User Identity Protection in Automatic Emotion Recognition through Disguised Speech

Citation for published version:
Haider, F, Albert, P & Luz, S 2021, 'User Identity Protection in Automatic Emotion Recognition through Disguised Speech', *AI*, vol. 2, no. 4, pp. 636-649. https://doi.org/10.3390/ai2040038

Digital Object Identifier (DOI):
10.3390/ai2040038

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
AI

Article

User identity protection in automatic emotion recognition through disguised speech

Fasih Haider *, **Pierre Albert** and **Saturnino Luz**

Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh EH16 4UX, UK; {Fasih.Haider, Pierre.Albert, S.Luz}@ed.ac.uk
* Correspondence: Fasih.Haider@ed.ac.uk

Abstract: Ambient Assisted Living (AAL) technologies are being developed which could assist elderly 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 analyzing speech data captured in the user's home. However, the collection of data in home settings presents acute privacy protection challenges. To address this issue, we propose a low cost system for recording disguised speech signals which can protect user identity by using pitch shifting. The disguised speech so recorded can then be used for training machine learning models for affective behaviour monitoring. Affective behaviour could provide an indicator of the onset of mental health issues such as depression and cognitive impairment, and help develop clinical tools for automatically detecting and monitoring disease progression. In this article, acoustic features extracted from the non-disguised and disguised speech are evaluated in an affect recognition task using six different machine learning classification methods. The results of transfer learning from non-disguised to disguised speech are also demonstrated. We have identified sets of acoustic features which are not affected by the pitch shifting algorithm and also evaluated them in affect recognition. We found that while the non-disguised speech signal gives the best unweighted average recall (UAR) of 80.01% the disguised speech signal only causes a slight degradation in performance, reaching 76.29% UAR. The transfer learning from non-disguised to disguised speech results in a greater drop in UAR (65.13%). However, feature selection improves the UAR (68.32%). This work forms part of a large project which includes health and wellbeing monitoring and coaching.

Keywords: privacy preservation; affect recognition; health technologies; emotion recognition; ambient assisted living; social signal processing

Citation: . *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Health and wellbeing monitoring using Ambient Assisted Living (AAL) technologies involves developing systems for automatically detecting and tracking a number of events that might require attention or coaching. In the SAAM project [1], we are employing AAL technologies to analyse activities and health status of older people living on their own or in assisted care settings, and to provide them with personalised multimodal coaching. Such activities and status include mobility, sleep, social activity, air quality, cardiovascular health, diet [2], emotions [3] and cognitive status [4]. While most of these signals are tracked through specialized hardware, audio and speech are ubiquitous sources of data which could also be explored in these contexts. Speech quality and activity, in particular, closely reflect health and wellbeing. We have explored the potential of speech analysis for automatically recognizing emotions [3], cognitive difficulties [4] and eating-related events [2] in the SAAM AAL environment [5]. AAL technologies and coaching systems such as SAAM, which focus on monitoring of everyday activities, can benefit from recognition of these audio events in characterizing contextual information against which other monitoring signals can be interpreted. However user privacy remains one of the major challenges in collecting audio data in home environments for the development of health monitoring technology.

1.1. Mental Health and Affective Speech

The literature suggests that older people with cognitive impairment have difficulty accessing semantic information [6]. Since successful communication is essential for meaningful social interaction, this takes a toll on the patients' and their carers' wellbeing. This has an impact on the emotional life of these people. Speech monitoring for mood and cognitive changes may help inform interventions targeted at alleviating such impacts.

In addition to their role in cognition [7], the expression of emotions and their recognition are key aspects of communication [8]. Emotional information can be conveyed in different ways, from explicit facial and verbal expression (e.g. smile, pout, happy statement) to more subtle non-verbal cues, such as intonation, modulation of vocal pitch and loudness of emotional expression. These non-verbal cues are generally referred to as emotional prosody.

In a previous study [9], we found that there are differences in automatically inferred affective behaviours regarding expressions of *sadness*, *anger* and *disgust* among people with and without cognitive impairment (Alzheimer's Disease, AD). Although these results need further study, they suggest that speakers with AD exhibit a deficit in the expression of those emotions, reflected on voice volume, speech rate and pitch. The proposed Affective Behaviour Representation (ABR) and emotion classification scores are able to predict cognitive deficit in such situations with an accuracy of 63.42%. However, in that study there was a mismatch between the dataset used to generate the features for recognition (*emoDB* [10]) and the data on which these features were used (*Pitt Corpus* [11]). Thus, prediction accuracy is likely to have been hindered by the facts that (1) the Pitt Corpus was not explicitly designed to elicit emotions, (2) that the two datasets were recorded under different acoustic conditions, (3) that the speakers were selected from different demographics, and (4) that they are in different languages [9].

1.2. Privacy-Concerns Related to Speech

Privacy concerns constitute a major obstacle in developing and deploying digital technologies for monitoring cognitive health. Individual and societal concerns about privacy and data security have been translated in regulations. In the European Union, the GDPR [12] has set new standards for the collection and management of personal information. Speech data is classified as personal data[1]: it can be used to identify age, gender, subject identity and health status [13]. Sensitive data also encompass additional data such as content-free features which could potentially be used for the identification of a person. The potential of such features as biometric markers further widens the importance of their protection. Concern about privacy is shared by users, who are reluctant to consent to being constantly recorded at their homes and/or while speaking through phones or computers. The balance between the benefits from an analysis of spoken interaction is often offset by the associated threat to privacy.

Ethical requirements for health-studies have reflected these changes in regulation. They have raised awareness on the need for careful risk analysis for studies involving the collection and use of speech-related data. In the context of AAL and in-situ studies, speech analysis usually requires sending data over networks with different levels of security and associated risks, setting the additional possibility of a data breach if intercepted and compromised. While the security of the network can be improved by reducing the transit and exposure of sensitive data through a local pre-processing [14,15], the risk posed by the presence of sensitive data remains.

[1] As defined in Art. 4(14) of the GDPR and Article 3(13) Directive 2016/680

A possible way to mitigate these problems is to obfuscate the identity of the user while the data is collected by changing the pitch of their speech [16]. However, changing the signal can also degrade its analysis: pitch shifting disturbs the acoustic patterns of speech which could be indicative of cognitive impairment.

Hence developing a digital technology using acoustic information should take theses issues into account. In this study, we also propose a framework using feature engineering to address the disturbance of acoustic features caused by pith alteration for affect recognition as shown in Figure 5d.

1.3. Speech Disguising

Speech Disguising is a way to alter speech to hide someone's identity [16]. Zheng et al. [17] subjectively analyse the automatic speech disguise technologies i.e. pitch shifting, vocal tract length normalization (VTLN) and voice conversion (VC) using 30 trials. They found that the speech disguise technologies greatly confuse human evaluators, with an equal error rate around random guess (i.e. 50.00 % for pitch shifting, 46.67% for VTLN and 46.67 % for VC).

1.4. Contribution

We have previously developed a low-cost system [15,18] which records content-free, anonymised audio features for automatic analysis. In particular, we extract features such as the *eGeMAPS* set [19] which we have used to detect specific behaviours in the above-mentioned applications [2–4]. However, one of the limitation was that the previous system [15] delete the audio file after extracting the acoustic features from user's speech. It could work if the emotion is self-reported by user, and we do not have a plan to evaluate the new features (i.e. going to be proposed in future), but not for situations where other humans needs to annotate the audio files with emotions to generate data for machine learning model training. So that preserving audio file is also important while preserving privacy. While speech disguising technologies could help preserve the user's privacy to some extend, a question arises: "what are the effects of speech disguising on acoustic information for emotion recognition"? In this study, we extend our previous work and propose to collect the disguised speech by altering the pitch of the speech signal to protect the identity of a user for development and deployment of machine learning based application. For testing (i.e. deployment), this approach also guarantees the user's spoken content privacy in addition to identity protection. This is because the acoustic features are computed using different statistical functionals at the utterance level rather than at frame level, which makes it impossible to extract or re-build content information through, for instance, synthesis of speech from the extracted features or automatic transcription [20].

To the authors' best knowledge, this is the first study and evaluation of disguised speech for the development and deployment of affect recognition technologies based on acoustic features. Hence, the contributions of this article are:

- Identification of acoustic features which are not affected by disguising speech;
- Evaluation of acoustic features extracted from the disguised speech for affect recognition, and comparison with features extracted from non-disguised speech;
- Demonstration of transfer-learning of acoustic features from non-disguised speech to disguised speech for affect recognition, and analysis of their generalisability.

2. Materials and Methods

This section describes the system and algorithms which have been used for proposing emotion recognition using disguised speech.

2.1. Emotion Recognition System

This section describes hardware and software components of the system used to extract acoustic features and collect disguised speech. The collected disguised speech could be presented to human annotators (e.g. crowd-sourced annotation i.e labelling stage) for annotation of emotions. The system's architecture is shown in Figure 1 where the voice activity detection module detects audio segments based on energy of audio signal. After that, we use pitch shifting algorithm [21] for speech disguising and saves the audio segments. Later, we extract acoustic features using openSMILE [22] and train machine learning models (development module) for emotion recognition. At the end, we test the machine learning model (affective and emotional processing module).

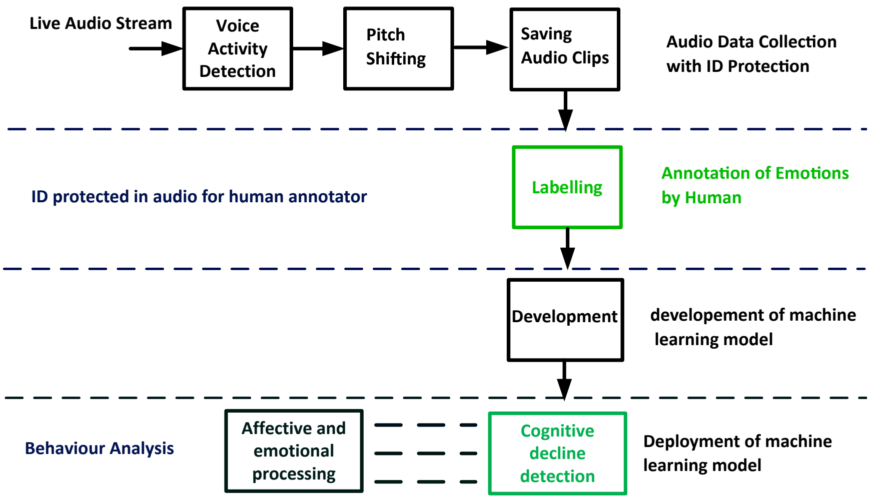


Figure 1. Proposed approach: the affective and emotional processing module will provide input to the cognitive decline recognition module. The 'labelling' and 'cognitive decline detection' are not part of this study. The pitch shifting parameters are only known to and set by the data collection technician and/or user. The Human annotator doesn't have that information.

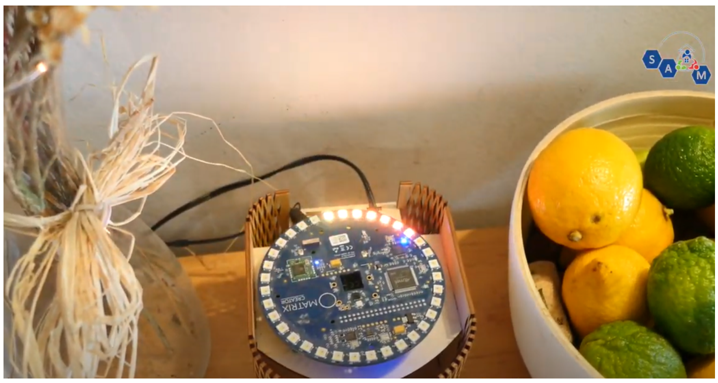


Figure 2. Matrix Creator and Raspberry Pi 3 B+

2.1.1. Hardware Components

The hardware consists of a Matrix Creator board, constituted of a microphone array, an inertial measurement unit, and several other sensors, mounted on a Raspberry Pi 3 B+, as shown in Figure 2. 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.

2.1.2. Software Components

For voice activity detection, we employed the Auditok[2] Python binding. As changes are detected on disk due voice recording (using the watchdog library[3]) the OpenSMILE [23] toolkit processes the audio file of disguised speech and saves the speech features in the attribute-relation file format (ARFF). The extracted acoustic features are then processed by a machine learning model for emotion recognition.

2.2. Data sets

The Berlin Database of Emotional Speech (EmoDB) corpus [10] is a data set commonly used in the automatic emotion recognition literature. It features 535 acted emotions in German (5 male and 5 females), based on utterances carrying no emotional bias. The corpus was recorded in a controlled environment resulting in high quality recordings. Actors were allowed to move freely around the microphones, which affected absolute signal intensity. In addition to the emotion, each recording was labelled with phonetic transcription using the SAMPA phonetic alphabet, emotional characteristics of the voice, segmentation of the syllables, and stress. The quality of the data set was evaluated by perception tests carried out by 20 human participants. In a first recognition test, subjects listened to a recording once before assigning one of the available categories, achieving an average recognition rate of 86%. A second naturalness test was performed. Documents achieving a recognition rate lower than 80% or a naturalness rate lower than 60% were discarded from the main corpus, reducing the corpus to 535 recordings from the original 800. The data sets is annotated for 6+1 emotions: anger, disgust, fear, joy (happiness), sadness, boredom + neutral.

2.3. Identity Protection

To disguise the identify of the subjects, we apply pitch shifting algorithm while maintaining the duration of speech signal using Praat [21]. The audio data with identity protection along with script for pitch shifting is made available through our git repository[4]. We have used a factor of 2 for pitch shifting with time step of 0.01 seconds, minimum pitch of 75 Hz, and maximum pitch of 600 Hz. The pitch shifting parameters are only known to and set by the data collection technician and/or user. The Human annotator does not have that information. An example of non-disguised and disguised audio segment (i.e. spectrogram representation) is shown in Figure 3 and 4 respectively, where the durations of the non-disguised and disguised speech are the same.

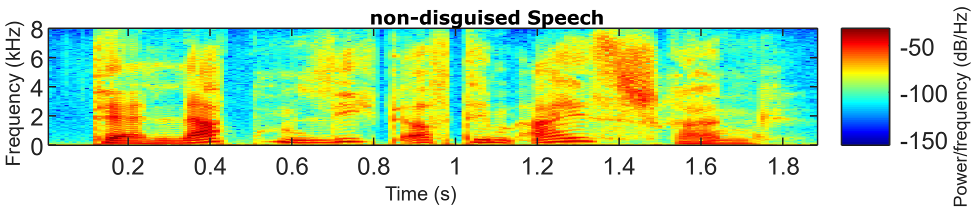


Figure 3. An example of a speech utterance's spectrogram from the EmoDB dataset of a male subject.

2.4. Acoustic Features

Acoustic feature extraction was performed on the non-disguised and disguised speech segments using the openSMILE v2.1 toolkit which is a "source-available" software suite for automatic extraction of features from speech, widely used for emotion and affect recognition in speech [24]. The extracted features are also made available through

[2] https://pypi.org/project/auditok/ – accessed April 2021
[3] https://github.com/gorakhargosh/watchdog – accessed April 2019
[4] https://git.ecdf.ed.ac.uk/fhaider/pitchshifting4affectrecogntion

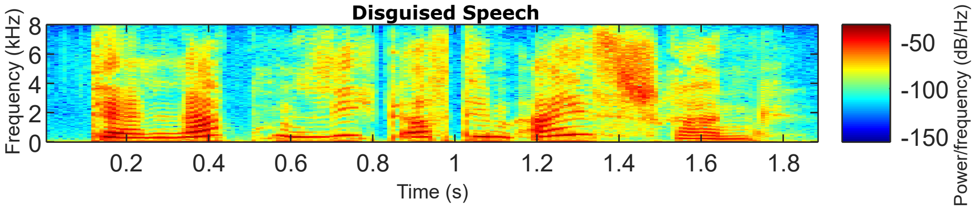


Figure 4. An example of a speech utterance's spectrogram from the EmoDB dataset of a male subject after applying pitch shifting algorithm for identity protection.

the above mentioned git repository. The following is a brief description of the acoustic feature sets used in the experiments described in this paper:

2.4.1. emobase:

This feature set contains the mel-frequency cepstral coefficients (MFCC), voice quality, fundamental frequency (F0), F0 envelope, line spectral pairs (LSP) and intensity features with their first and second order derivatives. Several statistical functions are applied to these features, resulting in a total of 988 features for every speech segment [24].

2.4.2. ComParE:

The *ComParE 2013* [23] feature set includes energy, spectral, MFCC, 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. Statistical functionals are also computed, bringing the total to 6,373 features.

2.4.3. eGeMAPS:

The *eGeMAPS* [19] feature set resulted from an attempt to reduce the somewhat unwieldy feature sets above to a reduced set of acoustic features based on their potential to detect physiological changes in voice production, as well as theoretical significance and proven usefulness in previous studies [22]. 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, for a total of 88 features per speech segment.

2.5. Statistical Analysis

To investigate the possible differences in acoustic characteristics between the non-disguised and disguised speech signals, we first performed a normality test using the one-sample Kolmogorov-Smirnov procedure. This test showed that the data (i.e. acoustic features) follow a normal distribution ($p < 0.001$). We then performed a t-test between the acoustic features extracted from the non-disguised speech signals and the acoustic features extracted from the disguised speech signal. We observed the following:

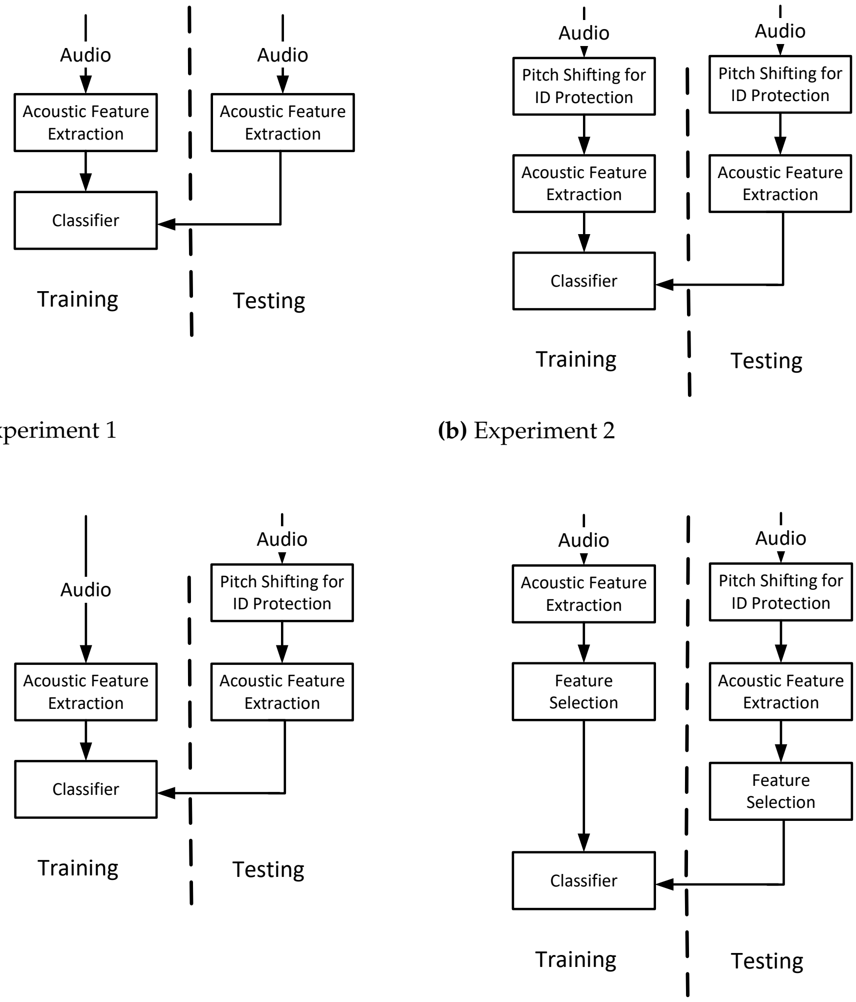
1. for the emobase feature set, there are 257 features out of 988 for which no statistically significant differences ($p > 0.05$) between the non-disguised and disguised speech signals were found. Parts of different functional of Mfcc, fftMag, ZCR, energy, loudness and intensity are not affected by the speech alteration.
2. For the ComParE feature set, we found that 2491 features out of 6373 show no statistically significant differences ($p > 0.05$) between non-disguised and disguised speech signals. Some mfcc, fftMag, audiospec, HNR, ZCR, energy, RASTA, jitter and shimmer functionals are not affected by the speech alteration procedure. The full lists of emobase and ComParE features tested is available through the above mentioned git repository.

3. For the eGeMAPS feature set, we have noted that there are 24 features out of 88 which have no statistically significant differences ($p > 0.05$). The full list of those features is shown below:
   - $F0semitoneFrom27.5Hz\_sma3nz\_pctlrange0-2$
   - $F0semitoneFrom27.5Hz\_sma3nz\_meanRisingSlope$
   - $F0semitoneFrom27.5Hz\_sma3nz\_stddevRisingSlope$
   - $F0semitoneFrom27.5Hz\_sma3nz\_stddevFallingSlope$
   - $loudness\_sma3\_meanRisingSlope$
   - $spectralFlux\_sma3\_stddevNorm$
   - $mfcc1\_sma3\_stddevNorm$
   - $mfcc2\_sma3\_stddevNorm$
   - $mfcc3\_sma3\_stddevNorm$
   - $logRelF0-H1-H2\_sma3nz\_stddevNorm$
   - $logRelF0-H1-A3\_sma3nz\_stddevNorm$
   - $alphaRatioV\_sma3nz\_amean$
   - $alphaRatioV\_sma3nz\_stddevNorm$
   - $hammarbergIndexV\_sma3nz\_amean$
   - $slopeV0-500\_sma3nz\_stddevNorm$
   - $slopeV500-1500\_sma3nz\_stddevNorm$
   - $spectralFluxV\_sma3nz\_stddevNorm$
   - $mfcc1V_sma3nz\_stddevNorm$
   - $mfcc2V_sma3nz\_stddevNorm$
   - $mfcc3V_sma3nz\_stddevNorm$
   - $mfcc4V_sma3nz\_stddevNorm$
   - $loudnessPeaksPerSec$
   - $MeanUnvoicedSegmentLength$
   - $StddevUnvoicedSegmentLength$

2.6. Classification Methods

The classification experiments were performed using six different methods, namely decision trees (DT, where the leaf size is optimized through a grid search within a range of 1 to 20), nearest neighbour (KNN, where K parameter is optimized through a grid search within a range of 1 to 10), linear discriminant analysis (LDA), random forest (RF, with 1500 trees, where leaf size is optimized through a grid search within a range of 1 to 20), Naive Bayes (NB, with kernel distribution assumption optimized through a grid search for kernel smoothing density estimate, Multinomial distribution, Multivariate multinomial distribution and Normal distribution) and support vector machines: SVM, with a linear kernel (optimized by trying different kernel function i.e., linear, Gaussian, RBF and polynomial) with box constraint optimized by trying a grid search between 0.1 to 1.0, and sequential minimal optimization solver (optimized by trying different solvers i.e., iterative single data algorithm, L1 soft-margin minimization by quadratic programming and sequential minimal optimization). The prior-probabilities of the classifiers are set according to the class distributions.

The classification methods are implemented in MATLAB (http://uk.mathworks.com/products/matlab/ (December 2020)) using the statistics and machine-learning toolbox. The classifier hyper-parameters maximum ranges (such as $K = 10$) are set through trial and error. A leave-one-subject-out (LOSO) cross-validation setting was adopted, where the training data does not contain any information of the validation subjects. To assess the classification results, we used the Unweighted Average Recall (UAR) instead of overall accuracy as the dataset is imbalanced. The unweighted average recall is the arithmetic mean of recall for all seven classes.



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

(d) Experiment 4

Figure 5. Affect recognition system: Machine learning model training and testing where testing is performed in leave one subject out cross-validation settings.

3. Experimentation

This section describes the experiments and data partition to evaluate the proposed frameworks as shown in Figure 5.

3.1. Experiment 1

In this experiment, we extracted acoustic features over the non-disguised audio data. Later we trained the machine learning models for classification purpose. The validation is performed in leave-one subject out cross-validation setting as shown in Figure 5a.

3.2. Experiment 2

In this experiment, we extracted acoustic features over the transformed audio data where we hid the identity of a subject using pitch shifting algorithm. Later we trained the machine learning models for classification purpose. The validation is performed in leave-one subject out cross-validation setting as shown in Figure 5b.

3.3. Experiment 3

In this experiment, we trained the machine learning models using non-disguised speech and the validation is performed using disguised speech in leave-one subject out cross-validation setting as shown in Figure 5c.

3.4. Experiment 4

This experiment uses the selected acoustic features as described in Section 2.5, we trained the machine learning models using non-disguised speech and the validation is performed using disguised speech in leave-one subject out cross-validation setting as shown in Figure 5d.

4. Results

This section reports the results for the four experiments.

4.1. Experiment 1

The UAR for all feature sets and classification methods is shown in Tables 1. These results indicate that the ComParE feature set (80.01%) provides the best UAR, with the LDA classifier for emotion recognition. The confusion matrix is shown in Figure 6 for further insight (i.e. precision and recall for all 6+1 emotions) into the best result. The results indicate that the SVM provides the best averaged UAR of 73.42% across all the feature sets, and the ComParE feature set (57.76%) provides the best average UAR across the all classifiers.

Table 1: Experiment 1: Affect recognition results without identity protection where training and validation is performed on the non-disguised audio data. The Unweighted Average Recall (UAR%) is reported.

Features	RF	DT	KNN	NB	SVM	LDA	avg.
emobase.	0.6835	0.5052	0.2460	0.6051	0.7308	0.5574	0.5547
ComParE	0.7059	0.5368	0.2281	0.3953	0.7949	**0.8001**	**0.5768**
eGeMAPS	0.7063	0.4918	0.3885	0.4854	0.6858	0.6616	0.5699
avg	0.6986	0.5113	0.2875	0.4953	**0.7372**	0.6730	—

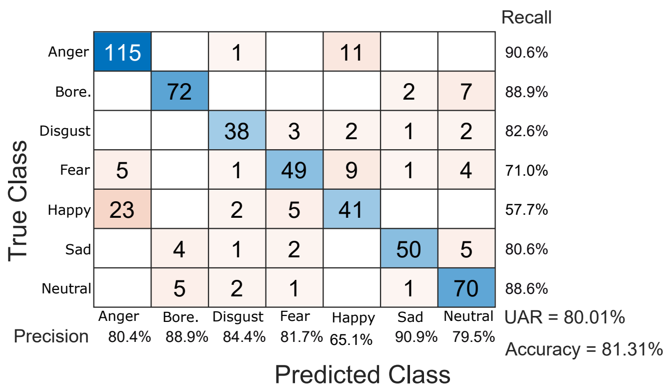


Figure 6. Confusion matrix of the best result for experiment 1 using LDA and Compare Feature set.

4.2. Experiment 2

The UAR for all feature sets and classification methods is shown in Table 2. These results indicate that the combination of the ComParE feature set and LDA again provides the best UAR score (76.29%). The confusion matrix for this is shown in Figure 7 where precision and recall for all 6+1 emotions are listed. In addition, SVM provides the best averaged UAR of 71.68% across all the feature sets and the eGeMAPS feature set (54.78%) provides the best average UAR across the all classifiers.

Table 2: Experiment 2: Affect recognition results with identity protection for training and validation subjects where training and validation is performed on the pitch-shifted audio data. The Unweighted Average Recall (UAR%) is reported.

Features	RF	DT	KNN	NB	SVM	LDA	avg.
emobase.	0.6657	0.4588	0.2759	0.5865	0.7358	0.5417	0.5441
ComParE	0.7063	0.5211	0.2016	0.2440	0.7388	**0.7629**	0.5291
eGeMAPS	0.6335	0.4529	0.3705	0.4818	0.6759	0.6720	**0.5478**
avg	0.6685	0.4776	0.2827	0.4374	**0.7168**	0.6589	—

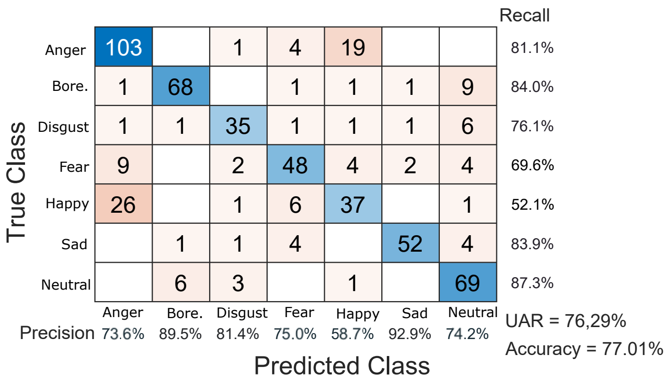


Figure 7. Confusion matrix of the best result for experiment 2 using LDA and Compare Feature set.

4.3. Experiment 3

The results for this experiment are shown in Tables 3. These results indicate that the ComParE feature set again provides the best UAR (65.13%), but this time the RF classifier proves to be the most effective. The confusion matrix is shown in Figure 8 where precision and recall for all 6+1 emotions are listed. RF provides the best averaged UAR of 57.33% across all feature sets, and the emobase feature set yields the best average UAR across all classifiers (45.95%).

Table 3: Experiment 3: Affect recognition results with identity protection for validation subjects, where training is performed on the non-disguised audio data and validation is performed on the pitch-shifted audio data. The Unweighted Average Recall (UAR%) is reported.

Features	RF	DT	KNN	NB	SVM	LDA	avg.
emobase.	0.5624	0.4172	0.2162	0.4838	0.6103	0.4673	**0.4595**
ComParE	**0.6513**	0.4479	0.2161	0.1429	0.1435	0.1344	0.2893
eGeMAPS	0.5062	0.3698	0.2623	0.3470	0.5391	0.1339	0.3597
avg	**0.5733**	0.4116	0.2315	0.3246	0.4310	0.2452	—

4.4. Experiment 4

The resulting UAR scores for all feature sets and classification methods used in this experiment are shown in Tables 4. As before, the ComParE/RF combination achieves the best result (68.32%). The confusion matrix is shown in Figure 9 where precision and recall for all 6+1 emotions are listed. As in the previous experiment, RF provideed the best averaged UAR (60.34%) across all the feature sets, and the emobase feature set yielded the best average UAR across classifiers (48.62%).

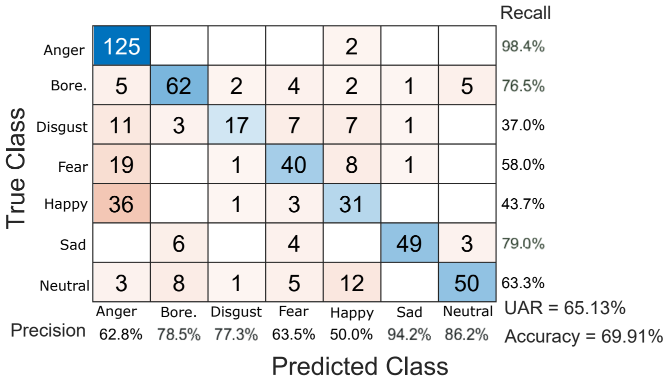


Figure 8. Confusion matrix of the best result for experiment 3 using RF and Compare Feature set.

Table 4: Experiment 4: Affect recognition results with identity protection, where training and validation is performed on selected acoustic features of the non-disguised audio data and validation is performed on the pitch-shifted audio data. The Unweighted Average Recall (UAR%) is reported.

Features	RF	DT	KNN	NB	SVM	LDA	avg.
emobase.	0.5731	0.4121	0.2665	0.5331	0.6250	0.5075	**0.4862**
ComParE	**0.6832**	0.4793	0.2541	0.1429	0.1839	0.1231	0.3111
eGeMAPS	0.5540	0.4467	0.2623	0.3305	0.4988	0.4375	0.4216
avg	**0.6034**	0.4460	0.2610	0.3355	0.4359	0.3560	—

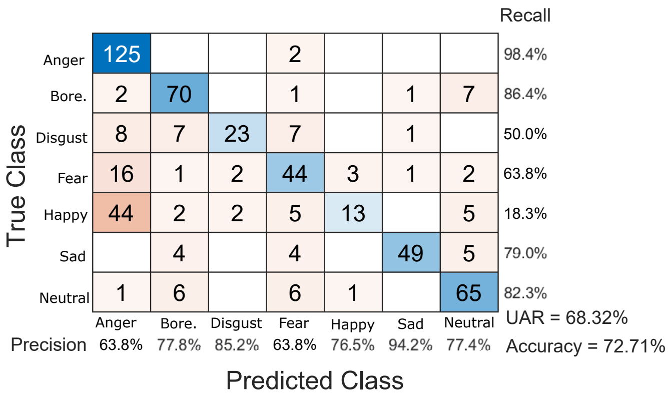


Figure 9. Confusion matrix of the best result for experiment 4 using RF and Compare Feature set.

5. Discussion

The summary of results is shown in Table 5. We note that the non-disguised speech (i.e. Experiment 1) provides the best UAR and accuracy but experiments 4 and 3 provide the best recall for Anger (98.43%) and Sad (83.87%) as shown in bold in Table 5. The 'Happy' emotion is miss-classified as 'Anger' and the miss-classification rate increases for disguised speech experiments, with the worst miss-classification rate occurring when feature selection is performed (Experiment 4). A similar patter is observed for the 'Disgust' category, which exhibits the greates performance degradation in disguised speech. However, feature selection provides better overall UAR (68.32%) than the full feature set (65.13%). Experiment 2 provides better UAR (76.29%) than experiments 3 and 4. One of the advantages of the architecture employed in experiment 2 is that the training and testing are both performed on the disguised speech, with the pitch shifted by the same factor (i.e. 2) for all speech utterances. A variable pitch factor may result in a different outcome.

To better understand the relationship between the experiments, we also plotted the Venn diagram shown in Figure 10. In this diagram, the brown area (labelled "Target")

Table 5: Results Summary: Accuracy (Accu.), Unweighted Average Recall (UAR) and recall of each emotion for the best best results of each experiment

Experiment	Accu.	UAR	Anger	Bore.	Disgust	Fear	Happy	Sad	Neutral
EXP.1	**81.31**	**80.01**	90.55	**88.89**	**82.61**	**71.01**	**57.75**	80.65	88.61
EXP.2	77.01	76.29	81.10	83.95	76.09	69.57	52.11	**83.87**	**87.34**
EXP.3	69.91	65.13	**98.43**	76.54	36.96	57.97	43.66	79.03	63.29
EXP.4	72.71	68.32	**98.43**	86.42	50.00	63.77	18.31	79.03	82.28

represents the annotated labels, the blue area represents the predicted labels of *Experiment 1*, the red area represents the predicted labels of Experiment 2, the green area represents the prediction obtained with the experiment 3, and finally the yellow area represents labels predicted with the experiment 4. The Venn diagrams suggest the information captured by different pitch profiles is not similar, as only 289 out of 535 instances are detected by all the experiments.

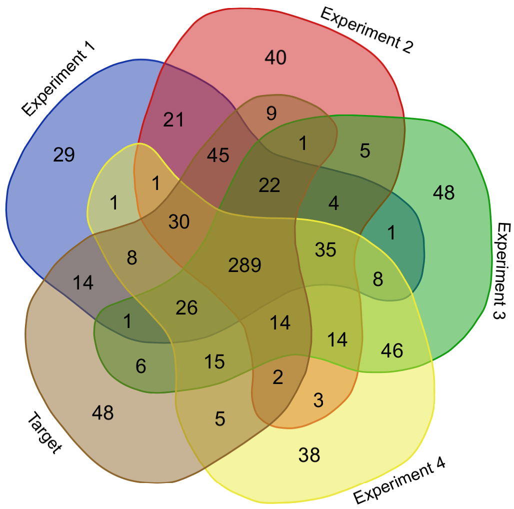


Figure 10. Venn Diagram.

Overall, the experiments show that despite some degradation in prediction accuracy, privacy preservation is compatible with emotion recognition in the settings proposed. We note that while previous studies have proposed affect recognition systems [19,25–28], this study presents an analysis of affect recognition on data that have been transformed to protect the identity of users.

5.1. Limitations

Some limitation of this study which we intend to address in future work include:

- the use of an off-the-shelf pitch shifting method which could have an influence on the performance of affect recognition system;
- the fact that pitch is shifted using a constant factor of 2, whereas a different factor or a variable factor could result in different results;
- feature selection is performed though a statistical approach, and more sophisticated feature selection methods [27] might improve the results further;
- the disguised speech for affect recognition system is evaluated using data which is collected in lab-settings instead of real-world settings;
- the hardware used for the proposed system is a combination of matrix creator and Raspberry Pi 3 B+ with 1.4 GHz 64-bit quad-core processor, which limits one's choice of audio processing and features extraction algorithms due to performance limitations.

6. Conclusion

AAL can benefit from unobtrusive, privacy-preserving systems for gathering and processing of speech at home. This paper described a framework for capturing disguised speech and training machine learning models while protecting the identity of users for automatic wellbeing monitoring tasks, in the context of an AAL-based coaching system for healthy ageing. This study also demonstrated that the acoustic information of disguised speech can be used for emotion recognition. We found that while the non-disguised speech signal gives the best Unweighted Average Recall (UAR) of 80.01% the disguised speech signal only causes a slight degradation of performance, reaching 76.29%. The transfer learning from non-disguised to disguised speech results in a reduction of UAR (65.13%). However, feature selection improves the UAR (68.32%). Privacy protection and preservation in audio and speech can be regarded from different perspectives, including the protection of a person's identity, protection of the content spoken, and protection from inferences one may be able to draw from the characteristics of a person's voice (such as cognitive or emotional status) [29]. A current limitation of the pitch shifting approach is that it only addresses the first (using pitch shifting for identity protection) and second (using statistical functionals of acoustic features instead of content) of these aspects. In future, we aim to address inference protection within a general framework. We also plan to evaluate humans' annotation performance on disguised speech.

Author Contributions: Conceptualization, Fasih Haider, Pierre Albert and Saturnino Luz; Data curation, Fasih Haider and Pierre Albert; Formal analysis, Fasih Haider; Funding acquisition, Saturnino Luz; Investigation, Fasih Haider, Pierre Albert and Saturnino Luz; Methodology, Fasih Haider; Project administration, Saturnino Luz; Software, Fasih Haider; Supervision, Saturnino Luz; Writing – original draft, Fasih Haider; Writing – review & editing, Fasih Haider, Pierre Albert and Saturnino Luz.

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

Conflicts of Interest: the Authors declare no conflict of interest.

References

1. Dimitrov, Y.; Gospodinova, Z.; Žnidaršič, M.; Ženko, B.; Veleva, V.; Miteva, N. Social Activity Modelling and Multimodal Coaching for Active Aging. Procs. of Personalized Coaching for the Wellbeing of an Ageing Society, COACH'2019, 2019.
2. Haider, F.; Pollak, S.; Zarogianni, E.; Luz, S. SAAMEAT: Active Feature Transformation and Selection Methods for the Recognition of User Eating Conditions. Proceedings of the 20th ACM International Conference on Multimodal Interaction; ACM: New York, NY, USA, 2018; ICMI '18, pp. 564–568. doi:10.1145/3242969.3243685.
3. Haider, F.; Luz, S. Attitude Recognition Using Multi-resolution Cochleagram Features. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3737–3741. doi:10.1109/ICASSP.2019.8682974.
4. Luz, S.; la Fuente, S.D. A Method for Analysis of Patient Speech in Dialogue for Dementia Detection. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); Kokkinakis, D., Ed.; European Language Resources Association (ELRA): Paris, France, 2018.
5. Hrovat, A.; Znidarsic, M.; Zenko, B.; Vucnik, M.; Mohorcic, M. Saam: Supporting active ageing-use cases and user-side architecture. 2018 27th European Conference on Networks and Communications (EuCNC), 2018.
6. Bondi, M.W.; Salmon, D.P.; Kaszniak, A.W. The neuropsychology of dementia. In *Neuropsychological assessment of neuropsychiatric disorders.*, 2 ed.; Oxford University Press: New York, NY, US, 1996; pp. 164–199.
7. Hart, R.P.; Kwentus, J.A.; Taylor, J.R.; Harkins, S.W. Rate of forgetting in dementia and depression. *Journal of Consulting and Clinical Psychology* **1987**, *55*, 101–105.
8. Lopes, P.N.; Brackett, M.A.; Nezlek, J.B.; Schütz, A.; Sellin, I.; Salovey, P. Emotional intelligence and social interaction. *Personality and social psychology bulletin* **2004**, *30*, 1018–1034.
9. Haider, F.; De La Fuente Garcia, S.; Albert, P.; Luz, S. Affective Speech for Alzheimer's Dementia Recognition. LREC: Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID); Kokkinakis, D.; Lundholm Fors, K.; Themistocleous, C.; Antonsson, M.; Eckerström, M., Eds. European Language Resources Association (ELRA), 2020, pp. 67–73.
10. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. Proceedings of the ninth European Conference on Speech Communication and Technology, 2005, pp. 1516–1520.

11. Becker, J.; Boller, F.; Lopez, O.; Saxton, J.; McGonigle, K. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology* **1994**, *51*, 585–594.
12. Parliament, T.E.; the Council. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) **2016**.
13. Nautsch, A.; Jiménez, A.; Treiber, A.; Kolberg, J.; Jasserand, C.; Kindt, E.; Delgado, H.; Todisco, M.; Hmani, M.A.; Mtibaa, A.; others. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language* **2019**, *58*, 441–480.
14. Dimitrievski, A.; Zdravevski, E.; Lameski, P.; Trajkovik, V. Addressing Privacy and Security in Connected Health with Fog Computing. Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good. Association for Computing Machinery, 2019, GoodTechs '19, p. 255–260. doi:10.1145/3342428.3342654.
15. Haider, F.; Luz, S. A System for Real-Time Privacy Preserving Data Collection for Ambient Assisted Living. INTERSPEECH, 2019, pp. 2374–2375.
16. Perrot, P.; Aversano, G.; Chollet, G. Voice disguise and automatic detection: review and perspectives. *Progress in nonlinear speech processing* **2007**, pp. 101–117.
17. Zheng, L.; Li, J.; Sun, M.; Zhang, X.; Zheng, T.F. When Automatic Voice Disguise Meets Automatic Speaker Verification. *IEEE Transactions on Information Forensics and Security* **2020**, *16*, 824–837.
18. Haider, F.; Luz, S. Affect Recognition Through Scalogram and Multi-Resolution Cochleagram Features. Proc. Interspeech 2021, 2021, pp. 4478–4482. doi:10.21437/Interspeech.2021-1761.
19. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; others. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* **2016**, *7*, 190–202.
20. Lajmi, L. An Improved Packet Loss Recovery of Audio Signals Based on Frequency Tracking. *Journal of the Audio Engineering Society* **2018**, *66*, 680–689.
21. Boersma, P.; Weenink, D. Praat: doing phonetics by computer [Computer program]. Version 6.0. 37. *URL http://www. praat. org/. Retrieved March* **2018**, *14*, 2018.
22. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; others. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* **2016**, *7*, 190–202.
23. Eyben, F.; Weninger, F.; Groß, F.; Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. Proceedings of the 21st ACM international conference on Multimedia. ACM, Association for Computing Machinery, 2013, pp. 835–838.
24. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia. ACM, Association for Computing Machinery, 2010, pp. 1459–1462.
25. Haider, F.; Luz, S. Attitude recognition using multi-resolution cochleagram features. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3737–3741.
26. Haider, F.; Salim, F.A.; Conlan, O.; Luz, S. An Active Feature Transformation Method for Attitude Recognition of Video Bloggers. INTERSPEECH, 2018, pp. 431–435.
27. Haider, F.; Pollak, S.; Albert, P.; Luz, S. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech & Language* **2020**, p. 101119.
28. Haider, F.; Pollak, S.; Albert, P.; Luz, S. Extracting Audio-Visual Features for Emotion Recognition Through Active Feature Selection. 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2019, pp. 1–5.
29. Pathak, M.A.; Raj, B.; Rane, S.D.; Smaragdis, P. Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise. *IEEE signal processing magazine* **2013**, *30*, 62–74.