Automatic Emotions Recognition in Audio Signals

Emilie Tavernier, Catherine Marechal, Laurie Conteville, Lamine Bougueroua

Automatic Emotions Recognition in Audio Signals

Emilie Tavernier, Catherine Marechal, Laurie Conteville, Lamine Bougueroua
EFREI Paris, AlliansTIC, Villejuif, France
emilie.tavernier@efrei.net
{catherine.marechal, laurie.conteville, lamine.bougueroua}@efrei.fr

Abstract

Automatic emotions recognition is a great research subject nowadays. It applies in different fields as robotic, data analysis, E-health. We can use different supports for emotions recognition: text, vision and voice. These three sources complement each other to analyse the emotional state of someone. In the field of health especially, this can be extremely helpful to detect pathologies early, like depression, so that doctors can prescribe rapidly appropriate healthcare.

This paper deals with classification methods of eight emotions with voice recordings. Machine learning algorithms are used. Particularly, the quadratic Support Vector Machines (SVM) gives the best results. We use supervised training (voice recordings are labelled according to the emotion) and features as the maximum, the means, and the variance of various acoustic properties… We use labelled databases of English voice recordings for our emotional classification study. The results depend on the number of emotions to be detected and on the type of emotion.

Keywords: Speech emotion recognition, IA, SVM, MFCC

I. INTRODUCTION

Affective Computing (AC) attempts to bridge the communication gap between human users and computer with "soulless" and "emotionless" feeling. The inability of systems to recognize, express and feel emotions limits their ability to act intelligently and interact naturally with us.

Moreover, the interest in understanding emotions is multi-disciplinary and covers a long history of research. The importance of modelling emotions has multiple benefits across applications such as E-health [1], E-learning [2], advanced driver-assistance systems [3], etc.

From a computational perspective, different vectors of emotions can be analysed. This includes facial expressions [4], speech [5] and multimodal approaches [6]. Facial emotion recognition especially is often used as all the people of the world share the same facial expressions for seven primary emotions (anger, contempt, disgust, enjoyment, fear, sadness, and surprise) [4]. However, depending on the country legislation, camera cannot be used in places considered private.

As an alternative, real-time voice analysis by algorithms could be considered less intrusive to determine the emotional state of someone. The aim of our study is to design and develop systems that can measure the emotional state of a person based on acoustic characteristics. In the practical case, sound signals recorded in real-time can be associated with predefined emotions. If negative emotions (sadness, anger…) are detected, an alarm can trigger adequate actions from humans or machines.

Currently, one of the main difficulties for emotion detection is the choice of features and their number. To reduce the features number, the PCA (Principal Component Analysis) method is largely used [7]. Among the most popular algorithms used for emotional detection are SVM and Hidden Markov Models (HMM) scheme [8, 9, 10] or both [11] with Mel-Frequency Cepstral Coefficients (MFCC) features extraction [12].

In this paper, we use global and "semi-local" features for which the signal is split into three parts. More complex methods of "semi-local" features as voiced/unvoiced, phonemes or phrasing [7, 13] are also used for speech emotion recognition.

In this paper, section II details the workflow and architecture of our emotion recognition system (software, data, algorithm…). In section III, we present our results and in section IV, we give the conclusion and perspective of our work.

II. METHODOLOGY

For this study, we used a database of 1379 labelled recordings (single sentence) by 20 actors from both sexes, with a sample rate of 48kHz. The emotions featuring in our database were anger, sadness, calm, surprise, joy, disgust, neutrality and fear. Recording has been clean by removing silences at the beginning and at the end of the signals and normalizing them so that they have all similar amplitude. There were no apparent needs for filtering noise.

Data were imported into MATLAB's Audio Datastore and features such as the *pitch*, *Mel-Frequency* and *Gammatone Cepstral Coefficients* (MFCC and GTCC) were extracted using

MATLAB's *audioFeaturesExtractor*[1]. Additionally, we used Praat[2] software to compute additional features: the *shimmer* and *jitter*. MATLAB's and Praat compute local features on small temporal windows resulting in an important volume of data. To keep computational cost reasonable, we transformed them in global features by computing statistics over them (mean, standard deviation, range…) or "semi-local" features (statistics covering sections of the signal). Like this, we obtained a set of 1188 features that we will now call 1188 aFE.

The features were then given as inputs to a classification model training algorithm. We tested all algorithms of MATLAB's Classification learner app on features sets with various sizes. For the rest of the study, we retained Quadratic SVM as the overall top performing one (closely tied with Cubic SVM). We used a 5-fold cross-validation and accuracy score to evaluate our models as our dataset were overall well balanced.

The figure 1 summarizes the overall architecture of the model, distinguishing the parts from MATLAB from external sources.

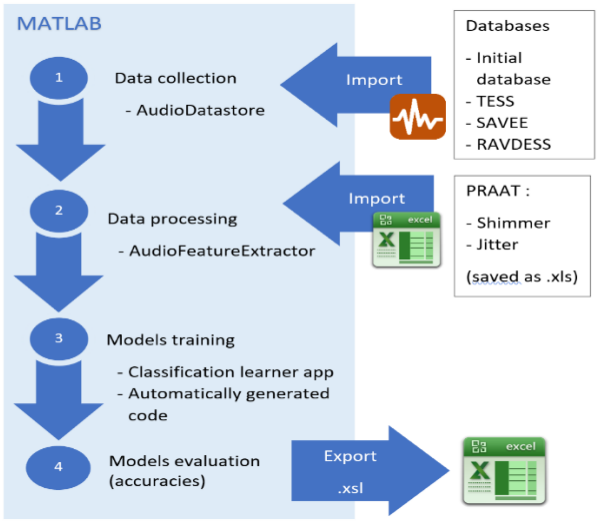

Figure 1. Architecture proposed

III. Results

While training models to classify the eight emotions (anger, sadness, calm, surprise, joy, disgust, neutrality and fear), we reached a maximum of 60.1% of accuracy on 1188 aFE set.

In order to identify more precisely which emotions where more challenging, we decided working on binary models instead.

A. Binary classification: 1 emotion vs others

First, we trained models to classify a single emotion against the rest (seven other emotions of the database). The sets were balanced. Here is a recap of the learning parameters for the experiment:

- Algorithm: Quadratic SVM
- Features: 1188 aFE (mean, max, median, std, range, var)
- Evaluation: 5-folds cross-validation
- Signal pre-processing: normalisation and silence removal

Additionally, to better understand our features individual influence on the models, we tried training models using only one feature at a time. With this, we wanted to see if some features were especially good to identify an emotion by themselves. These experiments resulted in about 12000 models whose accuracy score ranged from 29% to 74%.

We were not able to identify miracle features, but we used these scores to make rankings of the features per emotions. From there we trained binary models on each emotion using their top 10, 40 and 200 ranked features and compared them to our initial model trained on the set 1188 aFE.

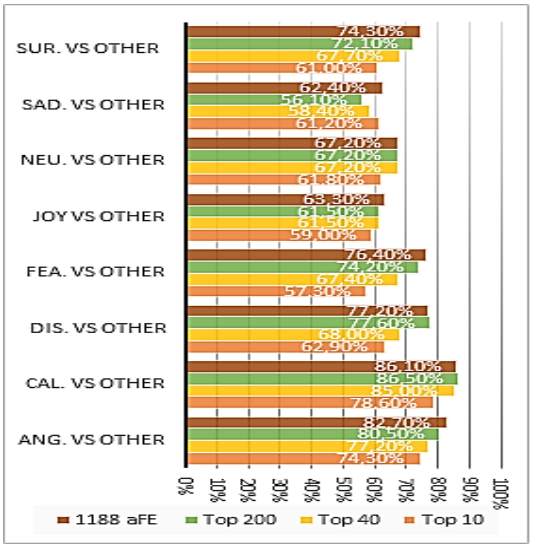

Figure 2. Accuracies obtained with various features sets

[1] https://fr.mathworks.com/help/audio/ref/audiofeatureextractor.html

[2] https://www.fon.hum.uva.nl/praat/

The results on figure 2 for the models trained with 200 features are quite close to the ones for the set 1188 aFE (expect for sadness). This clearly shows potential and calls for experimenting more on feature reduction, maybe comparing it with standard methods such as PCA.

B. Binary classification: 1 emotion vs 1 emotion

To have a clear vision of which emotions tend to be confused between each other, we trained binary models to classify one-emotion vs another. The training parameters for this section were the same as in part II.A.

VS	Cal.	Dis.	Fea.	Joy	Neu.	Sad.	Sur.
Ang.	87%	83%	91%	89%	89%	89%	84%
Cal.		90%	93%	92%	74%	77%	92%
Dis.			92%	83%	80%	78%	86%
Fea.				78%	78%	78%	76%
Joy					77%	82%	77%
Neu.						65%	84%
Sad.							83%

Figure 3. Accuracies obtained on binary models (1 emotion vs. 1 emotion)

The results in figure 3 show that most confusion happens for neutrality and sadness. We can notice also that surprise and fear tend to be confused too.

In order to improve these results, we tried to divide our signal to compute "semi-local" features on each portion. In fact, global features may be too generic and miss a lot of information and we got the idea that some emotions could be more audible at some specific time in a locution. For instance, someone surprised would likely rise his voice toward the end of a sentence.

The most interesting results regarding this hypothesis were obtained when dividing each signal in three equal parts and training models on each individual part. We remind that in this study, a signal corresponds to a sentence.

Considering the accuracies we obtained when training on one section at a time (beginning, middle and end), it appears overall that the beginning of the sentence is the least informative emotion wise while the end contains most of the relevant acoustic cues. Moreover, training models solely on the end section rather than the whole sentence resulted in a gain in accuracy for several emotions (figure 4). For instance, we gained 7% of accuracy for the binary model "anger vs calm" which initially scored 87% when training on the whole sentence. This method however resulted in a loss of accuracy for most models implying fear or joy.

We think that these results advocate for further experimentations using "semi-local" features and may justify in the future the effort to train a model to split sentences as a first layer to an automatic system for emotion recognition.

VS	Ang.	Cal.	Dis.	Neu.	Sad.	Sur.
Ang.		+7%	+5%	+3%	-4%	+5%
Cal.	87%		-1%	+6%	0%	0%
Dis.	83%	90%		+5%	+2%	+1%
Neu.	89%	74%	80%		+4%	+2%
Sad.	89%	77%	78%	65%		-1%
Sur.	84%	92%	86%	84%	83%	

Figure 4. Gains in accuracy for models trained on the end of a sentence (red/blue) over accuracies obtained on whole sentence (yellow/green).

IV. Conclusion and Perspectives

Many fields use Automatic Emotions Recognition as robotic, data analysis, E-health, analysing different supports as text, vision or/and voice. In the field of health, the emotion detection is very helpful to early detect signs of disease for a rapid patient care.

In this study, we experimented on automatic detection of eight emotions based on acoustic signals using classification methods. We studied emotions from 1379 audio-files recorded by actors. We used the supervised algorithms quadratic SVM, global or "semi-local" features for three kinds of models. One of them is the multi-classes model of eight emotions and we obtained 60.1% accuracy. The others are the binary models identifying one emotion among the others. From this, we established a ranking of features per emotions. We used this ranking to reduce the number of features. Finally, the binary models of one emotion vs another allowed us to identify which emotions are more difficult to distinguish from each other, for example neutrality and sadness or fear and surprise. We also established that the end of a sentence is the most important part to detect the emotion. This very interesting result allows reducing the analysed data.

For the future of this work, we consider swapping to open-source databases so that our result can be reproducible and compared to other papers. The RAVDESS[3] dataset is the closest

[3] https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

to the one we used and it would be a good candidate. TESS[4] and SAVEE[5] are other possibilities though they were recorded with less actors so the results obtain on these would risk being less generalizable.

Our attempt to reduce the number of features were encouraging though unconventional. It would be interesting to compare it to other algorithms for features selections such as PCA.

Finally, according to our results, it would be interesting to explore signal segmentation methods such as sentence detection to compute "semi-local" features for real-time emotions detection on speech analysis.

REFERENCES

[1] Tlija, A., Istrate, D., Bennani, A., et al., "Monitoring chronic disease at home using connected devices". In: 13th Annual Conference on System of Systems Engineering (SoSE), pp. 400–407. IEEE (2018), http://doi.org/10.1109/SYSOSE.2018.8428754.

[2] Trifa A., Hedhili A.,Lejoued Chaari W, "Knowledge tracing with an intelligent agent, in an E-learning platform". In the journal of "Education and Information Technologies", Springer. Pp. 711–741 (2019). https://doi.org/10.1007/s10639-018-9792-5.

[3] Hernandez, J., et al., "AutoEmotive: bringing empathy to the driving experience to manage stress". In: DIS 2014, 21–25 June 2014, Vancouver, BC, Canada. ACM (2014). http://dx.doi.org/10.1145/2598784.2602780.

[4] Paul Ekman Group, Fear, 2021, Paul Ekman Group LLC, https://www.paulekman.com/universal-emotions/. Accessed: 2021-02-18.

[5] Deng J., Xu X., Zhang Z., Frühholz S., Grandjean D., Schuller B. "Fisher Kernels on Phase-Based Features for Speech Emotion Recognition". In: Jokinen K., Wilcock G. (eds) Dialogues with Social Robots. Lecture Notes in Electrical Engineering, vol 427. 2017. Springer, Singapore. https://doi.org/10.1007/978-981-10-2585-3_15.

[6] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, et al., "Survey on ai-based multimodal methods for emotion detection" in High-Performance Modelling and Simulation for Big Data Applications, Springer, pp. 307-324, 2019. https://doi.org/10.1007/978-3-030-16272-6_11.

[7] Marie Tahon, Laurence Devillers. "Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges". IEEE/ACM Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers, 2016, IEEE/ACM Transactions on Audio, Speech and Language Processing, 24, pp.16 - 28. https://hal.inria.fr/hal-01404146.

[8] Milton A., Sharmy Roy S., et al., "SVM Scheme for Speech Emotion Recognition using MFC Feature". In : Internal Journal of Computer Application Vol. 69-No.9, May 2013. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.404.1324&rep=rep1&type=pdf.

[9] Sun, L., Fu, S. & Wang, F. Decision tree SVM model with Fisher feature selection for speech emotion recognition. J AUDIO SPEECH MUSIC PROC. 2019, 2 (2019). https://doi.org/10.1186/s13636-018-0145-5.

[10] Koduru, A., Valiveti, H.B. & Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23, 45–55 (2020). https://doi.org/10.1007/s10772-020-09672-4.

[11] Swain, M., Sahoo, S., Routray, A. et al. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. Int J Speech Technol 18, 387–393 (2015). https://doi.org/10.1007/s10772-015-9275-7.

[12] P. P. Dahake, K. Shaw and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 1080-1084, https://doi.org/10.1109/ICACDOT.2016.7877753.

[13] Moataz El Ayadi, Mohamed S. Kamel. and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases". Pattern Recognition, Volume 44, Issue 3, 2011, Pages 572-587, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2010.09.020.

[4] https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess?

[5] https://www.kaggle.com/ejlok1/surrey-audiovisual-expressed-emotion-savee