Machine Learning for Speech Emotion Recognition

Loop Q Prize

Alexandros Philippos Pouroullis

# 1. Introduction

Speech emotion recognition (SER) has become a very popular problem in the field of artificial intelligence because of its importance in getting computers to understand humans, because speech is our principal form of communication. By making advancements in SER, we can make human-computer interaction more natural.

# 2. The Problem

SER, of course, involves dealing with audio. Audio in its raw format isn’t particularly meaningful to algorithms because of the limited information it contains; audio is simply a time-series of amplitudes. That provides the model only with information about the intensity of the sound at different points in time. Though, deep learning, notable for its ability to extract relevant features automatically, may be able to discern useful patterns from the data without having to extract them manually. However, audio is also high-dimensional in its raw form, making it inefficient to feed into algorithms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Corpora | Samples | Language | Gender balanced | Accent (Homo-/heterogeneous) |
| CREMA | 6306 | English | Yes | Heterogeneous |
| RAVDESS | 1041 | English | Yes | Homogeneous |
| SAVEE | 397 | English | No | Homogeneous |
| TESS | 2366 | English | No | Homogeneous |

So instead of using raw audio, features are extracted from the audio. There are many features of different classes, typically prosodic features (pitch, loudness, rhythm, intonation) and characteristics of the vocal tract, that can be used for emotion detection [1].

Table 1: Summary of the corpora

Traditionally, features were hand-crafted and required domain-expertise to extract, but more recently, spectro-temporal features, like spectrograms and mel-spectrograms have become immensely popular, and they also have the added benefit of being able to be treated as images and fed into convolutional neural networks. (CNNs), which has resulted in impressive results.

# 3. The Corpora

One very important aspect in speech emotion recognition is the nature of the speech: is it simulated, natural, elicited? Moreover, what languages do the speakers speak in, are they male or female, do they have accents, and what words are spoken? The corpora provided for this project all involve speech spoken in English, with varied accents across corpora (in the case of CREMA, within the corpus), and, for the most part, arbitrary utterances unrelated to the emotional content.

These aspects have important implications, because will people in reality make their emotions as obvious as the actors do in the audio, and will the content of their speech be completely unrelated to its emotional content?

# Diagram Description automatically generated

Figure 1: Extracting the mel-spectrogram from an acoustic signal. Accessed from https://www.researchgate.net/publication/351469852\_Toward\_an\_Automatic\_Quality\_Assessment\_of\_Voice-Based\_Telemedicine\_Consultations\_A\_Deep\_Learning\_Approach

# 4. Feature Extraction

The two types of features I used were a) spectro-temporal features (spectrograms, mel-spectrograms, mel-frequency cepstral coefficients and its derivatives) and, what I call, static features (usually the mean, minimum or maximum of various quantities). All features were extracted using librosa [5].

# 4.1. Extracting the spectro-temporal features

Spectrograms, mel-spectrograms, and mel-frequency cepstral coefficients (MFCCs) are all related: they depict the intensity of frequency bands through time.

Exctracting spectrograms involves framing the signal into short frames, windowing them, applying the discrete Fourier transform on each frame, and then stringing them together, giving us information about the frequency, intensity and how these evolve through time. Though the logarithm is often taken of the spectrogram because of the very large dynamic range of audio. The mel-spectrogram is very similar, only it applies a triangular filter bank onto the log-spectrogram, such that each band is

perceptually equidistant by the human auditory system. The mel-spectrogram is

the preferred input for machine learning because it presents the algorithm with information that matches more closely how a human would perceive the speech.

MFCCs are also quite popular, but more difficult to interpret. They extend from the

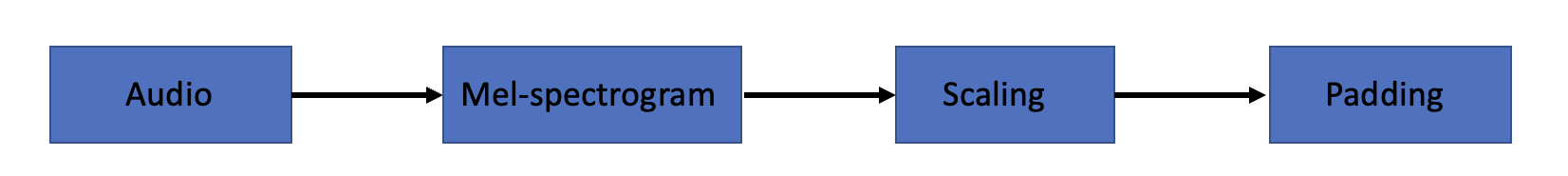
mel-spectrogram, taking the inverse Fourier transform (or in some implementations, the discrete cosine transform) of the mel-spectrogram to obtain a spectral envelope that is supposed to essentially give us information about the formants in the speech.

# 4.2. Extracting the static features

The reason I call them static features is because they are, for the most part, global statistics, or statistics that capture a single value for the entire audio file. They were usually means, minima or maxima. The feature extraction method was inspired by several sources, the principal one being a project by Xiaohan Zou, whose github page will be referenced at the bottom of this document [2]. The other features that were extracted were mel-energy spectrum dynamic coefficients [3] and the mean fundamental frequency [4]. The final feature space contains a total of 241 features. Some of the features include pitch, the zero crossing rate, energy, root-mean-square and others.

# 5. Metrics

The F1-score is preferred in model comparison because it captures the balance between precision and recall. However, it comes in different flavours which are worth mentioning.

The micro F1-score weights all samples equally; the macro F1-score weights all classes equally; and the weighted F1-score weights individual F1 scores in computing the average by the number of samples in each class. They all give very similar results if there’s no class imbalance, which in this case there wasn’t much of, so the differences between the different scores isn’t of much importance.

# 6. Model comparison

4 of the 5 models compared were trained on the static features, while the last model – the CNN – was trained on mel-spectrograms. The labels were integer-encoded in both cases.

Figure 2: Preprocessing of audio for the spectro-temporal features.

A picture containing graphical user interface

Description automatically generated

Figure 3: Preprocessing of audio for the static features.

Chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 4: CNN architecture. CL1=32 5x5 filters; CL2=64 3x3 filters; MPL + Dr = Max pooling (2x2) with 20% dropout. FC=Fully connected layer

Figure 5: Deep neural network architecture (Best performing model). FC=fully connected layer; Dr(0.2)=Dropout layer with 20% dropout.

Table 2: Scores of each model after 10-fold cross-validation. \*The CNN's metrics are based on a single train-test split due to long training times.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metrics | DNN | RF | KNN | SVM | CNN\* |
| Micro F1 | 0.66 | 0.63 | 0.55 | 0.63 | 0.59 |
| Macro F1 | 0.68 | 0.64 | 0.57 | 0.64 | 0.6 |
| Weighted F1 | 0.66 | 0.62 | 0.55 | 0.62 | 0.58 |

The best performing model in the end was the deep neural network trained on the static features.

# 7. Strengths of the solution

The solution was trained on audio from a variety of sources, of differing qualities and on speech with different accents, and still managed to perform somewhat well. Since it uses static features, it can handle audio of varying durations, making it quite convenient to use. Some of the audio also contained a fair amount of noise, which means that it’s still able to identify the relevant features when there’s reduced clarity. The model is relatively small and very quick to train, making it possible for more modest devices to run it.

It's likely to generalize well to other English speakers with differing accents, as well as to both genders because it was trained on speech with a variety of accents and with both male and female speakers. There were also many different speakers in the database, meaning that the model isn’t speaker-dependent.

# 8. Limitations of the solution

As was noted earlier, the corpora include speech with emotional content that is simulated. The issue with this is that the model is likely not going to generalize very well to natural speech. Moreover, the speakers were exclusively English speaking, so it’s not certain whether the model would be able to generalize to speakers of other languages.

The other limitation is that the model doesn’t consider lexical features. Words can give a strong indication of emotional content of speech along with prosody and vocal tract characteristics. In future work, it would be worth exploring a pipeline that combines a model for speech-to-text generation and a model that analyses the other features, spectral and such.

# 9. Conclusion

This project was, in truth, the most difficult and ambitious project I’ve ever worked on, but it proved to be the best experience I’ve ever had in machine learning and will be invaluable in my future endeavours. I sincerely thank you for considering my project and giving me the opportunity to work on a problem like this.

**References**

[1] Langari, S., Marvi, H. & Zahedi, M. 2020. Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*. 20:100424. DOI: 10.1016/j.imu.2020.100424.

‌[2] Xiaohan Zou’s github page: <https://github.com/Renovamen>

And the related project:

<https://github.com/Renovamen/Speech-Emotion-Recognition>

‌[3] Yashpalsing Chavhan, Manikrao Dhore & Yesaware Pallavi. 2010. *Speech Emotion Recognition Using Support Vector Machines*. Foundation of Computer Science. Available: https://www.researchgate.net/publication/43785303\_Speech\_Emotion\_Recognition\_Using\_Support\_Vector\_Machines [2022, June 14].

[4] Seehapoch, T. & Wongthanavasu, S. 2013. Speech emotion recognition using Support Vector Machines. *2013 5th International Conference on Knowledge and Smart Technology (KST)*. (January). DOI: 10.1109/kst.2013.6512793.

[5] McFee, B., Alexandros Metsai, McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., et al. 2022. librosa/librosa: 0.9.1. *Zenodo*. (February, 15). DOI: 10.5281/zenodo.6097378.

‌

‌

‌

‌