IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

A proposal for emotion recognition using speech features, transfer learning and convolutional neural networks

Roberto Móstoles[1], *David Griol*[2], *Zoraida Callejas*[2], *Fernando Fernández-Martínez*[3]

[1] Universidad Carlos III de Madrid. Madrid (Spain)
[2] Dept. of Languages and Computer Systems, University of Granada. Granada (Spain)
[3]Speech Technology Group, Center for Information Processing and Telecommunications, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Spain

100346034@alumnos.uc3m.es, dgriol@ugr.es, zoraida@ugr.es, fernando.fernandezm@upm.es

Abstract

In this paper, we present a proposal for emotion recognition using audio speech signal features consisting of two functionally independent systems. First, a voice activity detection module (VAD) acts as a filter prior to the emotion classification task. It extracts features from the input audio and uses a SVM classifier to predict the presence of voice activity. Secondly, the speech emotion classifier (EMO) transforms the power spectrum of the signal to a Mel scale and obtains a vector of its characteristics using a convolutional neural network. Emotion labels are assigned using this vector and a KNN classifier. The RAVDESS dataset has been used for training the models obtaining a maximum accuracy of 93.57% classifying 8 emotions.
Index Terms: speech emotion recognition, human-computer interaction, computational paralinguistics

1. Introduction

When people engage in natural conversational interaction, they convey much more than just the meanings of the words spoken. Their speech also conveys their emotional state and aspects of their personality [1]. Paralanguage refers to properties of the speech signal that can be used, either consciously or subconsciously, to modify meaning and convey emotion. Examples of paralinguistic features include those that accompany and overlay the content of an utterance and modify its meaning, such as pitch, speech rate, voice quality, and loudness, as well as other vocal behaviors, such as sighs, gasps, and laughter. Paralinguistic properties of speech are important in human conversation as they can affect how a listener perceives an utterance. Schuller and Batliner [2] presented a detailed survey of computational approaches to paralinguistics, examining the methods, tools, and techniques used to automatically recognize affect, emotion, and personality in human speech. Berkeham and Oguz have very recently presented a detailed survey of emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers for speech emotion recognition [3].

In this paper we propose a speech emotion recognition approach that consists of two independent modules: VAD and EMO. The Voice Activity Detection (VAD) module takes an audio signal as input, analyzes its features by means of a computationally efficient metric evaluation, and classifies its content using a SVM classifier. The result is a binary classification: 1 if it estimates that the audio contains speech activity, 0 otherwise.

Most techniques for speech emotion recognition use Cepstral coefficients in the Mel Frequencies (MFCC), which compute characteristics of both the semantic content of the audio signal (the encoded information) and the contextual content (the state of the interlocutor). The use of Deep Learning techniques for the classification of these characteristics is popular in the field [4, 5, 6]

The speech segments as detected by the VAD module are forwarded to the emotion classifier module (EMO), which transforms the power spectrum computed from the Short-Time Fourier Transform (STFT) of the signal following the Mel scale and obtains a vector of its characteristics using a Convolutional Neural Network (CNN). Transfer learning techniques have been employed to use a pre-trained CNN as a feature extraction model [7, 8]. The resulting feature vector has been used as input to a K-Nearest-Neighbor (KNN) classifier, which ultimately performs the classification task.

The remainder of the paper is structured as follows. Section 2 describes the datasets used to train and evaluate our proposal for speech emotion recognition. Section 3 presents the architecture for the proposed system. Sections 4 and 5 describe the main two modules (VAD and EMO), their practical implementation and the results of their evaluation. Finally, Section 6 presents the conclusions and guidelines for future work.

2. The datasets

Two datasets have been used, one for the VAD module and the other for the EMO module. The dataset used to develop the VAD module was developed by the authors and consists of over 2 hours of recorded speech and non-speech audios on different environments and under different communication contexts. Audio samples were then labeled manually, and the resulting audio files were split into segments of 200ms [9], which has been considered to be a wide enough time window in which perform reliable feature extraction (smaller windows leads to lack of information when computing features, while bigger windows may comprise speech and non-speech subsegments). Additional audio files were generated from the processed dataset using aggregation operations: the samples of every audio segment were shifted in time, compressed and exposed to noise, saving the result as new audios. This operations helps to extend the set of examples without adding excessive redundancy to the dataset. The final dataset contains 150k labelled entries.

For the EMO module the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) [10] has been used. We have considered 8 emotion categories: calm, happy, sad, angry, fearful, surprise, disgust, and neutral.

This dataset contains labeled files in 3 modality (full AV, video-only and audio-only) and 2 vocal channels (speech and song) from male and female actors. Since our focus was on speech emotion recognition, only audio-only speech samples

10.21437/IberSPEECH.2021-12

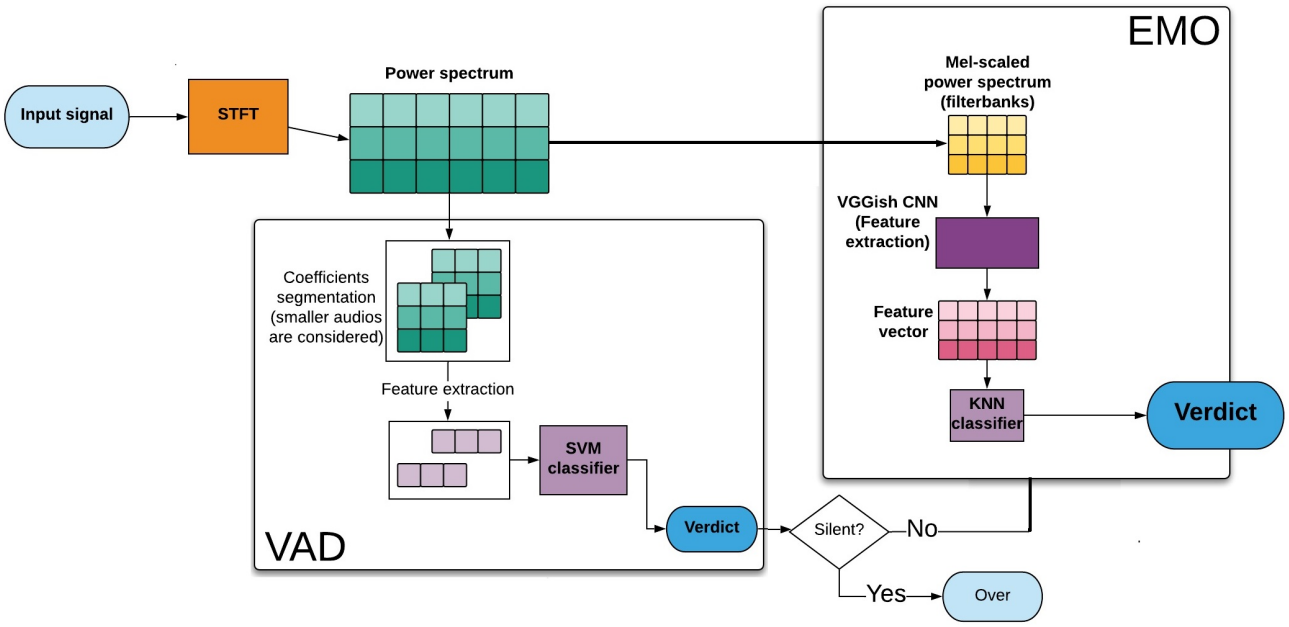

Figure 1: *Proposed architecture for emotion recognition from speech*

have been employed.

The database is gender balanced consisting of 24 professional actors, vocalizing lexicaly-matched statements in a neutral North American accent. Each expression is produced at two levels of emotional intensity (normal and strong intensity), with an additional neutral expression.

The set of 7,356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data.

In order to optimize the training work, the original audios have been segmented into 1s fragments, and each of these segments has been analyzed with the proposed VAD system, eliminating the audios in which there was not speech detected [11, 12, 13]. The audio files in both datasets have a sampling frequency of 16,000 Hz, 32 bits wav-audio format, mono channel.

3. Our proposal

Figure 1 shows the architecture proposed. As it can be observed, the first step consists of obtaining the STFT of the input signal. The STFT operates by splitting the original time signal into multiple smaller segments (frames) and then applying the Fourier Transform on each of them, thus, offering both time and frequency resolution.

The STFT shape and resolution greatly depends on the window features. The implementation used to develop the system allows fine tuning both the window features and shape of the resulting metrics of the STFT analysis, simplifying its usage across the system.

The VAD and EMO modules are described in the following sections. Although they are presented as modules of the proposed architecture, they can also work as functionally independent systems (voice activity detector and speech classifier respectively) that could be plugged into other architectures.

4. The Voice Activity Detection (VAD) module

If the audio segments used do not contain speech information the emotion analysis would be less precise. To avoid this situation, a computationally efficient real time voice activity detection module (VAD) has been implemented and placed as first step before emotion classification. Figure 1 shows the structure of the VAD module: it takes an audio signal as input, frames it into segments of 200 ms, computes the vector of statistical features of each segment and then feeds it to an SVM classifier to generate the final prediction as a binary output (0 if the segment is non-speech or 1 otherwise).

Hit-rate efficiency is achieved by having account of the elements present on each spoken communicative situation: the speaker (different speakers produce audios with different tonal features which are dependent of the individual's anatomy), the speech (the emotional context of the speech can further alter each individual's tonal features) and the environment (noisy conditions lowers the efficiency of the speech detection system).

4.1. Features

The proposed VAD system computes the following statistical metrics:

- **ZCR - Zero-Crossing Rate**: Assuming that the amplitude of a signal is defined in the range [-1, 1], the ZCR coefficient measures the number of times the amplitude changes sign (cross zero) in the time signal. A non-speech signal tends to show a higher ZCR values accounting for the noise being the only element present in the signal: the signal is steadier, and so, oscillates at all times around similar values.
- **HZCRR - High Zero-Crossing Rate Ratio**: While the Zero Crossing Ratio measures the stability of a signal's amplitude over the whole signal, the HZCRR measures the proportion of Zero Crossing Ratio across individual

segments of the signal. Formally is defined as the number of signal segments whose ZCR ratio is above 1.5 fold the average ZCR of the full signal, and represents a localized view of the ZCR.

- **SF - Spectral Flux**: This metric is defined as the mean difference in spectral power between two adjacent frames of the original signal. A higher spectral flux means a higher variance between frames, thus a higher chance for the signal to be speech.
- **STED - Short Time Energy Deviation**: STED measures the variance of the signal level by estimating the difference between the mean and minimum averaged spectral power across the signal frames.
- **SPSD - Short-Term Spectral Power Density**: While the STED coefficient measures the energy difference across time frames, the SPSD follows the same principle applying it to the frequency coefficients of the signal's STFT.
- **BE - Band Entropy**: Entropy measures the amount of information available in a source. The band entropy operates by estimating the entropy across the frequency power coefficients of each frame of the signal's STFT, and then averaging the result by the total number of frames. This feature is computed for multiple frequency bands, thus the original signal must be filtered in order to obtain representations of different frequency regions.
- **BP - Band Periodicity**: This metric is computed measuring the correlation of the waveform of adjacent signal frames. Like band entropy, this feature highly benefits from the multiband signal representation (human voice tends to oscillate in specific frequency bands), thus a filtered signal must be provided to properly compute the metric.

4.2. Classification

The VAD system operates on smaller segments of the original signal, thus the previously described features are computed multiple times for every audio signal that is feed into the system. Once all the feature coefficients are resolved, the VAD system uses the resulting feature vector as the input of a classifier, which ultimately performs the audio activity estimation.

For the classification task two base models have been considered: Random Forest (RF) and Support Vector Machines (SVM) [14, 15].

In order to adjust the models correctly, different critical settings have been considered for each of the classifiers and compared using a cross-validation study with the samples generated. The optimal configurations obtained for the classifiers were: Random Forest (200 trees, quadratic selection of characteristics), linear SVM (C=1000) and polynomial SVM (C=2000).

Figure 2 shows the accuracy results obtained with the selected SVM classifier.

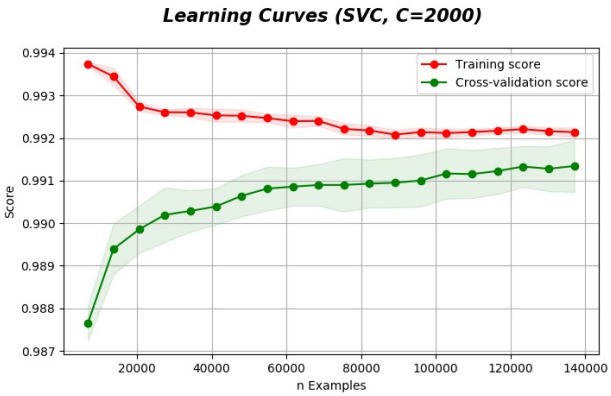

Figure 2: *Results obtained for the VAD module with the selected SVM classifier*

5. The emotion classifier module (EMO)

As Figure 1 shows, the first step in the proposed emotion classifier is to transform the power spectrum computed from the STFT of the signal following the Mel scale.

Then, MFCC coefficients are taken as input into the VGGish model to extract the input features used for the emotion recognition[1]. VGGish is a CNN model developed by the Sound Understanding team of Google to compute 128-dimension features vectors from Mel scaled input spectrograms. The model mixes multiple convolution-pooling layers to produce the feature vector.

The model consists of several convolution layers with pooling on the maximum value in the output of each one of them, consisting of a last layer totally connected for classification.

5.1. Transfer Learning

Implementing a CNN architecture specifically built to recognize emotions from a given input vector requires high amounts of data to generate an efficient model. Transfer learning techniques can greatly optimize the generation of Deep Learning models by using the network structure implemented in an existing model (with similar classification functions to the desired model), removing its output layer and implementing one specific to the classification task in hand. The resulting model can then be trained with the relevant data to fit the wanted model. Alternatively, instead of a classification network, the reference model can be used to extract features from an input vector, and then use the resulting feature vector as input for another Machine Learning classifier, which ultimately performs the classification task.

To use the VGGish model in the proposed application, a standard CNN model has been built with the same layer architecture used by VGGish. The last layer has been replaced by a Machine Learning classifier, which performs the emotional classification task on the characteristics extracted by the neuron network.

The training of the models was completed using the RAVDESS dataset, segmenting the original audios into fragments of one second and eliminating the audios in which there was not a sufficient amount of speech. Additionally, aggregation operations have been applied to the resulting audios (audio displacement, compression and noise introduction) to extend the number of available examples, which has made it possible to increase to a total of 30,000 files preserving the nature of the emotion that they represent.

[1]The definitions of the VGGish model implemented in Keras with tensorflow backend are available at https://github.com/DTaoo/VGGish. Last access: February 2021

Table 1: *Results of the evaluation of the set of classifiers for the EMO module*

Alternative	Model	Hyperparameters	Precision	Delta
CNN	KNN	N = 3	93.57%	3.2%
CNN	SVC	Penalty = 1.75	93.69%	5.67%
CNN	MLP (1) Layers: (80);	Learning rate: 0.05	92.66%	7.34%
CNN	MLP (2) Layers: (80, 80);	Learning rate: 0.05	93.24%	6.76%
MFCC	SVC	Penalty = 1.0	93.11%	5.3%
MFCC	KNN	N = 2	92.36%	7.64%

5.2. Classification

Starting from the log-Mel scale power spectrum, the resulting coefficients are supplied as input to the proposed VGGish model. The model extracts the characteristics of the input data, being these coefficients those that are finally supplied to the classifier for the detection of emotions. For the classification task of the VGGish output, 3 different models have been considered:

- KNN - K Neighbors: The model uses the concept of majority voting fixing classes to each classifiable object depending on the class of its neighbors: the most common neighbor class is the class assigned to the element.
- SV - Support Vector Machines: Support Vector Machines recursively transforms the dimensionality of the inputs to find hyperplanes that splits and classifies the whole space.
- MLP - Multilayer Perceptrons: Represents a neural structure formed by interconnected layers of nodes. The input is transformed while propagating through the layers to the output, where is finally classified.

As with the VAD module, multiple configurations have been considered on each classifier to select the one providing the best results. Table 1 shows the results obtained with the different classifiers. The Delta parameter refers to the differences between the test and validation sets.

All the selected models provide satisfactory results, however, the overfitting shown by the MLP models of the CNN alternative, as well as the KNN classifier considered during the classification of the MFCC coefficients, discards them as candidates for the final emotion classifier. Therefore, the model chosen for its integration in the system is the KNN, which shows a fairly close convergence between the test and training sets as well as offering an accuracy close to 94%. This result improves the baselines obtained for the RAVDESS corpus in recent proposals [16, 4, 14]. Figure 3 shows the confusion matrix obtained for this classifier.

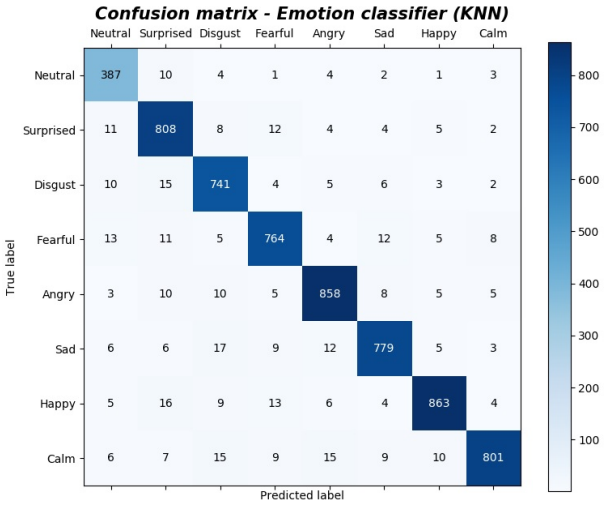

Figure 3: *Confusion matrix for the KNN classifier*

6. Conclusions and future work

In this paper we have presented a proposal for emotion recognition using speech features, transfer learning and convolutional neural networks. The proposed architecture integrates two independent subsystems that collaborate in the task of classifying an audio according to its content: VAD and EMO.

The VAD subsystem determines the existence of speech components in the audio using a support vector machine, which takes the metrics previously computed by the subsystem and decides on the presence of voice in the signal.

The EMO subsystem (classifier of emotions), acts on the basis of the decision reached by the VAD and transforms the power spectrum of the signal to a Mel scale to obtain a vector of its characteristics using a network of convolutional neurons. Then, an emotion tag is assigned by a KNN classifier.

We have implemented our proposal and used it with the RAVDESS dataset, obtaining an accuracy of 93.57% with 8 emotion categories. As future work we plan to evaluate our proposal with more datasets.

7. Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR project:https://menhir-project.eu) and by the Spanish Ministry of Economy, Industry and Competitiveness through CAVIAR (MINECO, TEC2017-84593-C2-1-R).

8. References

[1] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*. Springer, 2016.

[2] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. Wiley, 2013.

[3] M. B. Akçay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56 – 76, 2020.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[5] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion

recognition based on DNN-decision tree SVM model," *Speech Communication*, vol. 115, pp. 29–37, 2019.

[6] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[7] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding," *Expert Systems with Applications*, vol. 139, p. 112851, 2020.

[8] J. C. Hung, K.-C. Lin, and N.-X. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Applied Soft Computing*, vol. 84, p. 105724, 2019.

[9] B. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in *Proc. of IEEE International Conference on Signals and Systems (ICSigSys)*, Bandung, Indonesia, 2019, pp. 40–44.

[10] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *Plos One*, vol. 13, no. 5, p. e0196391, 2018.

[11] F. Saki and N. Kehtarnavaz, "Emotion Detection Using MFCC and Cepstrum Features," in *Proc. of the 4th International Conference on Eco-friendly Computing and Communication Systems*, Orlando, FL, USA, 2015, pp. 29–35.

[12] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *Proc. of 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, Florida USA, 2016, pp. 29–35.

[13] Z. Tuske, P. Mihajlik, Z. Tobler, and T. Fegyo, "Robust voice activity detection based on the entropy of noise-suppressed spectrum," in *Proc. of 9th European Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, 2005, pp. 245–248.

[14] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[15] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[16] M. Kwon and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors*, vol. 20, pp. 1–15, 2019.