Speech Emotion Recognition from Social Media Voice Messages Recorded in the Wild

Lucía Gómez-Zaragozá(✉), Javier Marín-Morales(✉), Elena Parra(✉), Jaime Guixeres(✉), and Mariano Alcañiz(✉)

Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, Valencia, Spain
{lugoza, jamarmo, elparvar, jaiguipr, malcaniz}@i3b.upv.es

Abstract. Speech is the most natural way for human communication, carrying the emotional state of the speaker that plays an important role in social interaction. Currently, many instant messaging apps offer the possibility of exchanging voice audios with other users. As a result, a great amount of voice data is generated every day, representing a new challenging approach for speech emotion recognition in real environments. In this study, we investigated emotion recognition from voice messages recorded in the wild using machine-learning algorithms. Unlike most research in this field, which use databases based on emotions evoked in lab environments, simulated by actors or subjectively selected from radio or TV talks, we created an ecological speech dataset with audios from real WhatsApp conversations of 30 Spanish speakers. Four external evaluators labelled each audio in terms of arousal and valence using the Self-Assessment Manikin (SAM) procedure. Pre-processing techniques were applied to the audios and different time and frequency domain features were extracted. Supervised machine learning classifiers were computed using feature reduction and hyper-parameter tuning in order to recognize the affective state of each voice message. The best recognition rate was obtained with Support Vector Machines, achieving 71.37% along the arousal dimension and 70.73% along the valence dimension. These results support the use of emotion recognition models on daily communication apps, helping to understand social human behavior and their interactions with devices in the real world.

Keywords: Speech emotion recognition · Speech database · Vocal social media

1 Introduction

Speech is the most natural and efficient way of communication for humans. It conveys not only linguistic information but also the emotional state of the speaker, which is a key factor for daily human interactions, as it is interpreted and used by the listener to adapt the behavior in response. Currently, speech emotion recognition (SER) is a growing research area that aims to recognize the emotional state of a speaker from the speech signal. It has potential applications both for the study of human-human communication and human-computer interaction (HCI) [1].

C. Stephanidis and M. Antona (Eds.): HCII 2020, CCIS 1224, pp. 330–336, 2020.
https://doi.org/10.1007/978-3-030-50726-8_43

In today's social media era, instant messaging tools such as WhatsApp or Facebook Messenger have spread worldwide, allowing the users to exchange text, voice, image and video messages. These applications do not only facilitate the communication with relatives and acquaintances but also are trending to compete with face-to-face interactions, especially among younger generations [2]. The communication in instant messaging tools is mainly performed via text or audio. To date, there has been extensive research in the study of emotions in text-based interactions [3], where the lack of non-verbal emotional cues is compensated by using emoticons, letter repetition or typed laughter, among others [4]. However, speech emotion recognition in mobile environment, and particularly in the context of instant messaging tools, is still at an early stage. One likely reason is that the task of recognizing emotions in real-world conditions is still a challenge.

Currently, the main approach for emotion recognition is based on supervised machine learning techniques, in which the database selection is a primary issue. The vast majority of SER research use databases that can be classified in three categories: acted, induced and natural/spontaneous [5]. The first include speeches that are portrayed by professional or semi-professional actors who simulate emotions while pronouncing pre-determined isolated utterances. Induced datasets contain speeches produced in controlled situations designed to elicit a certain emotional state, for example watching a video, listening a story or conducting a guided discussion. The least frequent category are natural speech emotion databases, in which audios are recorded in real-world situations (such as real psychologist interviews) or they are obtained from movies and radio or TV programs (for instance reality shows or talk shows).

The databases described above have several limitations for its application in the recognition of real-life emotions, as described in diverse studies [5–7]. Acted databases are a popular method, as they are easier to create. However, acted emotions differ from natural emotions, tending to be more exaggerated and stereotypical. Induced databases include speeches that are more similar to real expression of emotions, but the methodology used to obtain them has some limits: each subject may react different to the same stimuli and a further subjective evaluation is needed in order to determine the sample's emotion, in addition to the ethical implications of inducing emotion. Regarding spontaneous databases, the recordings usually have conditions such as background noise and overlapping voices that are typical in natural environments, known as in-the-wild settings. Nevertheless, the emotions may not be spontaneous if the subject is aware of being recorded, as an interview or a radio show. In the case of hidden-recording, due to the artificial situation (lab or studio settings), the subject may subconsciously keep their expressions under control or express them in an unnatural way. It also important to note that recordings that are not produced in a conversational context lack some naturalness due to emotions are produced as a response to various situations. Furthermore, similar to induced databases, the samples showing emotional states are subjectively selected by evaluators and the databases involve legal issues and ethical problems that make public distribution difficult.

Therefore, there is a lack of research using databases that include audios that belong to historical private communications, showing the underdevelopment of SER models that can be applied to human-human audio messaging. To our knowledge,

Dai et al. (2015) presented the first suitable speech dataset for emotion recognition on voice instant messaging, consisting of vocal messages from the popular Chinese application WeChat [8]. Since their goal was to study the emotion propagation in a particular group, they collected voice historical data from nine familiar members in the same WeChat group in order to extract personalized features and use them for training a machine learning model. However, it is still a challenge to evaluate datasets with a larger number of audios, subjects and languages.

In this work-in-progress research, we investigated emotion recognition from voice messages using acoustic features and machine-learning algorithms. We collected the audio data from real conversations of 30 Spanish speakers conducted in the popular mobile app WhatsApp, in which the expression of emotions is considered to be more suitable than on other social media platforms [9]. We obtained 12 audios for each of the subjects, with an equal number of positive, neutral and negative valence recordings. Four external evaluators labelled each of the audios in terms of arousal and valence. Thus, we obtained an ecological dataset with audios recorded in the wild, on which we applied speech emotion recognition techniques.

2 Materials and Methods

2.1 Participants

The present study initially included 30 Spanish speakers between the ages of 18 and 55 years old. However, as explained in *Data Collection*, six participants were excluded for the analysis, leaving a total of 24 subjects (62.5% females) of ages (Mean ± SD) 31.7 ± 11.1 with no self-reported speech disorder. All methods and experimental protocols were performed in accordance with the guidelines and regulations of the local ethics committee of the Universitat Politècnica de València.

2.2 Data Collection

The data was collected using an online platform designed ad-hoc. The participants completed the study with their computer, following the instructions given in the platform. Once they accepted the informed consent, the participants answered a sociodemographic questionnaire. Then, they were requested to upload 12 voice messages according to two criteria: the audios should have been sent to other contacts prior to the study and one-third of them should have positive, neutral and negative valence, respectively.

Firstly, an expert manually identified the audios recorded in critical background noise conditions, rejecting from the study 6 participants whose majority of audios presented these states. To avoid any possible bias derived of the self-assessment, the audios were assessed adopting the Self-Assessment Manikin (SAM) procedure [10], which consists of a non-verbal scale based on pictures that measures the valence, arousal and dominance related with an emotional response to a stimuli. Four evaluators used the 5-point SAM scale to rate each audio in terms of positive/high (>0), neutral (=0) and negative/low (<0) valence and arousal respectively. Only those samples in

which three out of four of the labels were in consensus were chosen for the study. To perform valence classification, 188 samples (49.5% positive valence) were considered, excluding neutral audios as an initial simplification. With regard to arousal classification, the data was unbalanced due to the fact that participants chose the audios on the basis of their valence. For this reason, we considered low and neutral arousal recordings as pertaining to the same group, resulting in 234 samples (59.4% high arousal).

2.3 Data Processing

The audio files, collected in .ogg format with sample rates of 41 kHz and 48 kHz, were processed following the pipeline in Fig. 1 in order to obtain two machine learning models for predicting valence and arousal independently. Each step is detailed below.

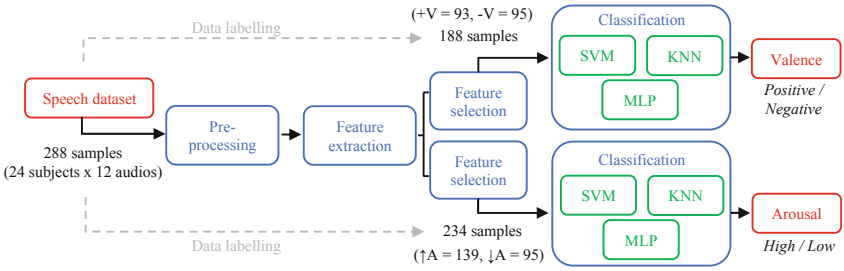

Fig. 1. Pipeline of the proposed speech emotion recognition procedure.

Pre-processing. The audio signals were normalized to range [−1, 1] using the standardisation method and then resampled to 48 kHz.

Feature Extraction. Long-term acoustic features were computed in two stages using the pyAudioAnalysis open source Python library [11]. First, the audio signal was divided into frames of 50 ms with 50% overlap. For each of them, the following features were computed: time domain cues (zero crossing rate, energy and entropy of energy) and frequency domain cues (spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 Mel-Frequency Cepstral Coefficients (MFCCs), 12-element chroma vector and the standard deviation of the 12 chroma coefficients). Long-term features were finally computed as the statistics (mean and standard deviation) of the frame-based features extracted for the whole audio, assuming that their temporal variations carry the emotional content of the recordings.

Feature Selection. Due to the high-dimensional feature space resulting after data processing, random forest-based feature selection was applied in order to avoid overfitting. The algorithm rank the features according to the importance weights extracted from an artificial classification task and one feature is dropped in each iteration. The process continues until only one feature is considered, thus selecting the vocal cues that contain the most relevant emotion information from speech signals.

Classification. The following machine learning algorithms were applied for recognizing the affective state of voice data based on the extracted acoustic features: K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Multilayer Perceptron (MLP). We adopted cross-validation procedure for hyper-parameter tuning and feature selection. Specifically, we applied group k-fold cross-validation (k = 6) so that audios from the same subject were not included both in training and validation set.

3 Results

Table 1 and Table 2 show the performance of the three machine learning models that achieved best results after feature selection and hyper-parameter tuning, in terms of valence and arousal respectively. It includes the accuracy of each model, the true positive rate (TPR), the true negative rate (TNR) and the number of features included in the model (N-features).

Table 1. Best accuracy results for each model in terms of valence.

Model	Accuracy (%)	TPR	TNR	N-features
KNN	68.06	63.02	72.43	57
SVM	70.73	70.34	70.35	5
MLP	62.85	63.74	59.03	39

Table 2. Best accuracy results for each model in terms of arousal.

Model	Accuracy (%)	TPR	TNR	N-features
KNN	67.95	86.33	41.78	23
SVM	71.37	79.59	59.28	6
MLP	65.81	75.87	54.30	78

4 Discussion

One of the most critical factors to create an automatic speech emotion recognition system is database selection. Most previous studies performed SER using speech corpus whose application for real-life emotion recognition is rather limited. Here, we collected the speech dataset from voice messages from real conversations, where the participants were not aware that their audios were going to be part of a study and thus, samples can be considered as natural expressions of emotions. In addition, voice data was originally recorded in the wild so the audios presented background noise and only those subjects whose majority of audios had critical noise conditions were dropped from the study. We performed a comparison of different classification models for valence and arousal recognition from acoustic features extracted from the voice messages.

Different classification algorithms are used in the literature for recognizing the affective state of voice data based on the extracted acoustic features. Particularly, SVM is one of the most widely used methods [6, 7], as the results obtained here seem to support.

The results in Table 1 show that SVM obtained the best recognition rate, achieving 70.73% accuracy in predicting positive or negative valence from the voice messages. Since it uses only five features, it avoids the possibility of overfitting, suggesting a promising result. KNN reached close accuracy, 68.06%, but including a large number of features needed.

Regarding arousal results in Table 2, the 71.37% SVM accuracy also outperformed the other two classification models, using also only six features. However, TNR are in general low, which may be caused by the annotation approach that consider both neutral and low arousal as pertaining to the same group.

However, some limitations need to be considered in this work-in-progress. Firstly, six participants were not included in the analysis due to critical noise conditions, which limited the number of speakers in the dataset. The unbalanced distribution of the audio data in terms of arousal led to a reassignment in the labels that may influenced the results. Another critical factor is the annotation method, which is in general a challenging task, as there are several approaches with respect to various factors: the classification of emotions (categories or dimensions), the emotion unit to label (phonemes, single words, sentences or complete utterances) and the evaluator (familiar members, experts or non-experts subjects).

The results highlight many point that need to be addressed in future research. The number of speakers should be increased, and the influence of the gender need to be considered since it could affect many features. In addition, the implementation of noise reduction techniques is also considered as part of ongoing research to deal with the challenge of recognizing emotions in real-world environments.

5 Conclusion

In this work-in-progress research, we collected our speech database using real voice messages from WhatsApp conversation of Spanish speakers. We emotionally labelled the audio samples in terms of valence and arousal. Global acoustic features were computed for each recording and a comparison of several classification models was performed for both valence and arousal prediction.

Preliminary results support the feasibility of using emotion recognition models on daily communication apps. It may help to understand social human behavior and their interactions with devices in the real world, improving personalization and adaptive interfaces in social networks.

Acknowledgments. This work was funded by the European Commission (H2020-825585 HELIOS).

References

1. Khanna, P., Sasikumar, M.: Recognizing emotions from human speech. In: Pise, S.J. (ed.) Thinkquest 2010, pp. 219–223. Springer, New Delhi (2010). https://doi.org/10.1007/978-81-8489-989-4_40
2. Venter, E.: Challenges for meaningful interpersonal communication in a digital era. HTS Teol. Stud. **75**, 1–6 (2019). https://doi.org/10.4102/hts.v75i1.5339
3. Zucco, C., Calabrese, B., Agapito, G., Guzzi, P., Cannataro, M.: Sentiment analysis for mining texts and social networks data: methods and tools. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **10**, e1333 (2019). https://doi.org/10.1002/widm.1333
4. Sherman, L., Michikyan, M., Greenfield, P.: The effects of text, audio, video, and in-person communication on bonding between friends. Cyberpsychology J. Psychosoc. Res. Cybersp. **7**, Article 1 (2013). https://doi.org/10.5817/cp2013-2-3
5. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. **21**, 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z
6. Akçay, M.B., Oğuz, K.: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. **116**, 56–76 (2020). https://doi.org/10.1016/j.specom.2019.12.001
7. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020
8. Dai, W., Han, D., Dai, Y., Xu, D.: Emotion recognition and affective computing on vocal social media. Inf. Manag. **52**, 777–788 (2015). https://doi.org/10.1016/j.im.2015.02.003
9. Waterloo, S.F., Baumgartner, S.E., Peter, J., Valkenburg, P.M.: Norms of online expressions of emotion: comparing Facebook, Twitter, Instagram, and WhatsApp. New Media Soc. **20**(5), 1813–1831 (2017). https://doi.org/10.1177/1461444817707349
10. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994). https://doi.org/10.1016/0005-7916(94)90063-9
11. Giannakopoulos, T.: pyAudioAnalysis: an open-source python library for audio signal analysis. PLoS One **10**(12), e0144610 (2015). https://doi.org/10.1371/journal.pone.0144610