PLAVIDA, an Annotation Tool for Audio and Video in African Languages

Go Issa Traoré(✉), Borlli Michel Jonas Some, Ousmane Ouédraogo, and Lucien Kalmogo

Université Nazi BONI, Bobo Dioulasso, Burkina Faso
goissatraore@yahoo.fr, sborlli@gmail.com, oueo5587@gmail.com, lucienkalmogo21@gmail.com

Abstract. PLAVIDA, PLatform for Audio and VIdeo Data Annotation is a platform designed to facilitate audio and video data annotation. To perform sound classification tasks with Machine Learning algorithms, we need annotated data on these sounds. It is on the basis of this annotated data that these algorithms will learn to make classifications. However, the community lacks labelled audio data on African languages. PLAVIDA will allow researchers the opportunity to create a multimedia labelled databases which can be used as input in Artificial Intelligence models. This could boost research around audio classification in several African languages. We have used python and Android IONIC/Angular technology to develop this tool. The innovation in PLAVIDA, is the possibility given to illiterate people to be able to interact with, when we want to labelle sound or video in African local languages. The tool can be then used both by literate and illiterate people. The type of labelling we are faced on concern the emotional perception people can have when listening or watching a media. It incorporates an annotation logic based primarily on the maximum rate of the same emotional perception over all. In the case where there is no majority vote, the user profile criterion is used. The data annotated using this application can be exported in XML, CSV or JSON format. These types of format are the data formats used to create Artificial Intelligence models.

Keywords: Annotation tool · Audio and video data annotation · Database · African languages

1 Introduction

Audio classification is the process of listening to and analysing audio recordings using computer tools. Otherwise known as sound classification, it is at the heart of various modern Artificial Intelligence technologies, including automatic speech recognition, virtual assistants, text-to-speech applications and so on. It is also used in radio stream analysis, video archiving, audio coding, music classification, auditory scene recognition [7], etc. This work requires well-prepared data based on precise prediction objectives. This allows to apply the appropriate machine

A. Sere et al. (Eds.): AFRICOMM 2023, LNICST 588, pp. 141–153, 2025.
https://doi.org/10.1007/978-3-031-81573-7_11

learning algorithms for the analysis and prediction of a specific aspect in audio. The most important limiting factor in the study of audio classification is the lack of labelled databases. Indeed, the community lacks audio databases:

- of large-sized;
- incorporating a large range of annotations;
- in other languages than English;
- and, most importantly, freely available to all.

To obtain labelled audio data in African languages, we have created an application which allows to annotate a voice with the emotions expressed in it. We called this application PLAVIDA (PLatform for Audio and VIdeo Data Annotation). We have created the application by integrating 3 African languages for the time being: Moore, Dioula and Fulfulde, which are the main languages spoken in Burkina Faso and for which we have voice data available from local radio stations. But the application is designed to be able to add several other languages and to annotate voice data in these languages. The aim of setting up this application is to be able to create speech corpora in various African languages in order to encourage research on audio data in these languages. PLAVIDA would be useful to researchers who wish to conduct studies on audio classification by allowing them to annotate large corpora of data in their language of choice.

Numerous studies have been carried out on audio data analysis [4,9,12,17]. To carry out this work properly, labelled data is required [1,11,14]. Data labelling must follow a rigorous procedure to ensure that each label associated with each piece of data is consistent. To achieve this, some applications have been created to facilitate labelling. Most recent annotation applications are based on Java, use XML for file exchange and have an object-oriented design [10]. MATE [8] is a tool that aims to simplify the tasks of annotating, displaying and querying speech or text corpora. It is designed to help humans to create linguistic resources and to facilitate the use of data by different groups, providing a tool that can be used with many different annotation schemas. Any annotation schema that can be converted to XML can be used with the tool. The tool is written entirely in Java, which means it is platform-independent. The software provides a number of predefined style sheets for use with particular annotation schemes. But its main strength is the ability to write new stylesheets for existing or new schemas. ELAN [16] is an annotation tool for audio and video recordings. With ELAN, a user can add an unlimited number of textual annotations to audio and/or video recordings. An annotation can be a phrase, word, glossary, comment, translation or description of any feature observed in the media. An annotation can be linked to a media item or refer to other existing annotations. Annotation content consists of Unicode text and annotation documents are stored in XML format. EUDICO [18] is an effort to enable multi-user annotation of a centralised corpus via a web interface. The tool should enable multimodal video annotation. EUDICO is based on an existing tool called Media Tagger which is used in various research institutes but requires a special hardware or software configuration. Anvil [10] (Annotation of Video and Language) is a tool for annotating audiovisual content incorporating multimodal dialogues. Anvil is highly generic (usable

with different annotation schemes), platform-independent, based on XML and has an intuitive graphical user interface. For project integration, Anvil allows voice transcripts to be imported and data to be exported in text and table formats for future statistical processing. Annotated data is stored in a single XML file. ATLAS [2]: Architecture and Tools for Linguistic Analysis Systems was born out of the need for applications covering corpus construction, evaluation infrastructure and multimodal visualisation. The main objective of ATLAS is to provide powerful abstractions of annotation tools and formats in order to maximise flexibility and extensibility. ATLAS was inspired by annotation graphs [3], a graph model for linear signals annotation (such as text and speech) indexed by intervals, for which efficient database storage and query techniques are applicable. Common Voice[1] is a web-based platform for collecting voices. It is open to the public and is fed by the voices of volunteer contributors from all over the world. For a language to be included on common voice, a certain number of sentences must be available. The phrases are read by the contributors to create the dataset. ANNEMO (ANNotating EMOtions)[2] is a web-based open-source tool for annotating affective and social behaviours from audiovisual data. In this tool, the annotator must connect to a web-based annotation interface using a unique identifier. The interface is divided vertically into two parts: a scrolling list of audiovisual recordings is presented on the left-hand side, while the video and annotation cursor are displayed one below the other on the right-hand side of the window. All annotation data is automatically saved on a server in the form of log files. This tool has been used to annotate the RECOLA database [15].

As a criticism of these works, first of all, each of these applications was created to annotate an audio database in a specific language, mainly in English, German and French and do not include African languages. Secondly these applications do not allow a sound to be annotated by a large number of annotators, which would not remove any doubt about the annotation of certain sounds that express very similar emotions. Thirdly accessibility to these applications poses a real problem for anyone wanting to set up a database in an African language. Common Voice, to which several languages can be added does not allow emotions annotation; it only allows voice collection. Then none of these tools provides annotated data in both CSV, XML and JSON formats. At last The labelling approach we want to adopt is not taken into account by any of the existing tools, so we cannot use them. It is to overcome all these limitations of existing platforms that we have set up PLAVIDA This paper first presents a methodology in which we present the architecture of PLAVIDA tool and the labelling approach that we have followed. Secondly we present the results by highlighting application's user interfaces, the final structure of the data annotated through the application and the experimentation of the application. Then, we discuss the results and finally we present the conclusion.

[1] https://commonvoice.mozilla.org/fr.
[2] https://diuf.unifr.ch/main/diva/recola/annemo.html.

2 Methodology

To implement this tool, a state of the art was first carried out. The aim of the state of the art was to identify existing tools for audio and video data annotation and to see the applicability of these tools for annotating audio and video in African languages. The existing tools did not take into account African languages and could not be well used to annotate these languages.

Then, we have defined the architecture of the tool we want to implement. This architecture has enabled us to understand the general structure of the system, the relationships between the elements that make up the application and the different functionalities that need to be developed. This structure is the result of a series of strategic decisions we made during the analysis and design phase.

2.1 PLAVIDA Architecture

The architecture of our solution is shown in Fig. 1 and consists of three levels: the application level, the logical level and the physical level.

The Application Level
It relays requests from the user (annotator, administrator) to the logical layer, and in return presents the information returned by the processing of this layer. To facilitate access to PLAVIDA, we have opted to use it through a smartphone. We have therefore implemented this layer through an Android application.

The Logical Level
The logical level includes annotation logic and an API for accessing the database. The annotation logic consists of the majority vote, the user profile and the language to which the annotator belongs. The API defines a set of procedures that describe the operations that the application performs on the data in response to user requests made through the presentation layer. These operations include creating, modifying, searching and storing annotations.

The Physical Level
It is the part that manages access to PLAVIDA data. The storage strategy we are using is an RDBMS (Relational Database Management System). The data can therefore be accessed via SQL queries formulated by the user.

Finally, in order to have audio and video data very well annotated, and based on the work of [6], we have defined the following labelling approach:

2.2 Labelling Approach

We count the number of annotations per sound. Every time a file is annotated, the annotation is recorded in a table with informations about the sound and the annotator. In this table, we have the following informations: the identifier of the audio, the name of the audio, the name of the emotion assigned to it by each annotator, the profile of the annotator and the number of annotations made to this sound. Once a sound has reached the maximum number of annotations, we

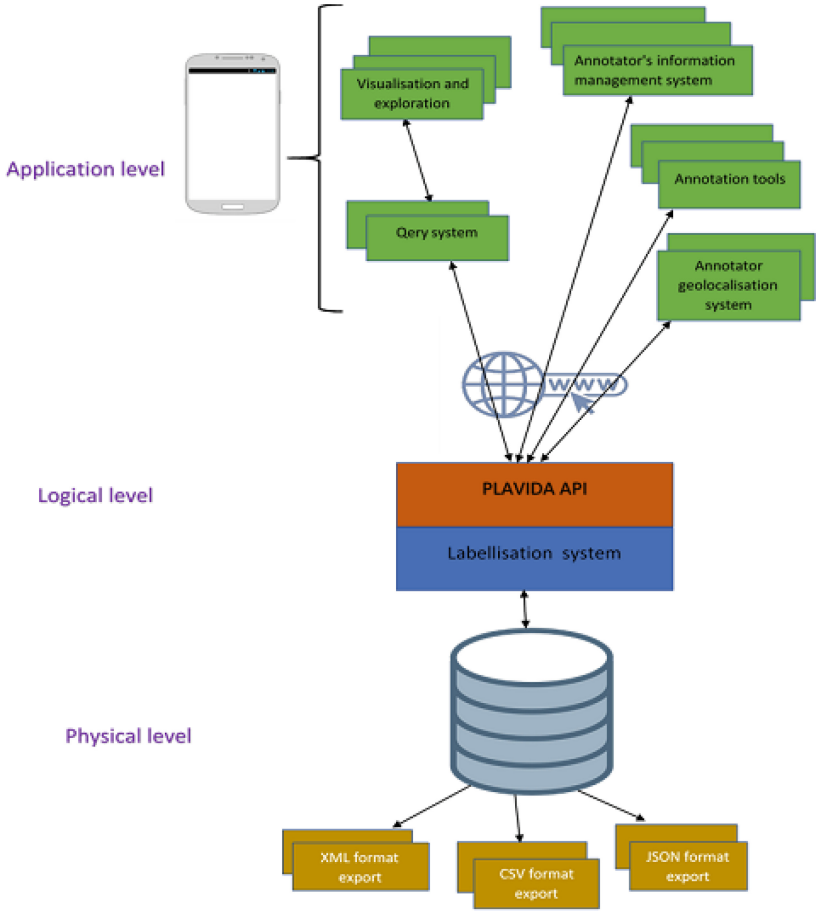


Fig. 1. PLAVIDA Architecture

evaluate that sound and its annotation is stored in a final table. This final table indicates the final emotion associated to this sound. The evaluation of a sound is based on two criteria, one of which is a priority and the other a secondary criterion. The priority criterion consists in choosing an emotion that has received a majority vote. In others words, the number of attribution of this emotion to this sound, as it was the case in the studies of [5]. From a total of 5 annotations for a sound, if at least 3 annotators attribute the same emotion to this sound, then we consider this sound expresses this emotion, whatever the profile of the annotators. But if there is an equality in the annotation of a sound with different emotions, for example on the 5 annotations if two emotions appear each twice in the annotation of a sound, we use the second criterion to decide and choose the emotion that seems to be better adapted. In this case, we use the annotator's profile. The annotation from a user with an illiterate profile will be taken into

account to the detriment of the annotation from a user with a literate profile. If the annotations are made by users with the same profile, it is difficult to distinguish between them. We cancel the annotations for this sound and add it back to the list of audio files to be annotated. This will allow other users to annotate this sound. This process is described by the diagram of the Fig. 2.

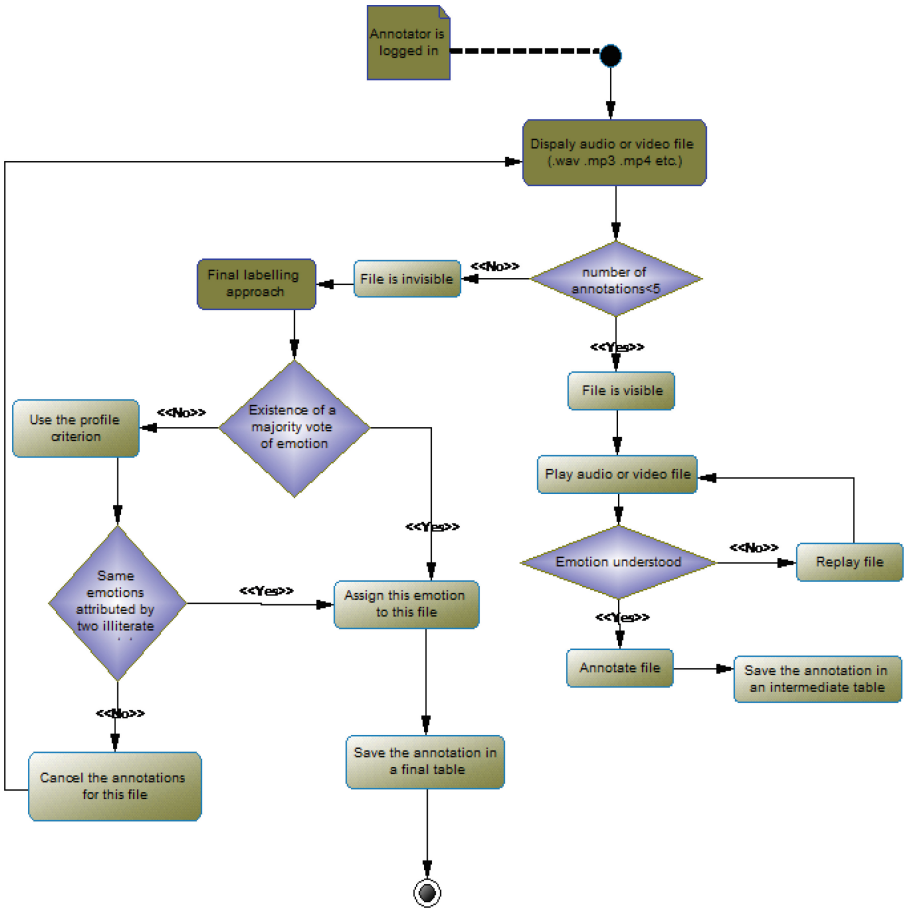


Fig. 2. Labelling logical description diagram

3 Results

3.1 User Interfaces

It consists of four (04) parts:

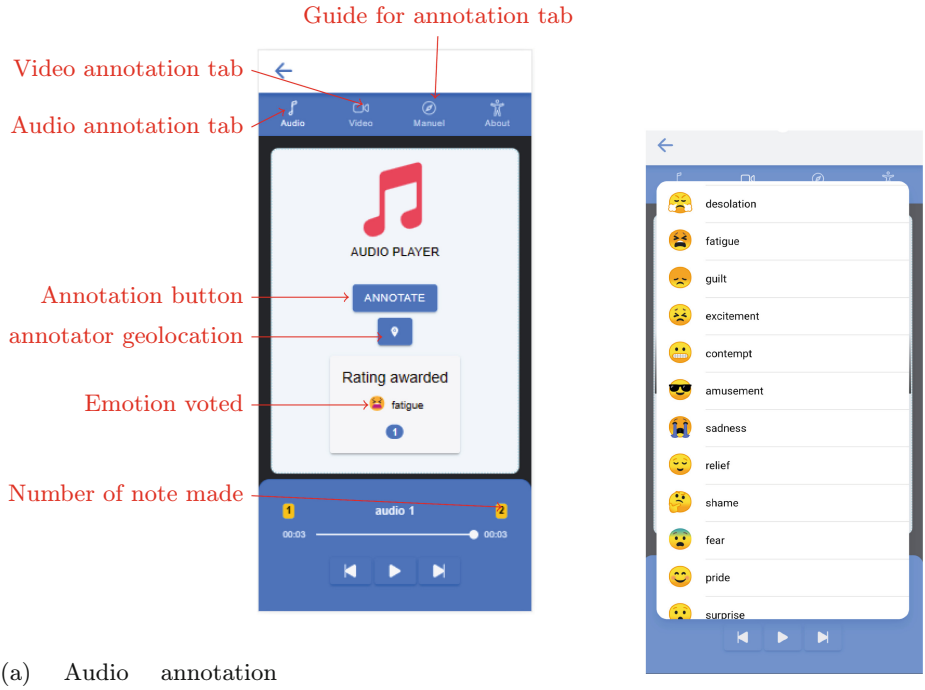
Registration and Login Interface: The annotator registration section provides the annotator's qualifications in the language they are annotating. This section also makes it possible to specify the annotator's profile: "literate" profile and "illiterate" profile. This is an important factor in the reliability of his annotation. In the case of a discrepancy on an annotation, we will consider the annotation made by the "illiterate" profile, who is a native speaker of the language. This is because it is assumed that this person uses the same language daily. To annotate a sound, the user must log in. Logging in allows us to retrieve the annotator's identifiers and find out who has annotated which sound. Figure 3a is the account creation page and Fig. 3b is the login page.

(a) Sign Up page (b) Sign In page

Fig. 3. Account creation and login page

Audio Annotation Interface: This section plays audio files randomly. It also allows users to listen to a sound and assign it an emotion using an emotion button. This button displays a list of emotions with their corresponding emoticons. The presence of the emoticons allows an 'illiterate' user to annotate the sounds without being able to read the name of the emotion. This interface also shows the number of annotations already made for a sound. A sound is only visible on this interface when its maximum annotation has not been reached. We have set the maximum annotation at 5. Some works have considered a maximum annotation of 3 annotators [6]. For reasons of annotation quality, annotators can only

annotate audio files in their mother tongue. Figure 4a is the annotation page and Fig. 4b is the list of emotions used to annotate a sound.



(a) Audio annotation page

(b) Emotion list

Fig. 4. Audio annotation page and motions list

Video Annotation: This is the part that concerns video annotation. In this part, the annotator must take into account three modalities: sound, visual and gestures. A list of videos also appears. On the basis of these three modalities, the annotator assigns an emotion to the video. A list of emotions is also available at this level with emoticons to enable an 'illiterate' person to annotate a video. For more information on multimodal annotation, the annotator should refer to the "annotation guide" section. The Fig. 5 represents the video annotation page.

Annotation Guide: The Annotation Guide screen presents three points:

- The first point explains the emotions and a description of the emotions that appear in the audio and video annotation pages. The aim of this section is to provide a clear understanding of certain emotions that are not well known by certain annotators. It also to highlight the nuances between similar emotions. This section also shows the correspondence between the names of the emotions and the emoticons.
- The second point explains how to annotate a sound and how the labelling will be done after the different annotations;

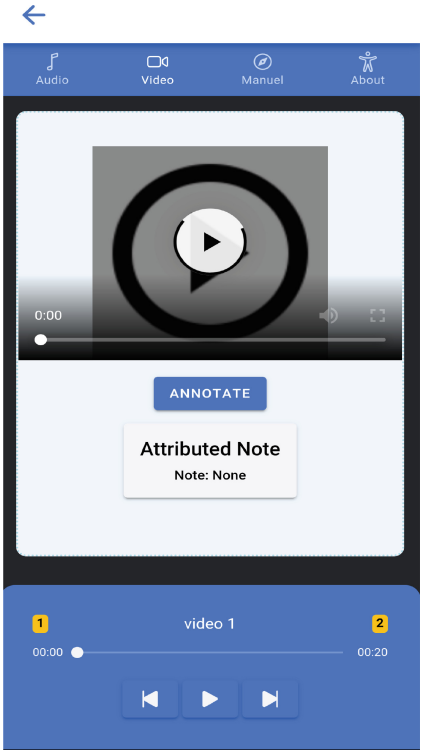


Fig. 5. Video annotation page

– The third point describes multimodal annotation. It presents a description of the three modalities: sound, gesture and visual. It also explains how to perform multimodal annotation.

These presentations are intended to guide annotators who have no knowledge of emotions and annotations. They are not intended to influence any annotator. We recommend annotators to assign emotions to sound and video freely and according to their feelings.

3.2 Description of Final Data Structure

As the aim is to have annotated data available in several languages to facilitate machine learning work, we propose 3 data formats in which anyone wishing to work with their data can export the final table for Machine Learning tasks. These 3 formats are CSV, XML and JSON. These are data formats the most widely used as input to Machine Learning algorithms. The final data contain the informations shown in the Table 1.

Figure 6a shows the json format of the exported data, Fig. 6b shows the xml format and Fig. 6c the csv format.

Table 1. Description of final table content

Column name	Description
Id_audio	corresponds to the audio identifier
audio_language	corresponds to the language in which the audio was spoken
audio_reference	corresponds to the name used to identify an audio in the list of audio files. It is unique for each audio
emotion_name	corresponds to the emotion attributed to an audio
emoji	Corresponds to the code associated with an emotion in csv format

(a) JSON format (b) XML format

(c) CSV format

Fig. 6. Overview of the three possible data export formats

3.3 Experimentation

At the time we are writing this paper, the platform has been used by 20 students (literate profile) from the Nazi BONI University and 20 people (illiterate profile) to annotate the audio in three languages: Moore, Dioula and Fulfulde. A total of 100 audio files of 10 min were entered into the platform. These audio data was collected from YouTube and segmented in order to facilitate the identification of an emotion. For efficiency reasons, each annotator could annotate a maximum of 10 audios during the test. We have not yet done a large-scale evaluation of the software functioning in the context of a large-scale annotation project, which is essential to fully demonstrate the validity of our approach. However, local audio is currently being collected from local radio stations in order to create a corpus of

annotated data in one of the main languages spoken in Burkina Faso. From the 100 audio files, 75 were annotated 5 times on 5 through the application. 3 were annotated 3 times on 5 and the 2 others did not receive any annotation. The 75 files annotated 5 times were evaluated following the process described through Sect. 2.2. The annotated data was successfully exported in CSV format, JSON format and XML format. Note that there were audio files for which, on the 5 possible annotations for an audio, two different emotions were each attributed twice and a third different emotion. These audio files express emotions that are very similar. For example, relief and satisfaction in a speech, hubris and pride in a speech, or shame and embarrassment in a speech, and so on.

4 Discussion

The choice of 5 annotators per sound is motivated by the nature of the audio and video data that will be introduced into the application for annotation. The authors, who limited themselves to 3 annotations [6], recorded the audio and video data themselves in an environment that was already very well prepared. These sentences are written in such a way that they already express the emotions, i.e. the environment is very clear and the quality of the data is high. However, we believe that the domain of expression of the data used by these authors is very restricted and that their emotion recognition models could show limitations when we will use them on data collected in a popular domains such as radios and telephone calls. We will provide the ability to annotate data recorded in popular domains such as speech recorded during radio and television broadcasts. Given that the identification of emotions in these types of audio and video will not be as simple for an annotator as in existing work, we have therefore chosen 5 annotations per sound to allow better labelling of this data.

The other aspect which also justifies the choice of 5 annotators is the target of the annotators. Indeed, most of the existing works target some annotators who are only natives of the language to be annotated and who would have a good experience (high age). But given the multitude of African languages and the large amount of data that will be collected from radio and television, this approach does not make it easy to obtain very large quantities of annotated data. We are going to make it possible to annotate the data for a larger number of people, such as pupils, students, shopkeepers, farmers and many others. To ensure that the data is properly annotated, we need to go beyond 3 annotations per sound. In addition, the annotation made by an illiterate profile should be prioritised in the case of a tie in the 5 annotations, because an illiterate profile annotator has a better experience in his language since he only speaks this language on a daily basis.

In order to make PLAVIDA easier to use and more accessible to annotators, we decided to use Android technology, which is widely used in Africa and is currently growing [13].

For a test phase, we used 100 audio files. This number of files may seem small for a full evaluation of the application, but it gives an idea about how the

application works and its labelling approach. The results of this experimental phase satisfy the objectives set for the introduction of this tool.

5 Conclusion

PLAVIDA, the audio data annotation platform, has been presented: from its architecture to the structure of the final annotated data, passing through the labelling logic. It consists of a frontend implemented by an Android application which uses an API to access the backend managed by the python language. PLAVIDA is an easy, intuitive platform for annotating audio and video data in African languages. It has a convivial interface that is easy to use by all categories of people (literate and illiterate). It also allows annotated audio and video to be available in 3 data formats: CSV, XML and JSON. PLAVIDA would be useful to researchers who wish to conduct studies on audio classification by allowing them to annotate large corpora of data in their language of choice. The basic functionality of the application, which is to annotate audio and video, is operational. The application is easy to install on smartphones and robust. However, there are many functionalities that could be the subject of future work on the application. The introduction of languages and audio data by users needs to be developed, while ensuring the reliability and security of these data. The ability to annotate data offline should be integrated into the application to make it easier to use. Audio and video segmentation functionality needs to be added to the application backend. We are working to integrate these functionality into the application in the near future.

References

1. Bagher Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1208. https://aclanthology.org/P18-1208
2. Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., Liberman, M.: ATLAS: A flexible and extensible architecture for linguistic annotation. http://arxiv.org/abs/cs/0007022
3. Bird, S., Liberman, M.: A formal framework for linguistic annotation. http://arxiv.org/abs/cs/9903003
4. Boukabous, M., Azizi, M.: Multimodal sentiment analysis using audio and text for crime detection. In: 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5 (2022). https://doi.org/10.1109/IRASET52964.2022.9738175
5. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
6. Etienne, C.: Apprentissage profond appliqué á la reconnaissance des émotions dans la voix. https://theses.hal.science/tel-02479126

7. Flocon-Cholet, J.: Classification audio sous contrainte de faible latence. Theses, Université de Rennes (2016). https://theses.hal.science/tel-01395495
8. Isard, A., McKelvie, D., Mengel, A., Møller, M.B.: The MATE workbench annotation tool, a technical description
9. Jiang, H., Wu, X., Xie, X., Wu, J.: Audio public opinion analysis model based on heterogeneous neural network. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 449–453. https://doi.org/10.1109/ICCECE51280.2021.9342052
10. Kipp, M.: Anvil-a generic annotation tool for multimodal dialogue. In: Seventh European Conference on Speech Communication and Technology. Citeseer (2001)
11. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE **13**(5), e0196391. https://doi.org/10.1371/journal.pone.0196391. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391
12. Luitel, S., Anwar, M.: Audio sentiment analysis using spectrogram and bag-of-visual- words. In: 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), pp. 200–205 (2022). https://doi.org/10.1109/IRI54793.2022.00052
13. Ouedraogo, I., Some, B.M.J., Benedikter, R., Diallo, G.: Mobile technology as a health literacy enabler in African rural areas: a literature review. preprint, In Review (2021). https://doi.org/10.21203/rs.3.rs-243773/v1. https://www.researchsquare.com/article/rs-243773/v1
14. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964. ISSN: 2379-190X
15. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2013). https://doi.org/10.1109/FG.2013.6553805
16. Sloetjes, H., Wittenburg, P.: Annotation by category - ELAN and ISO DCR
17. Yin, Y., Hanes, D.W., Skiena, S., Clouston, S.A.P.: Quantifying healthy aging in older veterans using computational audio analysis. J. Gerontol. Ser. A glad154 (2023). https://doi.org/10.1093/gerona/glad154
18. Zheng, Y., Peng, J.E.: ELAN (EUDICO linguistic annotator). RELC J. **53**(2), 469–474 (2022). https://doi.org/10.1177/00336882221089052