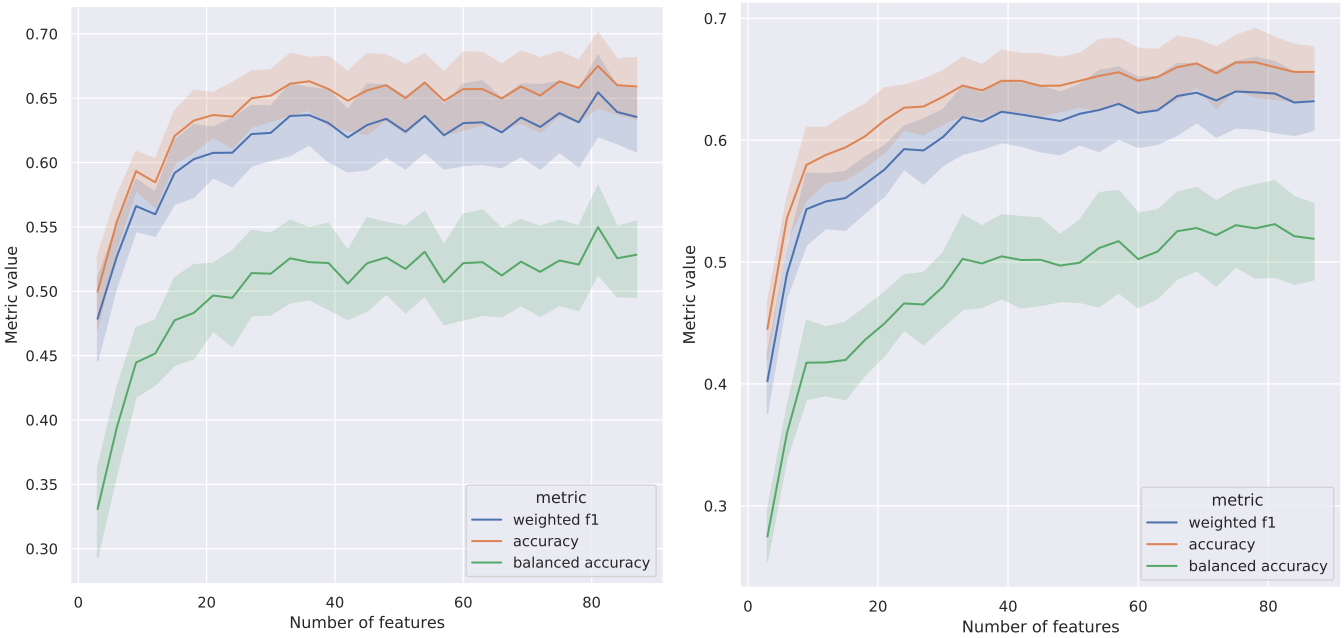
This research aimed to find out more about the relationship between audio signal features of a conversation and the social relationship. Firstly, the dimension-reduced features were plotted in 3D graphs. Unfortunately in these graphs there do not seem to be clear clusters



(a) eGeMAPS extracted features (3-dimensional PCA)

(b) eGeMAPS extracted features (3-dimensional t-SNE)

Fig. 2. Dimension reduced extracted features - PCA vs. t-SNE



(a) Feature importance on a decrease of impurity

(b) Permutation feature importance

Fig. 3. Comparing feature importance algorithms

for the labels. However, the 'Colleague' labels can be most distinguished from the big group.

After testing the different classification algorithms, it became clear that the Random Forest (RF) performed the best on this data. For this reason, the rest of the research was performed using the RF algorithm. The balanced accuracy of 0.54 for this multi-class classification indicates a model that found relationships between the audio features and the social relationship. The SVM and KNN clearly

Table 4. Top 15 feature importances sorted from high to low (mean decrease in impurity) - Random Forest

Group	Parameter	Functional	MDI
Spectral (balance)	Spectral Flux	stddevNorm	0.040
Energy/Amplitude related	Loudness	stddevNorm	0.032
Spectral (balance)	Spectral Flux V	stddevNorm	0.030
Spectral (balance)	Spectral Slope V 0-500Hz	amean	0.026
Spectral (balance)	Spectral Slope V 0-500Hz	stddevNorm	0.024
Spectral (balance)	Spectral Slope UV 0-500Hz	amean	0.022
Spectral (balance)	MFCC 1 V	amean	0.022
Energy/Amplitude related	Loudness	percentile20.0	0.022
Spectral (balance)	Spectral Flux UV	amean	0.022
Spectral (balance)	Hammarberg Index UV	amean	0.019
Frequency related	Pitch	amean	0.018
Spectral (balance)	MFCC 1	amean	0.017
Spectral (balance)	Alpha ratio UV	amean	0.016
Frequency related	Formant 3 bandwith	amean	0.016
Energy/Amplitude related	Harmonics-to-noise ratio	amean	0.016

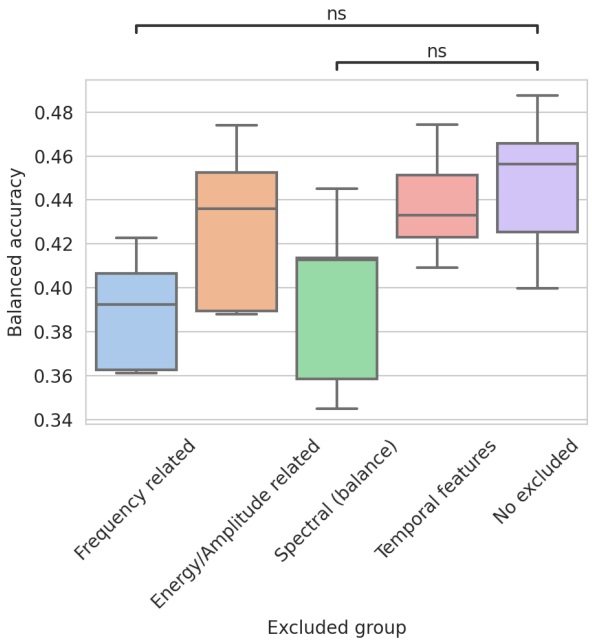


Fig. 4. Balanced accuracy of Random Forest after removing a feature group

benefit from the normalisation of the data but Random Forest works just as well with non-normalized data as with normalized data. Afterwards, this could have been foreseen by knowing how tree-based models (such as RF) work. As expected, there is a wide gap between the accuracy and the balanced accuracy or weighted f1. This emphasizes the importance of using a metric that takes imbalanced data into account.

As the Random Forest was trained on the training set (80 per cent) and tested on the test set (20 per cent), a confusion matrix was created by comparing the true labels in the test set with the predicted labels. In this confusion matrix, the 'Friend' label, which has the most data entries, is predicted more often than the other 5 outvoted labels. The 'Colleague' label was also predicted very well by the model. The prediction of the Parent-Offs label performs very poorly but this label was also the smallest in the data set.

Testing different feature importance methods was a good idea. The expectation was that permutation importance would work better because of the bias towards continuous and high cardinality features in Mean Decrease in Impurity (MDI) [28]. However, MDI turned out to give better feature importances, supported by Figure 3.

When comparing the feature groups, there was not one feature group that is significantly more important than the other ones. However, the Spectral and Frequency related parameters seem more responsible for the performance than the other two groups. As Frequency is a smaller group of features, we can conclude that these are more important than the Spectral (balance) features.

Looking at the top 15 features, it is noticeable that there are a lot of Spectral (balance) parameters that have an important contribution to the model. The Spectral Flux, Loudness and Spectral Slope are the most important parameters of the model.

Spectral Flux is defined as the difference in the spectra of two consecutive frames [10]. This corresponds with earlier research in which Spectral Flux was found to be the best overall speech arousal feature [37]. The fact that loudness is important is understandable when thinking about our voice usage in different social contexts. Spectral Slope has proven to be important for stress detection [30] and this is apparently also a determinative acoustic feature in different social interactions.

6 CONCLUSIONS

In this paper, an approach was presented on how to predict a social relationship from a conversation based on commonly used acoustic features. The used audio features were based on the Extended Geneva Minimalistic Acoustic Parameter Set [10] and extracted from a subset of the Ego4D data set [14]. The following contributions are presented:

- Social relationships are predictable using audio features.

- Classic ML algorithms such as SVM or Random Forest can be used for building predictive models for analyzing social interactions.
- Acoustic features that are used for emotion prediction can be used for analyzing social relationships in conversation as well.

7 FUTURE WORK

Looking at the work that is presented in this paper, there are things that could be considered in future research.
In this research, the imbalance of the data set has an effect on the performance of the model for the outvoted labels. In future research, a more balanced data set could be used to improve this. Also, in order to improve the performance on the same data set, techniques could be used to balance the set. Under-sampling the biggest class could improve the performance of the model but leaving data out is probably not desired. Using techniques like SMOTE [7] can help the model achieve higher performance by oversampling minority classes.
In this research, the only parameter set that was used was the eGeMAPS-set. In future research, different parameter sets could be compared. There are probably acoustic features that were not considered in this research but that play a role in social interactions. Also data augmentation could be done to make the model better. By adding for example noise or background sounds to the audio, the model might perform better on never seen data.

REFERENCES

[1] N S Altman. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression". In: *The American Statistician* 46.3 (Aug. 1992), pp. 175–185. ISSN: 0003-1305. DOI: 10.1080/00031305.1992.10475879. URL: https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879.

[2] Mohamed R Amer et al. *Human Social Interaction Modeling Using Temporal Deep Networks*. Tech. rep.

[3] Andrey Anikin and César F. Lima. "Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations". In: *Quarterly Journal of Experimental Psychology* 71.3 (2018), pp. 622–641. ISSN: 17470226. DOI: 10.1080/17470218.2016.1270976.

[4] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: (June 2020). URL: http://arxiv.org/abs/2006.11477.

[5] Leo Breiman. *Random Forests*. Tech. rep. 2001, pp. 5–32.

[6] Jonathan A. Caballero et al. "The sound of im/politeness". In: *Speech Communication* 102 (Sept. 2018), pp. 39–53. ISSN: 0167-6393. DOI: 10.1016/J.SPECOM.2018.06.004.

[7] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2011), pp. 321–357. DOI: 10.1613/jair.953. URL: http://arxiv.org/abs/1106.1813%20http://dx.doi.org/10.1613/jair.953.

[8] Corinna Cortes, Vladimir Vapnik, and Lorenza Saitta. "Support-vector networks". In: *Machine Learning 1995 20:3* 20.3 (Sept. 1995), pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018. URL: https://link.springer.com/article/10.1007/BF00994018.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. "OpenSMILE - The Munich versatile and fast open-source audio feature extractor". In: *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*. 2010, pp. 1459–1462. ISBN: 9781605589336. DOI: 10.1145/1873951.1874246.

[10] Florian Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7.2 (Apr. 2016), pp. 190–202. ISSN: 19493045. DOI: 10.1109/TAFFC.2015.2457417.

[11] H.M. Fayek, M. Lech, and L. Cavedon. "Evaluating deep learning architectures for Speech Emotion Recognition". In: *Neural Networks* 92 (2017), pp. 60–68. DOI: 10.1016/j.neunet.2017.02.013.

[12] Simone Felicioni and Mariella Dimiccoli. "Interaction-GCN: A Graph Convolutional Network based framework for social interaction recognition in egocentric videos". In: (Apr. 2021). URL: http://arxiv.org/abs/2104.14007.

[13] Damien François, Vincent Wertz, and Michel Verleysen. *The permutation test for feature selection by mutual information*. Jan. 2006, pp. 239–244.

[14] Kristen Grauman et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video". In: (Oct. 2021). URL: http://arxiv.org/abs/2110.07058.

[15] Joshua J. Guyer, Leandre R. Fabrigar, and Thomas I. Vaughan-Johnston. "Speech Rate, Intonation, and Pitch: Investigating the Bias and Cue Effects of Vocal Confidence on Persuasion". In: *Personality and Social Psychology Bulletin* 45.3 (Mar. 2019), pp. 389–405. ISSN: 15527433. DOI: 10.1177/0146167218787805. URL: https://journals.sagepub.com/doi/full/10.1177/0146167218787805.

[16] Tin Kam Ho. "Random Decision Forests". In: *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*. ICDAR '95. USA: IEEE Computer Society, 1995, p. 278. ISBN: 0818671289.

[17] Debi Laplante and Nalini Ambady. "On How Things are Said: Voice Tone, Voice Intensity, Verbal Content, and Perceptions of Politeness". In: *Journal of Language and Social Psychology* 22.4 (Dec. 2003), pp. 434–441. ISSN: 0261927X. DOI: 10.1177/0261927X03258084.

[18] Zihe Liu et al. "A Multimodal Approach for Multiple-Relation Extraction in Videos". In: *Multimedia Tools and Applications* 81.4 (Feb. 2022), pp. 4909–4934. ISSN: 15737721. DOI: 10.1007/s11042-021-11466-y.

[19] D.M. Low, K.H. Bentley, and S.S. Ghosh. "Automated assessment of psychiatric disorders using speech: A systematic review". In: *Laryngoscope Investigative Otolaryngology* 5.1 (2020), pp. 96–116. DOI: 10.1002/lio2.354.

[20] Mustaqeem and Soonil Kwon. "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach". In: *Expert Systems with Applications* 167 (Apr. 2021). ISSN: 09574174. DOI: 10.1016/J.ESWA.2020.114177.

[21] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1901), pp. 559–572. ISSN: 1941-5982. DOI: 10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

[22] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[23] S. Poria et al. "Context-dependent sentiment analysis in user-generated videos". In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 1. 2017, pp. 873–883. ISBN: 9781945626753. DOI: 10.18653/v1/P17-1081.

[24] *Praat: doing Phonetics by Computer*. URL: https://www.fon.hum.uva.nl/praat/.

[25] Steffen Schneider et al. "wav2vec: Unsupervised Pre-Training for Speech Recognition". In: *Interspeech 2019*. Vol. 2019-September. ISCA: ISCA, Sept. 2019, pp. 3465–3469. DOI: 10.21437/Interspeech.2019-1873. URL: https://www.isca-speech.org/archive/interspeech_2019/schneider19_interspeech.html.

[26] Björn Schuller, Stefan Steidl, and Anton Batliner. *The Interspeech 2009 Emotion Challenge*. Jan. 2009, pp. 312–315.

[27] Björn Schuller et al. *The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism **. Tech. rep. 2013. URL: http://mistral.univ-avignon.fr/index.

[28] Erwan Scornet. "Trees, forests, and impurity-based variable importance". In: (Jan. 2020). DOI: 10.48550/arxiv.2001.04295. URL: https://arxiv.org/abs/2001.04295v3.

[29] Behjat Siddiquie et al. "Affect analysis in natural human interaction using Joint Hidden Conditional Random Fields". In: *Proceedings - IEEE International Conference on Multimedia and Expo* (2013). ISSN: 19457871. DOI: 10.1109/ICME.2013.6607590.

[30] Olympia Simantiraki et al. *Stress Detection from Speech Using Spectral Slope Measurements*. Nov. 2016.

[31] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. *PRACTICAL BAYESIAN OPTIMIZATION OF MACHINE LEARNING ALGORITHMS*. Tech. rep.

[32] Suramya Tomar. "Converting Video Formats with FFmpeg". In: *Linux J.* 2006.146 (June 2006), p. 10. ISSN: 1075-3583.

[33] P. Tzirakis et al. "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks". In: *IEEE Journal on Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309. DOI: 10.1109/JSTSP.2017.2764438.

[34] G Van Rossum and F L Drake. "Python 3 Reference Manual; CreateSpace". In: *Scotts Valley, CA* (2009), p. 242. URL: https://www.python.org/.

[35] J.C. Vásquez-Correa et al. "Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach". In: *IEEE Journal of Biomedical and Health Informatics* 23.4 (2019), pp. 1618–1630. DOI: 10.1109/JBHI.2018.2866873.

[36] Lucas N. Weaver et al. "Early mammalian social behaviour revealed by multituberculates from a dinosaur nesting site". In: *Nature Ecology and Evolution* 5.1 (Jan. 2021), pp. 32–37. ISSN: 2397334X. DOI: 10.1038/S41559-020-01325-8.

[37] Felix Weninger et al. "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common". In: *Frontiers in Psychology* 4 (2013). ISSN: 1664-1078. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2013.00292.

[38] S. Zhang et al. "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching". In: *IEEE Transactions on Multimedia* 20.6 (2018), pp. 1576–1590. DOI: 10.1109/TMM.2017.2766843.