Prosody recognition

Risch Guillaume, Delaunay Antoine, Nanecou Thomas

June 2022

1 Thanks

We would like to thank Madam Stephanini for the interest brought to our project and also, for her supporting our project through its construction.

We would like to thank Madam Jakubiek for her support too.

2 Introduction

We chose the subject of prosody because it's a way to recognize different feelings. It allowed us to start and understand AI and the work of scientist researcher. Our goal is to create a program able to know if a user is angry, sad, or happy and if it's a man or a woman. We hope our work can be helpful one day, especially to create a simulation of different interactions between humans and machines. Also, our project could be helpful to people who want to work on their social interactions, it could help them to understand the different intonations in their voices. We know the importance of this project because usually the recognition of emotions is done from facial analysis but never from a conversation. The company Dataiku have created a program with the same functionality, however in order to access it, you have to pay. Our wish is to propose a project which would be an open source and community to be improved and to become more and more efficient and to advance research in this field.

3 Background of the study

We have noticed that there are a lot of studies on prosody. Indeed, while searching about Hal and reading several research paper, we saw that studies often include intonations, pauses which can be influenced by adverbs and emotions that pass through a speech ([1][2][3]). There are also research works on the impacts of prosody or how to increase prediction to find emotions in a speech ([4][5][6][7]).

This allowed us to find a large dataset, named "Toronto emotional speech set (TESS)" to test all the different prosodies.

4 existing project

Before starting a new project it is important to know what have been done already in the world, to try to differentiate yourself. That's why we started our research to find out if there were already companies or people who have already worked on it. We found 3 people who have already worked on this subject and who have realized personal projects. These projects are just meant to be personal projects for training. They often have an accuracy of 70% ([12][13][14]). We also found a company that worked on it [14].It is a French company that has an accuracy rate of 60%-70% depending on the different emotions.

5 Analysis of the needs

As a consequence of this research, we know which part of our language we need to improve or create. Our goal is to be able to analyse a sentence. To do this, we will need several elements because we cannot just create a program that would answer randomly on each sentence. The first problem we encountered, was to make sure our program can retrieve the different informations of a sentence. As, the gender of the person, the different variations of F0, the number of pauses that the person makes etc... Moreover, our program must be as efficient as possible so that it responds in a minimum of time.
F0[11] the fundamental frequency refers to the approximate frequency of the periodic structure of voiced speech signals.
The other problem we encountered is that our program needed training to be able to do all this. So we chose to use a classifier.

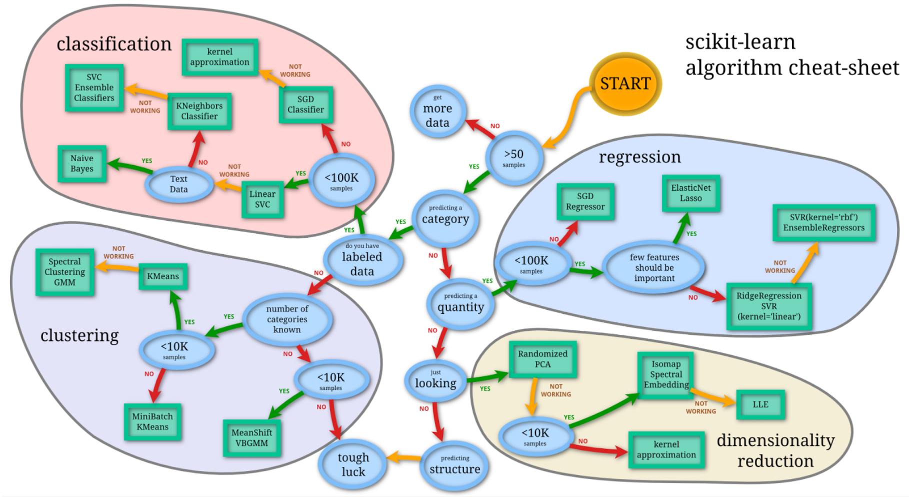


Figure 1: Features Graph

In order to train our machine, we are going to need a maximum of datas to be as accurate as possible. We have already started to look for different datasets that would allow us to train our program in the best way, based on the diagrams proposed by scikit-learn as illustrated in figure 1.
Once all these conditions are met, we will be able to train our classifier on our different datasets to be able to recognise weather the user is a male, a female, and to know how they feel.
We use the library pickle[8] to make us able to pass an audio to our program and make our classifier recognize if this audio is an happiness, fear or sadness extract.

6 Feasibility study

We chose to use python as programming language because it is the best language when it comes out to artificial intelligence and automatic language processing. One of the many strong aspects of python is its libraries which will help us realize our project. There are two main libraries that will be useful. The first is called pandas, made for modifying data structures [and operations for manipulating numerical arrays. The second one is sklearn, dedicated to learning and data prediction.

We also use google colab notebook to run our program. This tool is very useful because it's based on the power calculation of google. We also can work at the same time on the code without damaging the main program thanks to the different cells.

Carrying on our research, we came across a library which is extremely efficient in the study of languages, called myprosody[9]. Indeed, this library will allow us to collect different informations and it will be up to us to deduce whether this are necessary or not regarding the rules of features[10].

Unfortunately myprosody[9] didn't work, so we tried to debug by printing on each line a string to understand where is the problem and we find its caused by the library Parselmouth where Myprosody is based. We opened an issue on github to ask for help from its creator but we got no response. We started to code without any library but with the short time left it would be impossible to complete the project.

At last we find a new library call Disvoice[1] allowing us to extract features of F0[11] contour of the fundamental frequency.

7 data analysis

After the end of our project and especially after our program allows us to know an audio corresponds well to the emotion that we wanted to make him say. We passed to the second phase, the analysis of the data. It was very important for us to be able to analyze our data in order to identify which features were more important than others.
Despite the fact that our program cannot define whether the voice of a man or a woman, it still seemed important for us to know which features would define it. That is why we first started our data analysis on our old CSV.

This old CSV comes from an open source site[16]. It has many features : ”meanfreq”, ”sd”, ”median”, ”Q25”, ”Q75”, ”IQR”, ”skew”, ”kurt”, ”sp.ent”, ”sfm”, ”mode”, ”centroid”, ”meanfun”, ”minfun”, ”maxfun”, ”meandom”, ”mindom”, ”maxdom”, ”dfrange”, ”modindx”

So we decided to make tables of them and compare them to allow us to see which features are more essential.

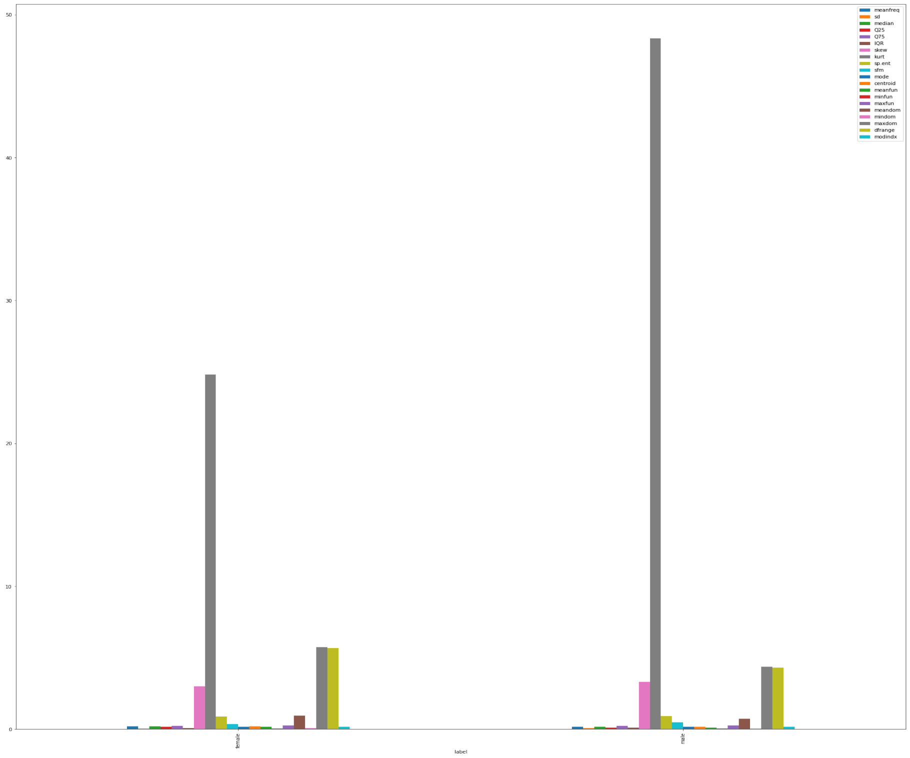


Figure 2: Women Man Features Graph

On figure 3 We notice that the kurt[19] features (kurtosis) are much more important for men, while the maxdom[20] and dfrange[21] features are higher for women. Otherwise we notice that the rest of the features are very equal between men and women.

Afterwards we could not continue on this CSV, indeed it does not correspond any more to our needs. This is where our program comes into play because it allowed us to recover CSVs from audio files. Thanks to this new CSV we have new features that do not correspond to the old ones. They are often based on F0.

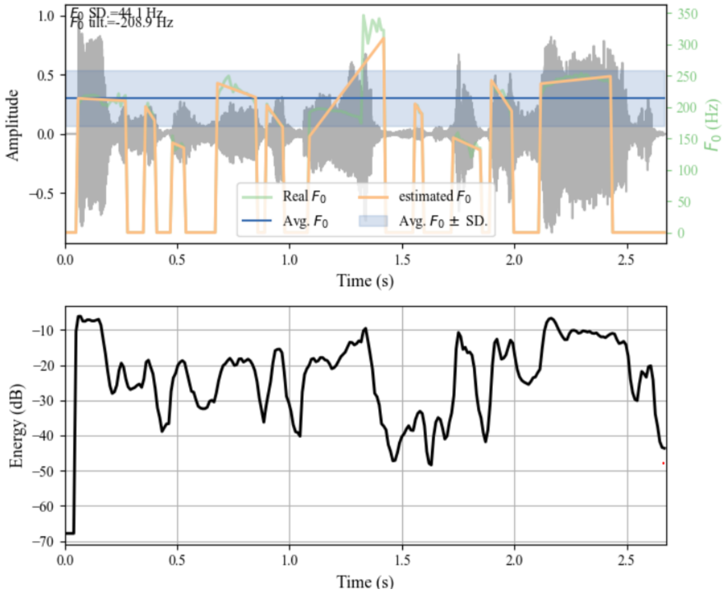


Figure 3: F0 Features

On figure 3 there are the different oscillations of F0 on a short speech.On green the real frequency which will be modified in several data like the average of the power in the speech.In yellow a simplification of the frequency who make us able to keep the most important moment of the audio[22].
This data are after cut in several features

”F0avg”,”F0std”, ”F0max”,”F0min”, ”F0skew”,”F0kurt”, ”F0tiltavg”,”F0mseavg”, ”F0tiltstd”,”F0msestd”, ”F0tiltmax”,”F0msemax”, ”F0tiltmin”,”F0msemin”, ”F0tiltskw”,”F0mseskw”, ”F0tiltku”,”F0mseku”, ”1F0mean”,”1F0std”, ”1F0max”,”1F0min”, ”1F0skw”,”1F0ku”,”lastF0avg”, ”lastF0std”,”lastF0max”, ”lastF0min”,”lastF0skw”, ”lastF0ku”,”Vrate”, ”avgdurpause”,”stddurpause”, ”skwdurpause”,”kurtosisdurpause”, ”maxdurpause”,”mindurpause”

After retrieving our CSV we thought it was also important to analyze the different features sorting them by emotions. This will allow us to differentiate the differences between the emotions.

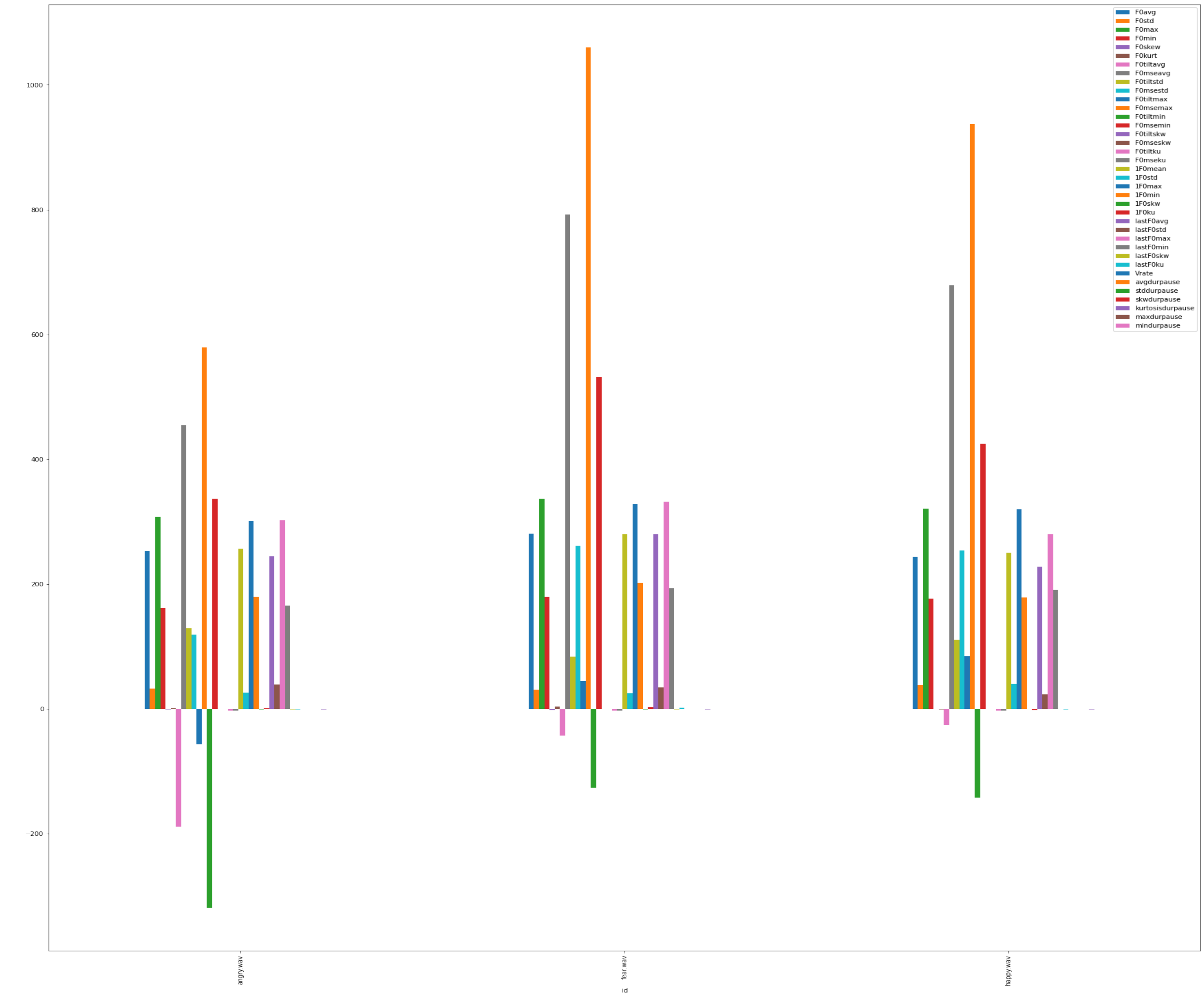


Figure 4: Emotion Features Graph

As we could see on this figure 4 some features differ enormously according to the different emotions. Indeed the features are completely different according to the emotions, we can see that the feature F0msemax is much more important for fear compared to happy and anger or the feature F0mseavg which is also more important for fear. On the contrary fear has much less important features for F0tiltmin.

But what may be surprising is that fear and happy have a much closer value spectrum than anger.

We could go on like this for a long time, comparing values is not very important, the main thing is to know if our program still manages to differentiate emotions and give us satisfactory results. That's why we make the confusion matrix that allows us to know if the program is doing a good job. In spite of the totally different data tables for anger and fear, we notice that in figure 5 our program puts 20 audio files in the fear category when they should be in anger. Thus we notice that our program has more difficulty in classifying the angry audio files than the others because it is on this line that the majority of the big errors are concentrated.

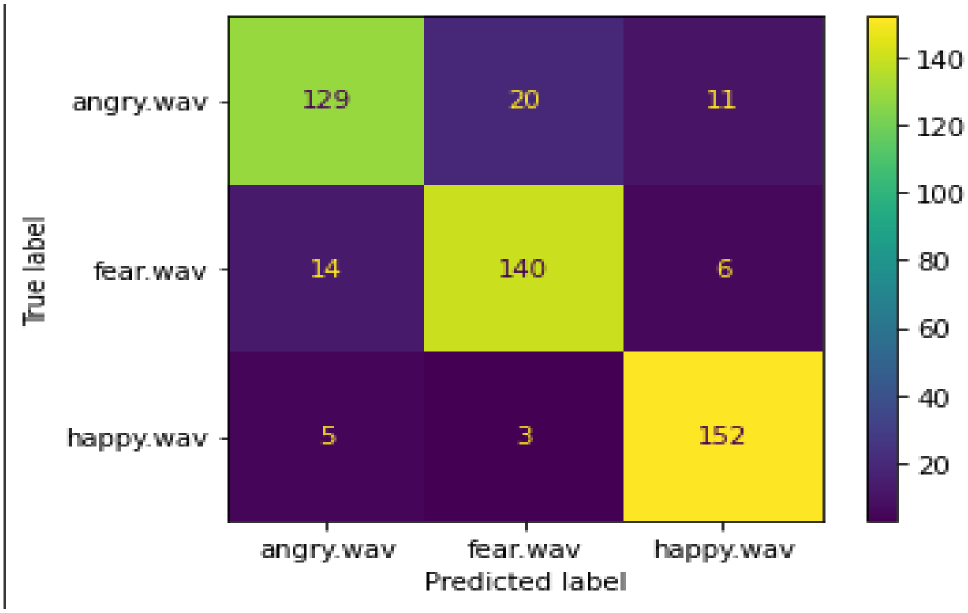


Figure 5: Confusion Matrix

But we still get a prediction of 89% at the end, which is very acceptable and fully satisfies us.

8 Project functioning

Our program works by creating a csv of the Tess data in the data processing step figure 6 These data are our training set but before we can exploit them we need to retrieve only the features that will be useful to us during the training, that's why we extract them. Finally comes the most interesting part, the one of deep learning, after having trained our program on our data set we serialize them to be able to save them and exploit them to allow our program at the moment we play an audio file to determine the nature of this one thanks to the different target (emotions) that it is supposed to find.

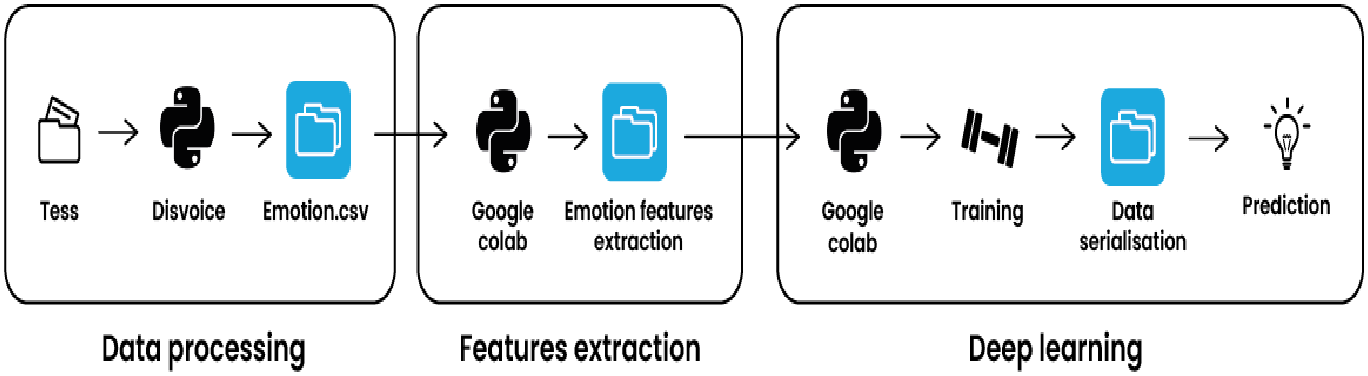
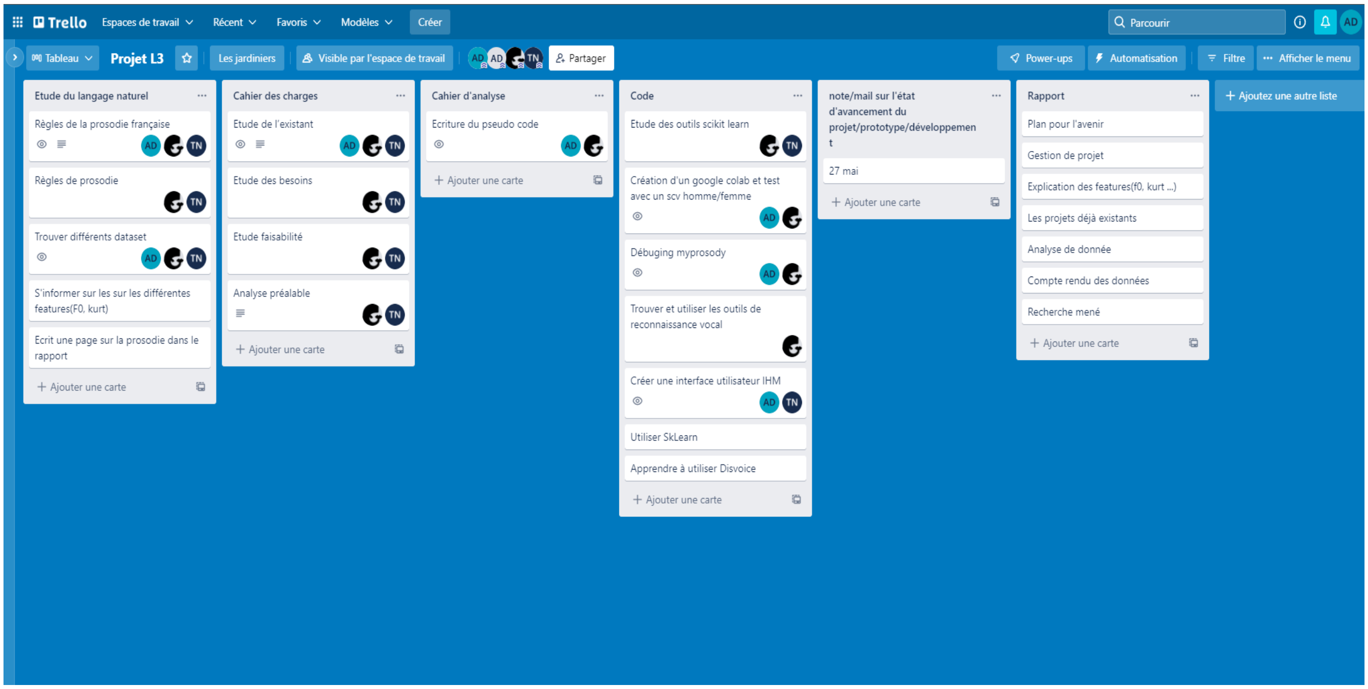


Figure 6: Project Steps

Data processing and Feature extraction are managed by google colab as this is the most resource intensive part of the program. The results provided by google colab can then be processed on our computers as the deep learning phase is much lighter.

9 Project Management

For the organization of the project, we tried to make it as simple as possible. Indeed, we are currently three in this group, this allows us to be in rotation on the project and thus to have always a person present to work. We were looking for a tool that would allow us to plan our tasks in advance, correct them and modify them. That's why we decided to choose Trello. This tool has the advantage of being easy to use, simple and functional. It allowed us to know which tasks we still had to do but also which ones we were behind on. The other advantage that Trello gave us was that we didn't need a manager to operate efficiently. We didn't all have the same schedule, some of us had to work on different hours, so we couldn't always communicate. Trello allowed us to overcome this problem



Etude du langage naturel
- Règles de la prosodie française
- Règles de prosodie
- Trouver différents dataset
- S'informer sur les sur les différentes features(F0, kurt)
- Ecrit une page sur la prosodie dans le rapport

+ Ajouter une carte

Cahier des charges
- Etude de l'existant
- Etude des besoins
- Etude faisabilité
- Analyse préalable

+ Ajouter une carte

Cahier d'analyse
- Ecriture du pseudo code

+ Ajouter une carte

Code
- Etude des outils scikit learn
- Création d'un google colab et test avec un scv homme/femme
- Débuging myprosody
- Trouver et utiliser les outils de reconnaissance vocal
- Créer une interface utilisateur IHM
- Utiliser SkLearn
- Apprendre à utiliser Disvoice

+ Ajouter une carte

note/mail sur l'état d'avancement du projet/prototype/développement
- 27 mai

+ Ajouter une carte

Rapport
- Plan pour l'avenir
- Gestion de projet
- Explication des features(f0, kurt ...)
- Les projets déjà existants
- Analyse de donnée
- Compte rendu des données
- Recherche mené

+ Ajouter une carte

Figure 7: Trello

To help us plan our project, we decided to use a Gantt chart. It was important for us to be able to use this chart because it would allow us to plan ourselves and our time well. Indeed this diagram has the advantage of allowing us to see on which part we are late and on which part we are ahead.

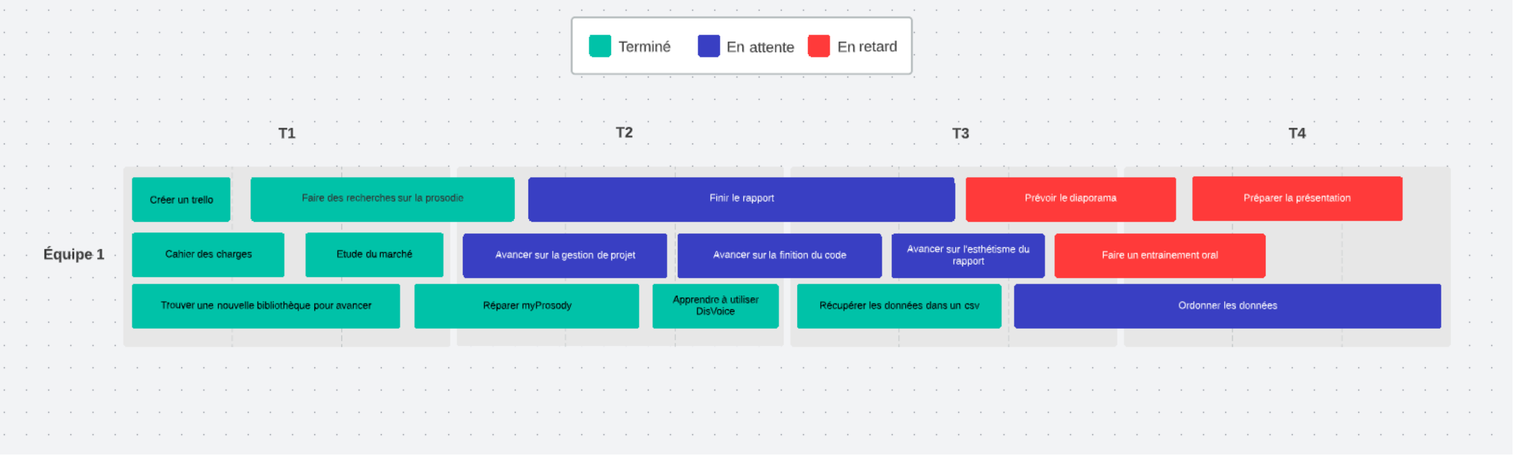


Figure 8: Gantt Chart

With these two tools we can say that we tried to implement an agile method to move this project forward. We wanted to use agile methods because in the world of IT, they are more and more present on different projects and allow to have a time management which is better. These agile methods are not intended to make us work with a strict schedule but more to allow us to flourish in a project. One of the other very important points is that it allows us to be free in our ideas and to be able to propose all possible ideas.

10 Vulnerabilities

We find some vulnerabilities due to the evolution of python it's really difficult to maintain a stable environment and what works on our version of python could conflict with new version.
Pickle can be a gate of security break because if untrusted are executed data as it could lead to malicious code being executed upon loading.

11 Outlook for the Future

To finish a project we think it is also important to know the different features we could have added if the given time was longer. For that we will imagine that the duration of our project is 6 months. We think that with 6 months of project we could have added different functionality.
First of all the fact to know if the interlocutor is a man or a woman. Indeed, we first tested with another training dataset that allowed us to know if the person speaking was a woman or a man. The problem was that this dataset did not have the same features as the one we used. That's why we started to think about what features were important to differentiate the voice of a woman from that of a man.
The problem is that we could not finish this part because of lack of time. So this is an objective that we could have achieved with more time.

At the beginning, we also thought about wanting our project to be done in real time. That was one of the features we wanted to achieve to try to surpass. After a lot of research we realized that this was not going to be possible.
It came that we were going to miss time because we had to take into account the training part on the subject, the experimentation part and finally the development. Taking all this into account we realized that we could not realize our project in real time.
We still found a solution that would work, the goal would be that our program would work in multi-threading. The biggest advantage with this way of doing things would be the speed of execution. The interlocutor would speak into his microphone and the program would analyze if he is angry, happy or scared.

These are all the goals we think we will achieve in 6 months, but we have thought about thinking bigger. Indeed the final touch of this project would be to make it an executable, the customer would not have to launch any file or external software, just launch the executable. This would make it much faster to use.

The last benefit we could have with extra time would have been on data collection. Our program is trained on 3 emotions at the moment, with more time we could have trained it on more emotions. But also we could have tried to find other emotions than the one in our dataset.

12 Conclusion

To conclude, we believe that our project is a success. Indeed, the goal of this project was to analyze the prosody of a sentence and to deduce emotions. We think we succeeded in our project because we are able to define which emotion is defined in an audio.
We have a success rate of 89% on our audios and our program is able to define precisely which emotion is given in a new audio. As said in the ”outlook for the future” part we think we could have trained our program on more emotions.
We are very happy with the results we have achieved, we have managed to make a program that uses machine learning work and give us very satisfactory results.
We can conclude that this project is a success.

13 References

[1] A. Cutler, D. Dahan, and W. van Donselaar, Prosody in the Comprehension of Spoken Language: A Literature Review

[2] Ivana Didirková George Christodoulides Anne-Catherine Simon, The Prosody of Discourse Markers alors and et in French. A Speech Production Study

[3] Emanuel A. Schegloff, Reflections on Studying Prosody in Talk-in-Interaction

[6] Francois Mairesse Joseph Polifroni Giuseppe Di Fabbrizio, Can Prosody Inform Sentiment Analysis? Experiments on Short Spoken Reviews

[7] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, Recognising real-istic emotions and affect in speech: State of the art and lessons learntfrom the first challenge

[8] S. Yacoub, S. Simske, X. Lin, and J. Burns, Recognition of emotionsin interactive voice response systems

[9] C. M. Lee and S. S. Narayanan, Towards detecting emotions in spokendialogs

[10] https://docs.python.org/3/library/pickle.html

[11] https://github.com/Shahabks/my-voice-analysis

[12] Remya Susan John, Starlet Ben Alex, M. S. Sinith, and Leena Mary, Significance of Prosodic Features for Automatic Emotion Recognition

[13] Gyögy Szaszak, Miklos Gabriel, Tulics Akos, Mate Tüdik, Analysing F0 discontinuity for speech prosody enhancement

[14] https://github.com/aswintechguy/Deep-Learning-Projects/tree/main/Speech%20Emotion%20Recognition
%20Sound%20Classification

[15] https://www.youtube.com/watch?v=LaYGr4ErXn0

[16] https://www.youtube.com/watch?v=iWk9kYGQn_U

[17] https://blog.dataiku.com/speech-emotion-recognition-deep-learning

[18] https://www.kaggle.com/datasets/primaryobjects/voicegender

[19] kurtosis: is a measure of the "tailedness" of the probability distribution of a real-valued random variable

[20] maxdom: maximum of dominant frequency measured across acoustic signal

[21] dfrange: range of dominant frequency measured across acoustic signal

[22] Pavel Kral, Jana Kleckova, Christophe Cerisara,Analysis of Importance of the prosodic Features for Automatic Sentence Modality Recognition