INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Affect Recognition through Scalogram and Multi-resolution Cochleagram Features

Fasih Haider and Saturnino Luz

Usher Institute, Edinburgh Medical School, the University of Edinburgh, Edinburgh, UK
Fasih.Haider@ed.ac.uk, S.Luz@ed.ac.uk

Abstract

An approach to the categorization of voice samples according to emotions expressed by the speaker is proposed which uses Multi-Resolution Cochleagram (MRCG) and scalogram features in a novel way. Audio recordings from the EmoDB, EMOVO and Savee Data-sets are employed in training and testing of predictive models consisting of different sets of speech features. This study systematically evaluates the performance of the feature sets most commonly used in computational paralinguistic tasks (i.e. *emobase*, *eGeMAPS* and *ComParE*) in addition to MRCG- and scalogram-derived features and their fusion, across five different classifiers. The datasets used in this evaluation include speech in three different languages (German, Italian and English). MRCG features outperform the feature sets most commonly used in computational paralinguistic tasks, including *emobase*, *eGeMAPS* and *ComParE*, for the EmoDB (unweighted average recall, UAR $= 59.15\%$) and SAVEE (UAR $= 36.12\%$) datasets, while *eGeMAPS* provides the best overall UAR (33.84%) for the EMOVO dataset. A support vector machine (SVM) classifier yields the best UAR for EmoDB (80.05%) through fusion of *emobase*, *eGeMAPS*, *ComParE* and MRCG, and for EMOVO (40.31%), through fusion of *emobase*, *eGeMAPS* and *ComParE*. For SAVEE, random forests provide the best result (46.55%) using the *ComParE* feature set.
Index Terms: Emotion recognition, Affective computing, Social signal processing

1. Introduction

Speech signals are used in a number of automatic prediction tasks, including emotion recognition [1, 2]. Emotional states can also have an influence on health and intervention outcomes [3]. Positive emotions have been linked with health improvement, while negative emotions may have negative impacts on wellbeing [4]. For example, long term accumulations of negative emotions are predisposing factors for depression, while positive emotions-related humour and optimism have been linked with positive health outcomes, such as effects on the immune system or association with cardiovascular diseases [5]. Emotion recognition is used in health technology applications, as in the assessment of depression [6, 7], health monitoring in smart environments [8], or dementia diagnosis and prognosis (see [9] for a survey). Ambient Assisted Living (AAL) technologies are being developed which could assist older people to live healthy and active lives. These technologies have been used to monitor people's daily exercises, consumption of calories and sleep patterns, and to provide coaching interventions to foster positive behaviour. Speech and audio processing can be used to complement such AAL technologies to inform interventions for healthy ageing by analysing acoustic data captured in the user's home. In a recent study [10], we proposed a novel 'affective behaviour representation' feature vector for Alzheimer's dementia recognition using the EmoDB dataset. Results show that classification models based solely on affective behaviour (trained on German language EmoDB data [11]) attain 63.42% detection accuracy on the ADReSS dataset (a subset of Pitt corpora in English language) for Alzheimer's dementia recognition[12].

Automatic identification of emotions in speech is a challenging task, and identifying relevant acoustic features and systematic comparative evaluations have been difficult [13]. In 2016, eGeMAPs [14] was proposed as a synthesis of features extensively used in the literature, according to their theoretical significance and potential to reflect the affective processes that underlie displays of emotion through voice. This feature set aimed to provide a common ground of emotion-related speech features, and has since become a *de-facto* standard. The set of target emotions has mostly been fixed around the 'Big Six' (see Section 3.1), and evaluations are frequently performed on a number of publicly available corpora. In the health domain, feature selection methods for speech processing have been applied to determine the most discriminative features in support of automation efforts. Multimodal emotion recognition has also received increasing attention in the past few years [15, 16], involving analysis of facial, speech, body movements and biometric information [17, 18].

Several studies [19, 17, 20, 21] extract audio features through standard presets, including IS10, GeMAPS, eGeMAPS, and emobase. While the accuracy levels attained by classifiers using those feature sets show promise, there is room for improvement. Despite not figuring among those commonly used sets, the MRCG feature set has been used successfully recently for voice activity detection, attitude recognition and dementia speech recognition. To the best of our knowledge, MRCG and scalogram derived features have not been evaluated systematically for emotion recognition tasks using datasets from different languages. These features are computed using different statistical functionals at the utterance level rather than at frame level, which makes it impossible to extract content information through, for instance, synthesis of speech from the extracted features or automatic transcription [22].

We have developed a low-cost system (hardware shown in Figure 1 [23]) which can extract audio features upon detection of voice activity, and store audio features for further processing while protecting the spoken content. These privacy preserving features are being tested in the context of a larger project which includes health and well-being monitoring and coaching [23]. This presents a comprehensive evaluation of different acoustic feature sets, including, scalograms and the recently proposed MRCG feature set extracted from three different types of languages (English, German and Italian) to inform the design of such systems. It also extends our previous work [24, 25, 26], where we evaluated the MRCG features on attitude and Alzheimer's dementia speech recognition tasks and the capacity of our privacy-based feature extraction system [23] to

http://dx.doi.org/10.21437/Interspeech.2021-1761

the next level, namely automatic detection of emotions using resource constrained devices.

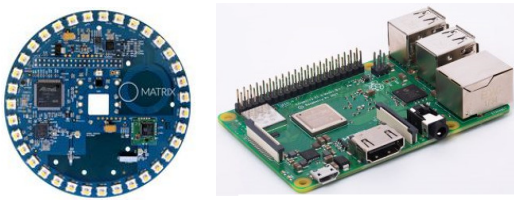

Figure 1: *Matrix Creator and Raspberry Pi 3 B+*

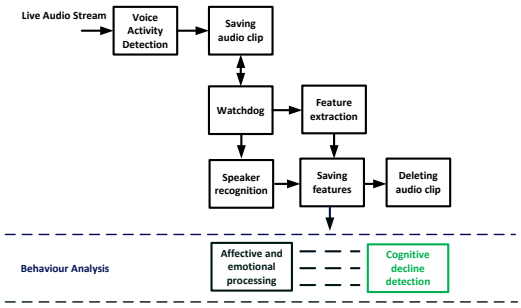

Figure 2: *System architecture, where the affective and emotional processing module will provide input to the cognitive decline recognition module.*

2. Emotion Recognition System

This section describes hardware and software components of the system used to extract acoustic features and recognised emotions while preserving the user's spoken content privacy. The system's architecture is shown in Figure 2.

2.1. Hardware Components

The hardware consists of a Matrix Creator board, consisting of a microphone array, an inertial measurement unit, and several other sensors, mounted on a Raspberry Pi 3 B+, as shown in Figure 1. This setup is meant to be installed in a room where social activity and dialogue interaction occurs frequently, such as a dinning room or a sitting room.

2.2. Software Components

For voice activity detection, we employed the Auditok[1] Python binding. Based on watchdog[2] input, the OpenSMILE [27] toolkit and a user recognition module are used to process the audio file and save the speech features in the attribute-relation file format (ARFF). The user recognition module is trained by registering each user [28]. The extracted acoustic features are then processed by a machine learning model to identify the emotion.

3. Experimentation

We conducted an evaluation of different datasets and acoustic features in order to inform the development of the emotion recognition module for the above described hardware.

[1]https://pypi.org/project/auditok/ – accessed April 2019

[2]https://github.com/gorakhargosh/watchdog – accessed April 2019

Table 1: *Distribution of recordings across emotion categories.*

Dataset	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Boredom
EmoDB	79	127	46	69	71	62	-	81
SAVEE	120	60	60	60	60	60	60	-
EMOVO	84	84	84	84	84	84	84	-

3.1. Datasets

Three corpora were selected for their shared characteristics and public availability: EmoDB, SAVEE, and EMOVO. They consist of recorded acted performances, annotated using the well-known and widely used *Big Six* emotion categories: anger, disgust, fear, happiness, sadness, surprise + neutral, except in the older EmoDB data set where boredom was used instead of surprise. Their characteristics are summarised in Table 1 and briefly described below.

The Berlin Database of Emotional Speech, EmoDB [11] features 535 acted emotions in German, based on utterances carrying no emotional bias. The data were recorded in a controlled environment, resulting in high quality recordings. Actors were allowed to move freely around the microphones, which affected absolute signal intensity. The quality of the dataset was evaluated by perception tests carried out by 20 human participants for recognition rate and naturalness. Only recordings with average recognition rates of at least 80% and naturalness rates of at least 60% were included.

The Surrey Audio-Visual Expressed Emotion, SAVEE, [29] is an audio-visual dataset consisting of 480 British English utterances. Each actor did 15 recordings per emotion (3 common, 2 emotion specific, and 10 generic sentences different for each emotion) and 30 neutral recordings (the 3 common and every emotion specific sentences). No limitations regarding audio features are explicitly stated in the description of the data set. A qualitative evaluation of the database was run as a perception tests by 10 human subjects. The mean classification accuracy for the audio modality was 66.5%, 88% for the visual modality, and 91.8% for the combined audio-visual modalities.

The Italian Emotional Speech Database, EMOVO) [30] features recorded emotions from acted performances by 6 persons, based on both semantically neutral and nonsense sentences. Actors were allowed to move freely, affecting absolute signal intensity. A qualitative evaluation was performed using a discrimination test. Two phrases were selected and, for each, 12 subjects had to choose between two proposed emotions. The mean accuracy for the test was about 80%.

3.2. Acoustic Feature Extraction

The standard acoustic feature sets (emobase, ComParE, ans eGeMAPS) were extracted using the openSMILE v2.1 toolkit, which is widely used in emotion recognition from speech. The emobase feature set [31] contains Mel-Frequency Cepstral Coefficients (MFCC), voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives, and statistical functionals, resulting in a total of 988 features for every speech utterance. ComParE [27] features consists of energy, spectral, MFCC, and voicing related Low-Level Descriptors, including logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, yielding a total of 6,373 features per speech utterance. The *eGeMAPS* feature set [14] resulted from an attempt to reduce those somewhat unwieldy feature sets to a basic set of acoustic features based on

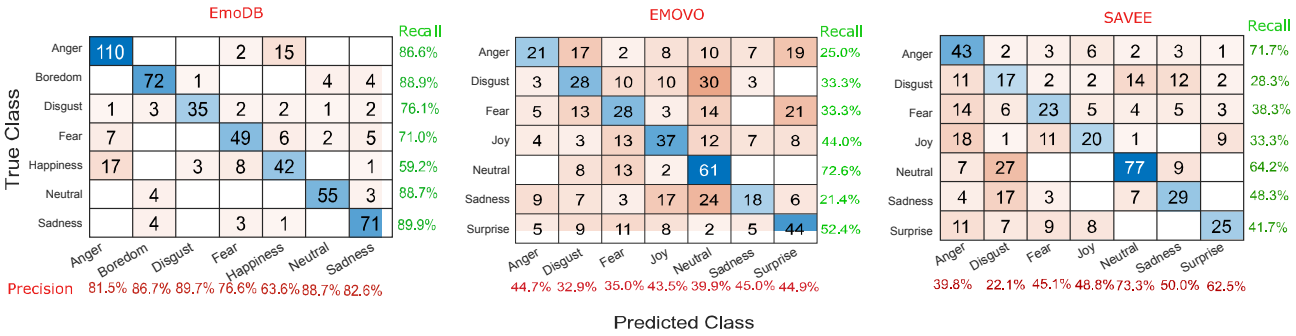

Figure 3: *Confusion matrices of the best results for each dataset.*

their potential to detect physiological changes in voice production, as well as their theoretical significance and proven usefulness in previous studies. It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, totalling 88 features per speech utterance.

In addition to these standard sets, we extracted MRCG and scalogram features for assessment and comparison. MRCG features were originally proposed by Chen et al. [32] and have since been used in speech related applications such as voice activity detection [33] speech separation [32], and more recently for attitude [25] and dementia speech [24] recognition. MRCG features are based on cochleagrams [34], which are generated by applying the gammatone filter to the audio signal, decomposing it in the frequency domain to mimic the human auditory filters. MRCG uses the time-frequency representation to encode the multi-resolution power distribution of an audio signal. Cocheleagram features were generated at four 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 were extracted from each speech 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 features. Statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) were then applied to those 768 MRCG features producing a total of 6912 features per speech utterance.

The scalogram features were extracted with the continuous wavelet transform (CWT) using filter bank [35]. The absolute value of the CWT coefficient and its first and second order derivatives were calculated. Statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) were then applied to these 150 scalograms and their first and second order derivatives producing a total of 4050 features per speech utterance.

3.3. Classification experiments

The classification experiments were performed using five different classifiers, namely: decision trees (DT, with leaf size optimized through a grid search within a range of 1 to 10), k-nearest neighbour (KNN, where K was chose through grid search from 1 to 10), Random Forest (RF, with 150 trees and leaf size optimized through grid search from 1 to 10), Naive Bayes (NB, with kernel distribution assumption optimized through grid search for kernel smoothing density estimate, multinomial distribution, multivariate multinomial distribution and normal distribution) and support vector machines, SVM (linear kernel, box constraint optimized through grid search between 0.1 and 1.0, and sequential minimal optimization solver chosen empirically among the iterative single data algorithm, L1 soft-margin minimization by quadratic programming and sequential minimal optimization). The prior probabilities of the classifiers were set according to the class distributions.

The classification methods were implemented in MATLAB using the statistics and machine-learning toolbox [35]. Leave-one-subject-out (LOSO) cross-validation was used, and we ensured that the training data did not contain any information of the validation subjects. To assess classification performance, we used the unweighted average recall (UAR) rather than overall accuracy, as the dataset is imbalanced. UAR is the arithmetic mean of recall over all seven classes.

4. Results and Discussion

The overall results for the different acoustic feature sets and their fusion as input to the above described classifiers are shown in Table 2. For EmoDB, the ComParE feature set provides better UAR (79.49%, with SVM) than the other four feature sets, and fusion of feature sets improve the results. For, EMOVO, the eGeMAPS feature set provides the best results (39.46% using RF). Fusion of feature sets, with inclusion of the new MRCG features improved the results even further. For SAVEE, ComParE produces the best results (46.55%, using SVM), and fusion of feature sets results in a decrease of UAR.

These results confirm that a higher UAR can be achieved using the fusion of feature sets on the EMOVO and EmoDB datasets, with performance on SAVEE remaining close to the highest single feature set configuration. Notably, the top performing fusion settings included MRCG, attesting to its potential to complement the paralinguistic feature sets. In fact, if the results are averaged over all classifiers, the MRCG feature set provides a better average UAR than all other feature sets, suggesting that MRCG provides more stable predictions. However, contrary to our expectations, the scalogram features performed poorly, and when fused with other features caused performance to degrade in most cases.

In a previous study [25], we evaluated MRCG for attitude recognition using vlogs data. The MRCG feature set was found to be the most reliable across different classifiers and to improve the performance when fused with emobase, eGeMAPS and ComParE for attitude recognition, in line with our current findings. For further insight into these results, we drew the confusion matrices of the best results, which are shown in Figure 3, with precision and recall figures shown for all the emotion categories. It can be seen that the best recall and precision for all emotion categories are obtained for EmoDB. The EMOVO data set provides better recall for neutral, disgust, happiness (joy)

Table 2: *UAR obtained using different feature sets along with the average UAR (mean) for each feature set over five classifiers.*

DataSet	Features	DT	KNN	NB	RF	SVM	$mean$
EmoDB	emobase	0.5052	0.2460	0.6224	0.6757	0.7308	0.5560
	eGeMAPS	0.4901	0.3885	0.4854	0.6428	0.6858	0.5385
	ComParE	0.5379	0.2281	0.3953	0.7048	**0.7949**	0.5322
	MRCG	0.4816	0.4819	0.6324	0.6319	0.7297	**0.5915**
	Scalogram	0.4091	0.1718	0.1183	0.5435	0.5663	0.3618
	emobase + *eGeMAPS*	0.4838	0.2488	0.5310	0.7043	0.7580	0.5452
	MRCG + Scalogram	0.5172	0.1731	0.1432	0.6164	0.6957	0.4291
	emobase + *eGeMAPS* + *ComParE*	0.5329	0.2281	0.3627	0.7142	0.7856	0.5247
	emobase + *eGeMAPS* + MRCG + Scalogram	0.4689	0.1926	0.1624	0.6434	0.7309	0.4396
	emobase + *eGeMAPS* + *ComParE* + MRCG	0.5115	0.2281	0.3656	0.6950	**0.8005**	0.5201
	emobase + *eGeMAPS* + *ComParE* + Scalogram	0.5023	0.2281	0.1627	0.6949	0.7846	0.4745
	All	0.4924	0.2281	0.1645	0.6899	0.7831	0.4716
EMOVO	emobase	0.2279	0.2007	0.1956	0.3418	0.3520	0.2636
	eGeMAPS	0.3639	0.2670	0.2959	**0.3946**	0.3707	**0.3384**
	ComParE	0.2500	0.1854	0.1956	0.3827	0.3741	0.2776
	MRCG	0.2228	0.2619	0.2942	0.3214	0.3656	0.2932
	Scalogram	0.2500	0.1837	0.1497	0.2959	0.3265	0.2412
	emobase + *eGeMAPS*	0.3112	0.2024	0.1956	0.3588	0.3588	0.2854
	MRCG + Scalogram	0.2279	0.1854	0.1531	0.3095	0.3282	0.2408
	emobase + *eGeMAPS* + *ComParE*	0.2704	0.1854	0.1565	**0.4031**	0.3656	0.2762
	emobase + *eGeMAPS* + MRCG + Scalogram	0.2432	0.1854	0.1446	0.3146	0.3282	0.2432
	emobase + *eGeMAPS* + *ComParE* + MRCG	0.2466	0.1905	0.1633	0.3639	0.3741	0.2677
	emobase + *eGeMAPS* + *ComParE* + Scalogram	0.2619	0.1854	0.1429	0.3707	0.3656	0.2653
	All	0.2619	0.1905	0.1429	0.3469	0.3724	0.2629
SAVEE	emobase	0.3286	0.2631	0.3310	0.3417	0.3905	0.3310
	eGeMAPS	0.3607	0.2131	0.3083	0.3702	0.3893	0.3283
	ComParE	0.3012	0.2167	0.2464	0.3810	**0.4655**	0.3222
	MRCG	0.2702	0.3226	0.4107	0.3893	0.4131	**0.3612**
	Scalogram	0.2643	0.1631	0.1464	0.2893	0.3429	0.2412
	emobase + *eGeMAPS*	0.3071	0.2798	0.3262	0.3762	0.3952	0.3369
	MRCG + Scalogram	0.2548	0.1631	0.1440	0.3571	0.4036	0.2645
	emobase + *eGeMAPS* + *ComParE*	0.3143	0.2167	0.2548	0.3714	0.4262	0.3167
	emobase + *eGeMAPS* + MRCG + Scalogram	0.2607	0.1690	0.1440	0.3738	0.3976	0.2690
	emobase + *eGeMAPS* + *ComParE* + MRCG	0.2905	0.2167	0.2738	0.4000	**0.4345**	0.3231
	emobase + *eGeMAPS* + *ComParE* + Scalogram	0.3310	0.2167	0.1381	0.3417	0.3786	0.2812
	All	0.2940	0.2167	0.1381	0.3881	0.3940	0.2862

and surprise than SAVEE. The SAVEE data set provides better precision for neutral, fear, happiness (joy), sadness, and surprise than EMOVO.

Overall, better results are obtained for EmoDB (80.05%) than for EMOVO (40.31%) and Savee (46.55%). Indeed, fusing the standard feature sets with the new MRCG features and using them as input to an SVM produces human-level classification performance on EmoDB. This could be due to the fact that these feature sets capture complementary information, which combined with SVM's robustness to high dimensionality and redundancy account for the relevant distinctions. Furthermore, as noted before, the annotation quality of EmoDB is likely higher than the other two datasets. EmoDB was was evaluated by 20 human coders with an average recognition rate of 86%, and audio recordings with the inter-coder agreement below 80% were removed. No such measure was taken for EMOVO or SAVEE.

For EMOVO, while the reported gold standard accuracy for the test set is 80% (see Section 3.1), one should note that rather than evaluating the full EMOVO data set, only two phrases were selected, and each coder had to choose only between two proposed emotions rather than seven. The fact that our machine learning approach to EMOVO classification is of a seven-class problem explains the much lower results obtained in our assessment in comparison to human performance. For SAVEE, 10 human subjects evaluated the data set and attained 66.5% accuracy for audio. Our machine learning based models provide promising results as compared to humans subjects. Although they are less accurate than human annotators, we use only acoustic information to automate the process of emotion recognition, while human annotators used both acoustic and linguistic information. By not relying on linguistic information, our method ensures that the content of the user's speech is protected.

5. Conclusion

This study evaluates different acoustic feature sets, including, the scalogram and the recently proposed MRCG feature set extracted from datasets of three different languages. By focusing solely on acoustic features, our classification methods are language-independent, and potentially privacy preserving. The MRCG feature set provides better results on average over a wide range of classification algorithms, including DT, KNN, NB, SVM and RF classifiers. In this sense, MRCG features appear to be more robust than other features. While the ComParE feature set provided better results using the SVM and RF classifiers, that feature set performed poorly with most other classifiers. Fusion of feature sets results in an overall improvement over individual feature sets except for the SAVEE dataset. We intend to employ the methods assessed in this work for emotion and mental health status assessment using data collected with the device described in section 2.1 in combination with the Tiny machine learning (TinyML) [36] framework for low-resource devices.

6. Acknowledgements

This research is funded by the European Union's Horizon 2020 research programme, under grant agreement 769661, SAAM (Supporting Active Ageing through Multimodal coaching).

7. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] A. Schuller, Björnand Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9–10, pp. 1062–1087, Nov. 2011.

[3] P. Ryan, S. Luz, P. Albert, C. Vogel, C. Normand, and G. Elwyn, "Using artificial intelligence to assess clinicians' communication skills," *BMJ*, vol. 364, 2019.

[4] N. S. Consedine and J. T. Moskowitz, "The role of discrete emotions in health outcomes: A critical review," *Applied and Preventive Psychology*, vol. 12, no. 2, pp. 59–75, Nov. 2007.

[5] J. E. Dimsdale, "Psychological stress and cardiovascular disease," *Journal of the American College of Cardiology*, vol. 51, no. 13, pp. 1237–1246, 2008.

[6] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: the continuous audiovisual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[7] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Appl.*, vol. 40, no. 16, pp. 6351–6358, 2013.

[8] L. Y. Mano, Faiçal *et al.*, "Exploiting IoT technologies for enhancing Health Smart Homes through patient identification and emotion recognition," *Computer Communications*, vol. 89, pp. 178–190, 2016.

[9] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, 2020.

[10] F. Haider, S. de la Fuente, P. Albert, and S. Luz, "Affective speech for Alzheimer's dementia recognition," in *LREC 2020 Language Resources and Evaluation Conference 11-16 May 2020*, 2020, pp. 67–73.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conf. on Speech Comm. and Techn.*, 2005.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176.

[13] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. on Aff. Comp.*, vol. 7, no. 2, pp. 190–202, 2016.

[15] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "EmotiW 2018: Audio-video, student engagement and group-level affect prediction," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 653–656.

[16] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: EmotiW 5.0," in *Procs ICMI'17*. NY, USA: ACM, 2017, pp. 524–528.

[17] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of ICMI'17*. New York, USA: ACM, 2017, pp. 553–560.

[18] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *2011 5th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2011, pp. 667–670.

[19] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of ICMI'17*. New York, USA: ACM, 2017, p. 569–576.

[20] S. Wang, W. Wang, J. Zhao, S. Chen, Q. Jin, S. Zhang, and Y. Qin, "Emotion recognition with multimodal features and temporal models," in *Proceedings of ICMI'17*. New York, USA: ACM, 2017, pp. 598–602.

[21] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings ICMI'17*. New York, USA: ACM, 2017, pp. 577–582.

[22] L. Lajmi, "An improved packet loss recovery of audio signals based on frequency tracking," *Journal of the Audio Engineering Society*, vol. 66, no. 9, pp. 680–689, 2018.

[23] F. Haider and S. Luz, "A system for real-time privacy preserving data collection for ambient assisted living." in *INTERSPEECH*, 2019, pp. 2374–2375.

[24] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[25] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagram features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3737–3741.

[26] 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.

[27] F. Eyben, F. Weninger, F. Groß, and B. 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.

[28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[29] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Norwich, UK*, Sept. 2009.

[30] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an italian emotional speech database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, 2014, pp. 3501–3504.

[31] F. Eyben, M. Wöllmer, and B. 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.

[32] J. Chen, Y. Wang, and D. 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.

[33] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, 2018.

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

[35] MATLAB, *version 9.6 (R2019a)*. Natick, Massachusetts: The MathWorks Inc., 2019.

[36] C. R. Banbury, V. Janapa Reddi, M. Lam, W. Fu, A. Fazel, J. Holleman, X. Huang, R. Hurtado, D. Kanter, A. Lokhmotov *et al.*, "Benchmarking tinyml systems: Challenges and direction," *arXiv e-prints*, pp. arXiv–2003, 2020.