ATTITUDE RECOGNITION USING MULTI-RESOLUTION COCHLEAGRAM FEATURES

Fasih Haider and Saturnino Luz

Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh
Edinburgh, UK

ABSTRACT

Attitudes play an important role in human communication. Models and algorithms for automatic recognition of attitudes therefore may have applications in areas where successful communication and interaction are crucial, such as healthcare, education and digital entertainment. This paper focuses on the task of categorizing speaker attitudes using speech features. Data extracted from video recordings are employed in training and testing of predictive models consisting of different sets of speech features. A novel attitude recognition approach using Multi-Resolution Cochleagram (MRCG) features is proposed. The results show that MRCG feature set outperforms the feature sets most commonly used in computational paralinguistic tasks, including *emobase*, *eGeMAPS* and *ComParE*, in terms of attitude recognition accuracy for decision tree, 1-nearest neighbour and random forest classifiers. Analysis of the results suggests that MRCG features contribute information not captured by these existing feature sets. Indeed, while the *ComParE* feature set provides slightly better results than MRCG features for support vector machine classifiers, the fusion of the existing feature sets with the new MRCG features improves on those results. Overall, with the addition of MRCG, the attitude recognition method proposed in this study achieves accuracy scores approximately 11 points higher than reported in previous studies.

Index Terms— Feature Engineering, Attitude Recognition, Affect Recognition, Multi-Resolution Cochleagram, Video Blogs.

1. INTRODUCTION

The emerging fields of social signal processing and affective computing seek to build models to automatically characterise human behaviours in interactive situations. This includes the detection of emotions and attitudes which can, among other things, influence communication effectiveness both in dialogue and in presentations. Methods developed in these fields have found applications in the analysis of clinician-patient communication [1], education [2], and entertainment [3, 4]. In this paper, the social signals of presenters in monological discourse is explored. This is done through the analysis of speech in a data set that consists of video "blogs" (vlogs).

Vlogs have become a popular form of online communication in recent years. While the "vlogger" (video blogger) does not receive feedback from viewers in real time, viewers can provide feedback asynchronously in the form of textual comments. Studies conducted on video blogs concluded that the non-verbal behaviour of the vlogger influences the level of attention gained by a video [5]. Therefore, a plausible application of automatic analysis of non-verbal behaviour in vlogs is providing feedback to the vlogger so that they can improve their vlogs. Other scenarios such as the above mentioned application to consultation skill training for clinicians could also benefit from automatic feedback. In addition, automatic recognition of attitudes could also help develop tools for recommendation, summarization and search of videos. In this study, we focus on attitude recognition, where an attitude is defined as a state that may permeate strong emotions [6].

In the discipline of affective computing, many techniques have been proposed for the detection of affective states in different contexts ranging [7, 8, 9]. However, analysis of vlogs has not been explored extensively in the literature. In one study, the facial expression, acoustic (speaking activity and prosodic) features, and multimodal information are used to predict personality traits in vlogs using regression analysis [10]. In a perceptual and acoustic analysis is performed for 12 different attitudes expressed by Portuguese speakers [11], results showed that audio-visual data provide better perception of attitudes than any single modality. An analysis of speaking time, F0 energy, speech rate, speech turns along with head motions, looking time, and proximity to camera by Biel et al. [4] showed that audio-visual non-verbal cues are significantly correlated with the median number of log views.

Allwood et al. proposed an automatic attitude detection system for multimodal dialogue systems using multimodal speech cues [12]. Madzlan et al. [13] used the acoustic and high-level visual features (i.e. facial landmarks) for a clas-

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661, SAAM project.

sification task to recognise the attitude automatically. They proposed a three-class problem grouping the attitudes in three classes: positive, negative and neutral [13], defining friendliness as a neutral attitude, amusement and enthusiasm as positive attitudes, and frustration and impatience as negative attitudes. Their results suggest that acoustic features provide better results (63.63% accuracy) than visual features (50.6%). However, they did not perform fusion of features. In a different study [14], Madzlan et al. analysed prosodic features of vloggers and found that these features (F0, voice quality and intensity) are correlated with a vlogger attitude, while in [15] they analysed audio-visual features for their attitude recognition.

In a previous study, we proposed an attitude recognition system using audio (*emobase* feature set) and visual (Fisher vector representation of dense histograms of gradient, dense histograms of flow and dense motion boundary histograms) features which achieved an accuracy of 58.72% [16]. In another study [3], we proposed an active feature transformation method for attitude recognition using MFCC features and achieved an accuracy of 56.61% in a 6-class attitude recognition task [3]. This study extends our previous work [16, 3]. Its main contributions are, therefore:

1. the presentation of a novel attitude recognition method and computational models[1],
2. an assessment of the discriminating power of different audio features (*emobase*, *eGeMAPS*, *ComParE* and MRCG feature sets) for the recognition of six attitudes (amusement, enthusiasm, friendliness, frustration, impatience and neutral) in speech data, and
3. a demonstration of the discriminating power of MRCG feature sets for attitude recognition. This is, to the best of our knowledge, the first such use of MRCG.

2. MULTI-RESOLUTION COCHLEAGRAM FEATURES

MRCG features have been proposed by Chen et al. [17] and have since been used in speech related applications such as voice activity detection [18] and speech separation [17]. MRCG features are based on cochleagrams [19]. A cochleagrams is generated by applying the gammatone filter to the audio signal, decomposing it in the frequency domain so as to mimic the human auditory filters. MRCG uses the time-frequency representation to encode the multi-resolution power distribution of an audio signal. Four cocheleagram features are generated at different levels of resolution. The high resolution level encodes local information while the remaining three lower resolution levels capture spectrotemporal information. A total of 768 features are extracted from each frame: 256 MRCG features (frame length of 20 ms and frame shift of 10 ms), along with 256 Δ MRCG and 256 $\Delta\Delta$ MRCG, meant to capture temporal dynamics of the signal [17].

[1]The data, models, and extracted features used in this study are available to the research community at `git@git.ecdf.ed.ac.uk:fhaider/attitudeRecognitionModels.git`

3. DATA SET

This study uses the video-blog dataset [13, 16]. This dataset contains 613 audio-visual segments from around 250 different videos that are annotated for the six different attitudes shown on Table 1. The data annotation was performed by two annotators with an inter-coder agreement of 75% as reported in [20]. The duration of video clips is around 1-3 seconds. This study uses the audio information only.

Table 1. Number of instances (speech utterances/video clips) for each attitude in the dataset

Attitude	Abbrev.	No. of Utterances
Amusement	A	103
Enthusiasm	E	107
Friendliness	Fd	100
Frustration	Fr	104
Impatience	I	103
Neutral	N	100

4. EXPERIMENTATION

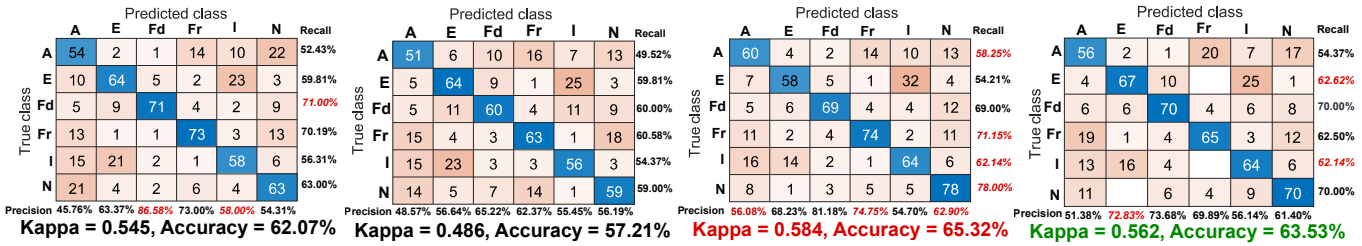
4.1. Feature Extraction

We employed the openSMILE [21] to extract the acoustic features which has been widely used for emotion recognition [7]. This study uses three openSMILE feature sets (*emobase*, *eGeMAPS* and *ComParE* feature sets) and MRCG feature set. The following is a brief description of each of the feature sets used in this study:

emobase: This acoustic feature set contains the MFCC, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives. In addition, many statistical functions are applied to these features, resulting in a total of 988 features for every speech utterance.

ComParE: The *ComParE* [22] feature set includes energy, spectral, Mel-Frequency Cepstral Coefficients (MFCCs), and voicing related Low-Level Descriptors (LLDs). LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. This feature set contains 6373 acoustic features for every speech utterance.

eGeMAPS: The *eGeMAPS* [23] feature set contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, hammarberg index and slope V0 features including many statistical function applied on these feature which resulted in-total of 88 features for every speech utterance.



(a) *emobase* (b) *eGeMAPS* (c) *ComParE* (d) *MRCG*

Fig. 1. Confusion matrix of attitude recognition using SVM classifier.

MRCG: The statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) are applied on the 768 MRCG features (detailed in Section 2) which resulted in total of 6912 features for every speech utterance.

4.2. Classification Methods

The classification is performed using five different methods namely Decision Tree (DT), Nearest Neighbour (KNN with K=1), Naive Bayes (NB) with kernel distribution, Random Forest (RF) and Support Vector Machines (SVM). DT, 1NN, NB and SVM (with linear kernel) classifiers are employed in MATLAB[2] using the statistics and machine learning toolbox in the 10-fold cross-validation setting. The RF classifier is employed using scikit learn[3] in 10-fold cross validation setting with 2,500 trees in the forest, with a leaf size of 50. KNN and DT are other non-parametric, non-linear methods, included for comparison.

5. RESULTS AND DISCUSSIONS

We conducted an experiment using the above described acoustic feature sets (*emobase*, *eGeMAPS*, *ComParE* and MRCG) and assessed the results in terms of accuracy. The data set is almost balanced for attitudes (classes), implying a "blind guess" accuracy of 16.67%, and majority class accuracy of 17.34%. The classification results of the feature sets using DT, 1NN, NB, RF and SVM classifier are reported in Table 2. Of the five classification methods, the results indicate that the SVM classifier provides the best results in all tested settings. The MRCG feature set provides better results than other feature sets for DT (43.76%), 1NN (39.22%), NB (51.22%) and RF (56.20%). The *ComParE* feature set provides a better result (65.32%) than others for the SVM classifier. However the mean accuracy of MRCG is higher across all classifiers, suggesting that the MRCG feature set is more reliable (50.79%) for different classification algorithms than other feature sets. While the *ComParE* feature set provides the best overall result, this feature set is the least robust (mean accuracy of 41.45%) to classifier change, as shown in Table 2.

Table 2. Accuracy (%) obtained using different feature sets along with the average accuracy (mean) for each feature set over five classifiers.

Features	*DT*	*1NN*	*NB*	*RF*	*SVM*	*mean*
emobase	39.38	25.77	49.27	53.43	62.07	45.98
eGeMAPS	*42.95*	*33.39*	40.36	52.73	57.21	45.33
ComParE	42.14	17.83	27.55	54.41	**65.32**	41.45
MRCG	**43.76**	**39.22**	**51.22**	**56.20**	63.53	***50.79***

To obtain further insight into the results, we draw the confusion matrix of the SVM classifier for the feature sets (Fig. 1). It can be seen that the *ComParE* feature set provides the best recall for A, Fr, I and N with an overall accuracy of 65.32% and Kappa [24] of 0.584. The MRCG feature set provides the best recall for E and I with an overall accuracy of 63.53% and Kappa of 0.562. The precision and recall for all the feature sets including overall accuracy and Kappa are also shown in Fig. 1.

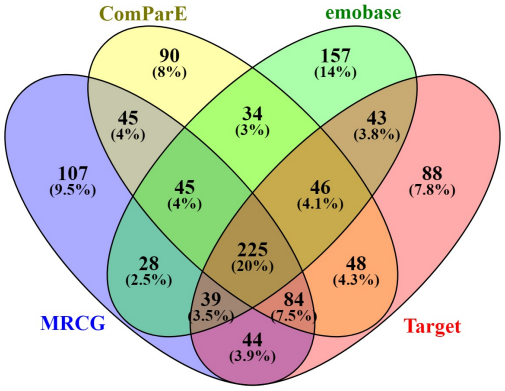


Fig. 2. Venn Diagram of the best results of three feature sets and annotated labels (Target).

To observe the relationship between top three feature sets for SVM classifier, we drew a Venn diagram as depicted in Fig. 2. In that Figure, the red ellipse (Target) represents the

[2]http://uk.mathworks.com/products/matlab/ (October 2018)
[3]http://scikit-learn.org/stable/index.html (October 2018)

Table 3. Fusion Results: Accuracy (%) obtained using different feature sets along with the average accuracy (mean) for each feature set over five classifier.

Features	DT	$1NN$	NB	RF	SVM	$mean$
eGeMAPS + MRCG	**48.14**	**37.76**	**44.89**	56.99	66.45	**50.85**
emobase + MRCG	42.79	27.55	47.16	56.03	66.77	48.06
ComParE + MRCG	46.68	17.83	28.04	56.70	68.56	43.56
emobase + *ComParE* + MRCG	46.52	17.83	28.04	57.17	**69.53**	43.82
emobase + *eGeMAPS* + MRCG	46.68	30.96	42.14	57.34	67.26	**48.88**
ComParE + *eGeMAPS* + MRCG	46.52	17.83	26.42	56.51	**69.53**	43.36
emobase + *eGeMAPS* + *ComParE*	42.14	17.83	25.77	55.57	69.04	42.07
emobase + *eGeMAPS* + *ComParE* + MRCG	46.03	17.83	26.09	**57.51**	**69.53**	43.40

annotated labels, the yellow ellipse represents the predicted labels by the *ComParE* feature set using the SVM classifier, the green ellipse represents the predicted labels by the *emobase* feature set, and finally the blue ellipse represents the predicted labels by the MRCG features using the SVM classifier. From the overlaps in this Venn diagram, it is observed that there are 88 instances (13 of A, 18 of E, 16 of Fd, 17 of Fr, 16 of I and 8 of N) which have not been recognised by any of the feature sets. However there are 225 instances (21 of A, 37 of E, 44 of Fd, 44 of Fr, 34 of I and 45 of N) which have been detected by all three feature sets.

The MRCG and *emobase* feature sets provide slightly less accurate results than the *ComParE* feature set but are nevertheless able to capture information which is not captured by the *ComParE* feature set, as shown by 43 (overlap of red: Target and green circle: *emobase*), 39 (overlap of green: *emobase*, red: Target and blue circles: MRCG) and 44 (overlap of red: Target and blue: MRCG circles) instances in our tests. This observation suggests that the fusion of feature sets could improve the results. Therefore we fused the feature sets and rerun the classification task. The results are shown in Table 3. It is observed that the fusion does indeed improve results for the DT, RF and SVM classifiers, even though a decrease in accuracy is observed for 1NN and KNN. The best accuracy overall (69.53%) is obtained with the fusion of features. This is higher than the best classifier with the *ComParE* feature set alone (65.32%). The confusion matrix of the best fusion result is shown in Fig. 3. The fusion decreases recall of one (E) out of six attitudes and precision of two (Fd and I) out of six attitudes but overall accuracy and Kappa results are improved.

In a previous study [16], we evaluated the *emobase* (acoustic) feature set along with visual (Fisher vector representation of dense histogram of gradient, dense histogram of flow and dense motion boundary histogram) features and found that the acoustic features provides better results than visual, achieving a maximum accuracy of 58.72% [16]. In another study [3], we evaluated MFCC features, principal component analysis of MFCC and a new method of active feature transformation of MFCC features, achieving 56.61%

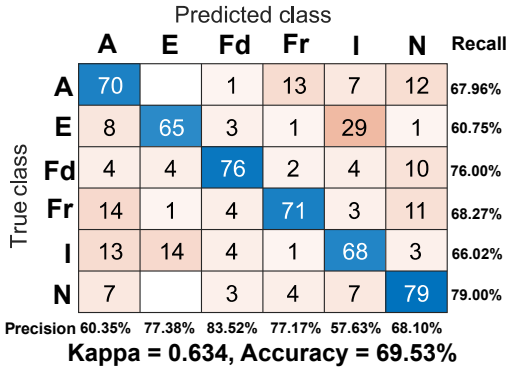


Fig. 3. Confusion matrix of attitude recogntion using feture sets (*ComParE* + *eGeMAPS* + MRCG) fusion for SVM classifier.

accuracy [3]. However, these previous studies did not evaluate the *ComParE*, *eGeMAPS* and MRCG feature sets for attitude recognition. This study demonstrates the usefulness of MRCG features for this attitude recognition task, achieving a maximum accuracy of 69.53% which is almost 11% higher than obtained in previous studies.

6. CONCLUSION

The MRCG feature set provides the superior results for a wide range of classification algorithms, including DT, 1NN, NB and RF classifiers. In this sense, MRCG features appear to be more robust than other features. While the *ComParE* feature set provided slightly better results using the SVM classifiers, that feature set performed poorly with most other classifiers. Fusion of feature sets results in an overall improvement over individual feature sets. The method proposed in this study also improves the accuracy up to around 11% as compared to previous studies. Future work includes applying different feature selection or transformation methods to assess whether dimensionality reduction can result in accuracy improvements over the full MRCG feature set, and evaluating the MRCG feature set on the analysis of patient-clinician communication as well as other emotion recognition datasets.

7. REFERENCES

[1] Padhraig Ryan, Saturnino Luz, Pierre Albert, Carl Vogel, Charles Normand, and Glyn Elwyn, "Using artificial intelligence to assess clinicians' communication skills," *BMJ*, vol. 364, 2019.

[2] Chunfeng Liu, Rafael A. Calvo, and Renee Lim, "Improving medical students' awareness of their non-verbal communication through automated non-verbal behavior feedback," *Digital Education*, p. 11, 2016.

[3] Fasih Haider, Fahim A. Salim, Owen Conlan, and Saturnino Luz, "An active feature transformation method for attitude recognition of video bloggers," in *Proc. Interspeech 2018*, 2018, pp. 431–435.

[4] Joan-Isaac Biel and Daniel Gatica-Perez, "Vlogsense: Conversational behavior and social attention in youtube," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 7, no. 1, pp. 33, 2011.

[5] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez, "You are known by how you vlog: Personality impressions and nonverbal behavior in youtube.," in *ICWSM*, 2011, pp. 446–449.

[6] Mark P Zanna and John K Rempel, "Attitudes: A new look at an old concept. s. 315-334 in: Bar-tal, d./kruglanski, aw (hrsg.), the social psychology of knowledge," 1988.

[7] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[8] Hayakawa Akira, Fasih Haider, Loredana Cerrato, Nick Campbell, and Saturnino Luz, "Detection of cognitive states and their correlation to speech recognition performance in speech-to-speech machine translation systems," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2539–2543.

[9] Carl Vogel and Liliana Mamani Sanchez, "Epistemic signals and emoticons affect kudos," in *3rd IEEE Conference on Cognitive Infocommunications*, Péter Baranyi, Ed., 2012, pp. 517–522.

[10] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 53–56.

[11] João Antônio De Moraes, Albert Rilliard, Bruno Alberto de Oliveira Mota, and Takaaki Shochi, "Multimodal perception and production of attitudinal meaning in Brazilian Portuguese," *Proc. Speech Prosody, paper*, vol. 340, 2010.

[12] Jens Allwood and Peter Juel Henrichsen, "Predicting the attitude flow in dialogue based on multi-modal speech cues," in *NEALT Proceedings. Northern European Association for Language and Technology; 4th Nordic Symposium on Multimodal Communication; November 15-16; Gothenburg; Sweden*. Linköping University Electronic Press, 2013, number 093, pp. 47–53.

[13] Noor Alhusna Madzlan, Yuyun Huang, and Nick Campbell, "Automatic classification and prediction of attitudes: Audio-visual analysis of video blogs," in *International Conference on Speech and Computer*. Springer, 2015, pp. 96–104.

[14] Noor Alhusna Madzlan, Jingguang Han, Francesca Bonin, and Nick Campbell, "Towards automatic recognition of attitudes: Prosodic analysis of video blogs," *Speech Prosody, Dublin, Ireland*, pp. 91–94, 2014.

[15] Noor Alhusna Madzlan, Jing Guang Han, Francesca Bonin, and Nick Campbell, "Automatic recognition of attitudes in video blogs-prosodic and visual feature analysis.," in *INTERSPEECH*, 2014, pp. 1826–1830.

[16] Fasih Haider, Loredana Sundberg Cerrato, Saturnino Luz, and Nick Campbell, "Attitude recognition of video bloggers using audio-visual descriptors," in *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, New York, NY, USA, 2016, MA3HMI '16, pp. 38–42, ACM.

[17] Jitong Chen, Yuxuan Wang, and DeLiang Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[18] Juntae Kim and Minsoo Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, 2018.

[19] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[20] Noor Alhusna Madzlan, Justine Reverdy, Francesca Bonin, Loredana Cerrato, and Nick Campbell, "Annotation and multimodal perception of attitudes: A study on video blogs," in *Proceedings from the 3rd European Symposium on Multimodal Communication, Dublin, September 17-18, 2015*. Linköping University Electronic Press, 2016, number 105, pp. 50–54.

[21] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[22] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[23] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "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.

[24] J Richard Landis and Gary G Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.