Speech Processing : a literature review

Go Issa TRAORE[1], Borlli Michel Jonas SOME[2]
{goissatraore@yahoo.fr[1], sborlli@gmail.com[2] }

Université Nazi BONI, Bobo-Dioulasso, Burkina Faso[1,2]

Abstract. In this paper, we have focused our research on the state of the knowledge on speech processing and the research perspectives that exist in this domain. This research was conducted on several digital libraries such as IEEE Xplore, ScienceDirect, arXiv, Springer Link, Papers With Code etc. The research focused on the types of speech classification, the techniques used to extract speech features, the Machine Learning (ML) techniques used and the speech data sources available. We found that studies focused mainly on emotion recognition, dialect identification in speech and speaker recognition. Mel Frequency Cepstral Coefficents (MFCC) is the main and most widely used for speech feature extraction. Neural networks dominate as ML techniques for speech classification. Speech databases available have been built in different contexts. Each database is specific to a given language, mainly English, German, Arabic, Chinese and French. There are almost no speech databases for low-resource languages, particularly african languages.

Keywords: speech processing, machine learning, speech features extractor, speech database.

1 Introduction

Speech is a time-varying signal that carries several layers of information. The information contained in speech is observed in both temporal (sec) and frequency (Hz) dimensions. Speech classification consists in extracting these informations and classifying them into predefined classes. Nowadays, the existence of speech data is no longer a problem for this type of work. Voice data is produced and stored by many audio-visual structures such as radio stations, television channels, mobile phone companies, social networks, etc. The success of certain social networks such as WhatsApp is largely based on the integration of voice. Google's voice assistants have also enjoyed undeniable success. To implement Google's voice assistants, Artificial Intelligence models for speech recognition were first built. But these types of work are the most difficult topics in data science [1]. It is also a complex task, involving two key issues: feature extraction and classification. This study focus specially on speech classification which is a set of problems or tasks in which a computer program classifies speech automatically into different categories, for example speech command recognition, speech activity detection, and speech sentiment classification. But, this field

JRI 2023, December 18-20, Ouagadougou, Burkina Faso

DOI 10.4108/eai.18-12-2023.2348132

of research is little known and very little exploited in some regions of the world, such as in Africa, in comparison with studies using textual data (text classification). However, a lot of information is conveyed in speech, particularly in local African languages, which are not studied.

This paper aims to present the different types of work existing in this field, the speech features extractor used for this purpose, the main technique use to extract feature from speech and databases used for speech classification. It also aims to give an overview of the potential research topics not yet exploited in this domain. The rest of the content is organized as follows: firstly, we present the issues and purpose of the study. Secondly, we present the materials and methods that enabled us to carry out this study. Then we present the results obtained. In the next section, we discuss these results. We finish with a conclusion.

2 Issues and purpose of the study

There exist studies which allows to analyze and understand human expressions and opinions in a given context. These studies include opinions analysis, sentiments analysis or someone characterization through his written. But these studies are based on textual format databases collected on social network (twitter, Facebook etc.) [2] [3] [4] [5] or on other forums (e.g. a forum on stock exchange shares etc.). To express for example an opinion on these forums and these social networks, written and spoken knowledge are required in the language in question. But, many languages are not written particularly in Africa and those which are written are hardly used in social network discussions. As a result, most analysis work is mainly based on English, French [6] and recently in Bambara. However, many opinions and sentiments are expressed through speech in local African languages which do not require the ability to read and write these languages. It is therefore interesting to go to these sources to analyze them in order to detect Humans expressions through speech. But before being able to carry out analysis work on speech, a certain number of questions must be asked in order to understand this domain such as: what types of studies exist on speech analysis ? What methods are used to extract features from speech? what algorithms are used to classify speech? what speech databases are already used by the scientific community?

The purpose of this study is to provide answers to the above questions mainly to give clear guidance to those who would like to do speech analysis.

3 Materials and methods

We used the bibliographic management software Zotero, in which we created folders by keyword and saved the articles. This allowed us to easily export the bibliographic references in .bib format. We followed four steps to conduct this study: the first was the selection of scientific publication sources; the second was the keyword search; the third was the selection of studies; and the fourth was extraction and analysis of speech data.

3.1 Scientific publication sources

Mainly six digital libraries were considered for this study. These libraries are the most popular and the most important in the computer science domain and contain important scientific resources on Machine Learning and speech processing. These libraries are :

IEEE Xplore[1], Springer Link[2], Papers with code[3], arXiv[4], ScienceDirect[5],HAL[6].

3.2 Keywords search

We have formulated a list of keywords according to which the research was conducted. These keywords were formulated in English and are summarised in the table 1.

Table 1: Keywords used

Keywords	Objective
Speech classification	know the types of studies on speech, the methods used to do this studies
Speech analysis	
Speech processing	
Speech emotion recognition	
automatic speech recognition	
Speaker recognition	
Dialect identification	
Speech to text	know the algorithmic approaches used to carry out this work
Speech features extraction Machine Learning and speech classification	
Speech database	know the existing speech data sources and their characteristics

3.3 Selection of studies

The differents stage of the selection are represented on the **figure 1** . Articles considered importants according to the keywords used were registered in the bibliographic management software Zotero to allow better management of bibliographic references.

[1] `ieeexplore.ieee.org`
[2] `link.springer.com`
[3] `paperswithcode.com`
[4] `arxiv.org`
[5] `sciencedirect.com`
[6] `https://hal.science/`

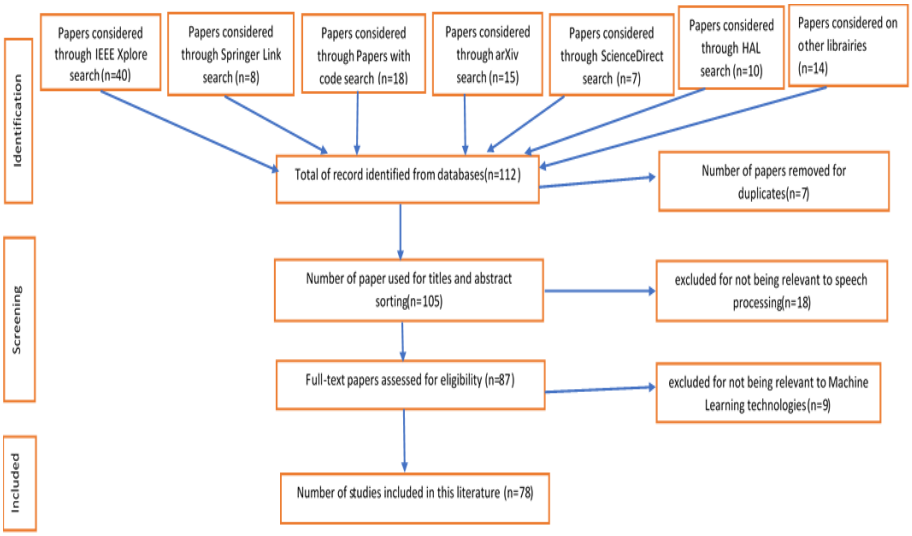

Fig. 1. Studies selection process workflow diagram

3.4 Extraction and analysis of speech databases

The speech databases that we have analysed are the most used and the most cited in the studies. The frequency of these databases in these studies shows their importance and relevance for speech analysis studies. After identifying these databases, we study them. The study consisted of reading the documentation of these databases, downloading and visualising their content and their structuration.

4 Results

The papers considered for this study include scientific articles, doctoral theses and a book. The results we obtained are divided into four categories: the types of work on speech, the speech features extraction methods, the most famous algorithms used in speech analysis and the speech data sources used.

4.1 The existing work on speech analysis

Most speech classification work focuses on speaker recognition, emotion recognition, dialect identification in languages and multimodal emotions analysis.

4.1.1 Speaker recognition

Speaker recognition is a task that identifies the speaker from multiple speech using machine [7]. These systems have two uses[8]:speaker verification and speaker diarisation. There is a lot of studies in this area. Mohammed Lataifeh et al. [9] developed a Quranic reader recognition system. They used the Quranic readings of thirty (30) famous Quranic readers from six major Arab countries namely Egypt, Saudi Arabia, Kuwait, Yemen, Sudan and the United Arab Emirates. Their system recognises Quranic readers based on the extraction of MFCC parameters.

In the same order, Muhammad Mohsin Kabir et al. [10] conducted a study on speaker recognition especially on its fundamental theories, recognition methods and opportunities. They identified three main sections of a speaker recognition system: data preprocessing, feature extraction and speaker modelling. They also identified three main approaches to speaker recognition: automatic speaker identification, speaker verification and speaker diarisation.

4.1.2 Speech emotion recognition

Speech emotion recognition is the process of recognising the emotion of a speech independently of its semantic content. In this sense, Muljono et al. [11] performed emotion recognition in Indonesian film audio. The audio was classified into four classes of emotions, namely: angry, sad, happy and neutral. They used mel-frequency cepstral coefficients (MFCC) to extract the features and SVM to do the classification of the data. Using spectrogram as a feature extraction method, deep convolutional neural network and EMO-DB as database, Abdul Malik Badshah et al. [12] proposed a model for speech emotion recognition. The model provides predictions for the seven classes of emotions: neutral, fear, anger, happiness, sadness, disgust and boredom. Abdul Qayyum et al. [**?**] proposed a model of speech emotion recognition using SAVEE database. To extract emotional characteristics, they used the Modulation Spectral Features (MSF) and the MFCC. To predict the emotions in the classes Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise, two different classifiers were chosen:Support Vector Machines(SVM) and Recurrent Neural Networks (RNN) .

4.1.3 Dialect identification

The term dialect refers to a regional or social variety of a language which is distinguished by its pronunciation, grammar or vocabulary. One of the advantages of dialect identification is that it allows us to discover the regional origin of the speaker or his social affiliation. Many studies have focused on the identification of dialects through speech. According to Sadam Al-Azani et al. [13] the problem of dialect detection from emotional videos is more difficult because these data carry several attributes that are difficult to be modelled such as feelings, thoughts, behaviours, moods, temperaments, etc. They proposed an Arabic dialect identification model. They used Egyptian, Levantine, Gulf and North African dialect classes. The data used was collected from YouTube and consisted in 59 native speakers from various Arab countries. Tanvira Ismail et al. [14] described a Gaussian mixture model (GMM) for identifying the Kamrupi dialect by extracting spectral features from speech data using Mel's Cepstral Frequency Coefficient (MFCC). Gu Mingliang et al. [15]

proposed a Chinese dialect identification model using a Clustered Support Vector Machine (CSVM). The database used in this study includes four main dialects: the North China dialect, the Wu dialect, the Guangdong dialect and the Fujian dialect.

4.1.4 Multimodal emotions analysis

Multimodal analysis consists in relating linguistic information produced in different modalities. Each of which contributes to the elaboration and perception of the communicated message. Thus, we can distinguish the verbal modality, which comprises several levels (phonemes, choice of lexicon, syntactic organisation, discursive organisation); the oral modality (prosody, voice quality) and finally the visual modality (gesture and facial expressions)[16]. Many recent studies have been conducted on this topic[17],[18]. Aman Shenoy et al. [19] proposed a model for multimodal emotion detection and sentiment analysis in conversations using a recurrent neural network (RNN). They took into account three important factors: the context of the conversation, the dependency between the emotional states of the listener and the speaker, and the relevance and relationship between the available modalities. Their model uses three types of data: textual, visual and acoustic. Jean-Benoit Delbrouck et al. [20] considered three modalities in their work: Linguistic (L), Acoustic (A) and Visual (V). Their model predicts feelings (negative, weakly negative, neutral, weakly positive and positive). Emotions are split into six classes: happy, sad, angry, disgusted, surprised and fearful.

Research is now focusing much more on multi-modal analysis of speech. Multimodal analysis allows to understand speech better because it takes several factors into account.

4.2 Speech features extraction methods

According to Jiang [21], there are two methods for processing speech data, one is to extract features directly from the original speech, and the other is to convert the speech into text.

4.2.1 Extraction of speech features

In recent years, techniques that process the speech signal have been developed with fewer requirements for Natural Language Processing (NLP) methods. These techniques have the advantage that the recognition is invariant to the language. To extract the speech features, many techniques exist: MFCC[22], wav2vec 2.0[23], HuBERT[24], spectrogram[25], Neural Networks[26]; etc. The MFCC (Mel-frequency cepstral coefficients) is the most popular representation of an speech signal [27]. MFCC components are the most representative feature of audio description [28]. It allows to performs better than experts in some case. It is the case in the study of Mohammed Lataifeh et al.[9]. we will present this method in more detail in the section 4.2.3. Other feature extraction methods exist: Abdul Malik Badshah et al. [12] used only spectrogram as a features extraction method to do emotion recognition in speech. Then, they used a deep convolutional neural network architecture on the Berlin EMO-DB database to predict emotions in the classes: neutral, fear, anger, happy, sad, disgust and boredom. Lim et al. [29] transformed the speech signal into a 2D representation using the Short Time Fourier Transform (STFT). Then, the 2D representation is analysed through CNNs and Long Short-Term Memory (LSTM) architectures to do speech emotion recognition. There are

also many combinatorial approaches associating MFCC. These methods are: MFCC-DBN (Deep Belief Network), MFCC-CNN (Convolutional Neural Network) and MFCC-RNN (Recurrent Neural Network) [28].The wav2vec 2.0[23] which is a self-supervised extractor is also a technique that is increasingly used in recent times.

The problem with MFCC is that it only works well when the quality of the data is very good. In the case of telephone or radio conversations, where there is very often noise and interaction, MFCC is no longer able to extract the audio characteristics very well. However, it is through these channels that people express the most. In this case, models such as PLP(Perceptual Linear Prediction), LPC(Linear Predictive Coding) and auto-encoder models such as BERT are better adapted for representing the audio signal than MFCC. Research is now focusing on combining the MFCC with other feature extractor methods in order to surpass the accuracies obtained with the use of the simple MFCC in the state of the art. As the MFCC is now the better features extractor, we will present how it works in section 4.2.3.

4.2.2 Transcription (speech to text)

There are three possible methods for real-time speech-to-text conversion: speech recognition, computer-assisted note-taking and computer-assisted real-time translation [30].

The most practical use of Speech To Text (STT) is for broadcasting and transcribing voice messages. Automatic speech recognition (ASR) is one of the applications of STT. After conversing speech to text, many techniques are used by researchers to perform classification. Among these techniques, Support Vector Machines (SVM) is one of the most widely used. Also many successes have been achieved in various fields with Naive Bayes Multinomial (NBM). This technique is known as a supervised statistical learning algorithm based on Bayes' theorem. It is generally used for textual classification [31].

4.2.3 Presentation of MFCC

The objective of using this method is to create the voiceprint from the speech signal. This voiceprint will allow us to have the characteristics of the speech signal. There are other methods for extracting speech features [32], [33], [34], [35], but the cepstral parameters obtained from the MFCC (Mel Frequency Cepstral Coefficients) method continue to be used for more than twenty years. The advantage of using MFCC is to improve the signal-to-noise ratio compared to the raw signal without the need for external paralinguistic expertise. It also has the advantage of being closer to the original audio signal [36]. The MFCC is composed of six (06) phases:

Phase 1: Speech signal is split into several overlapping windows;

Phase 2: In order to reduce the spectral distortion created by the overlap, a Hamming window is applied to the signal. See the formula 1 for Hamming window process.

$$w = 0.5 + 0.46 * \cos\left(\frac{2\pi * n}{N-1}\right) \tag{1}$$

where,

n = sample input index in time domain

N = number of input samples.

Phase 3: At this stage, the Fast Fourier Transform (FFT) is applied to the window to extract its constituent frequencies, we thus obtain the spectrum. FFT is donne by the formula 2.

$$F(n) = \sum_{k=0}^{N-1} U(n) * \exp\left(-jn\frac{2\pi}{N}k\right) \tag{2}$$

where,
exp(jn) = cos(n) + j*sin(n)
N = the number of input samples
F(n) = the k sequence of FFT output components
n = the output index in the frequency domain
U(n) = the n sequence of input sample
k = the input sample index of the time domain.

Phase 4: The spectrum produced by this decomposition is modulated before being filtered by a triangular filter bank following the Mel-scale [37]. This filter bank simulates the perception of frequencies by the human ear.The following formula 3 is used to obtain Mel-scale

$$mel(f) = 2995 * \log 10\left(1 + \frac{f}{700}\right) \tag{3}$$

where,
f=sample-rate.

Phase 5: After this filtering, the logarithm of the resulting values is calculated to obtain the spectral envelope in decibels. This is known as the Log Filter Bank (FB). The process of logarithm is carry out by the formula 4.

$$C[k] = \log 10\,(mel * spectrogram[k]) \tag{4}$$

where,
spectrogram[k] = the k-sequence of spectrogram coefficient
k = the spectrogram index in a frequency domain.

Phase 6: Finally, if we apply an inverse Fourier transform to these FB parameters using a Discrete Cosine Transform(DCT), we obtain the MFCC coefficients.

The MFCC coefficients represent the speech as so-called static information. To take into account the dynamics of the parameters, first and second time derivatives can be added [38]. The first derivative represents the rate of spectral variation, while the second derivative measures its acceleration. The DCT process to obtain MFCC is shown by the following formula 5.

$$S[j] = \sum_{n=0}^{N-1} s[n] * \cos\frac{\pi}{N}\left(n + \frac{1}{2}\right) * j \tag{5}$$

where,
N = number of input samples
S[j] = The j sequence of DCT output components

j = The DCT index output in frequency domain
s[n] = The n sequence of input samples
n = sample input index in the time domain.
The MFCC extraction steps are shown in datail by **Figure 2**.

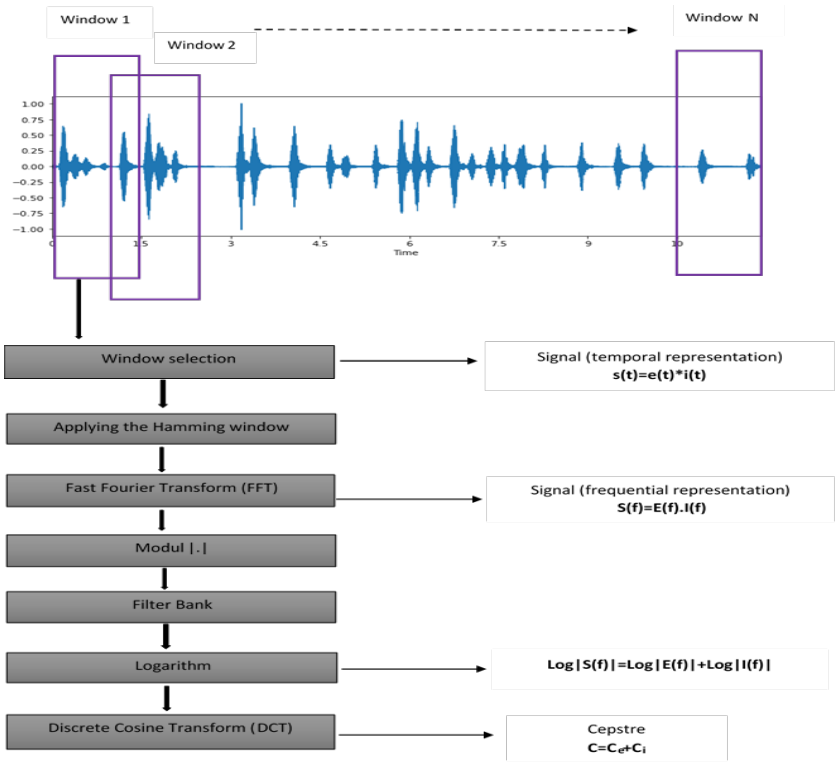

Fig. 2. MFCC extraction steps

4.3 Machine Learning algorithms used

Most of the studies on speech analysis use machine learning algorithms. According to Mohammed Lataifeh et al [9], machine learning models provide additional performance than experts in the field of speech data processing. These algorithms mainly include convolutional neural networks (CNN), recurrent neural networks (RNN) and Long Short Term Memory (LSTM) , or their combination [12], [29], [39]. There are also many studies that use SVM [11], [40].

In the paper of Wootaek Lim et al [29], they proposed a method for analyzing sequential speech data based on the concatenation between CNN and RNN. By applying their architecture on a public emotional speech database, the result of emotion recognition gives better results than conventional

methods such as simple CNN, simple RNN etc. To study whether the speech emotion recognition is language independent, Fardin Saad et al [41] used SVM to classify speech using English and Bangla. They predicted the emotions: joy, anger, neutral, sadness, disgust and fear in these two languages.

Nowadays, deep learning is used to solve many recognition problems, for example, image recognition [42], voice recognition [43] , face recognition [44], speech emotion recognition [45]. One of the main advantages of deep learning techniques is the automatic selection of features.

Research is focused on proposing new algorithms that combine several types of neural network. These new algorithms often give better results than using a simple type of neural network.

The limitation of supervised models, which are the most widely used, is that they base on what we give then as annotated data to do the classification. If the data is poorly annotated, the results will also be poor. The potential of auto-encoding, self-supervised learning and unsupervised learning models is not sufficiently exploited. These models are very interesting for overcoming data labelling problems and for low-resource languages such as African languages.

5 Description of some data sources used for speech analysis

We propose a description of the most used data sources in order to highlight their characteristics and their uses. Nowadays, there are many speech databases. These databases are of two types: databases built solely on voice and multimodal databases. Multimodal databases are databases that are labelled not only on the basis of voice, but by considering several modalities such as voice, visuals, gestures, the context of the conversation, and so on. These data sources are in the table 2.

Table 2: Presentation of some data sources used for speech analysis

Database name	characteristics	use
Some simple speech databases		
TESS(Toronto Emotional Speech Set)[46]	composed of 2,800 sound recordings made by two actresses (aged 26 and 64), Labelled by a group of 56 students, The TESS includes each of the seven emotions (anger, disgust, fear, joy, pleasant anger, sadness and neutral), data are in English. Data set is available on : `https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess`	used in speech recognition studies [Pc Thirumal et al, 2021], for speech emotion recognition [47], [48].
EMODB (Berlin Database of Emotional Speech)[49]	consisting of a total of 535 utterances, Recorded by 5 men and 5 women, It is composed of seven emotions: anger, boredom, anxiety, happiness, sadness, disgust and neutral data are in German. Data set is available on: `https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb`	It is used in speech emotion recognition studies [50],[51].

RAVDESS(Ryerson Audio-Visual Database of Emotional Speech and Song)[52]	Recorded by 24 people (12 men and 12 women), RAVDESS contains 7,356 files (total size: 24.8 GB), Eight emotions are expressed in this database: sad, happy, angry, calm, fearful, surprised, neutral and disgusted, data are in English. Data set is available on https://zenodo.org/records/1188976	Used in sentiment analysis in speech [53], for speech emotion recognition [54].
LibriSpeech[55]	A collection of approximately 1,000 hours of speech data, Each validation and test data set contains 20 male and 20 female speakers. data are in English. Data set is available on: https://www.openslr.org/12	Identification [56], speech recognition [57]
Some multimodal speech databases		
CMU-MOSEI(CMU Multimodal Opinion Sentiment and Emotion Intensity)[58]	it contains approximately 23,453 videos from over 1000 YouTube speakers on 250 topics, Videos are transcribed and correctly punctuated, it takes into account speech, face, context modalities etc, data are in English. Data set is available on: https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK	This database is used in multimodal sentiment analysis and for speech emotion recognition [59], [19]
CMU-MOSI (Multimodal Corpus of Sentiment Intensity)[60]	a collection of 2199 opinion videos, Each video is annotated with a sentiment in the range [-3,3] and consists of a collection of over 1000 speakers on YouTube data are in English. Data set is available on: https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK	It is used for multimodal sentiment analysis and emotion recognition. [61][62]
Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)[63]	Recorded from ten actors in dyadic sessions with markers on the face, head and hands, The corpus contains about 12 h of data. It consists of 151 videos of recorded dialogues, with 2 speakers. Contains 9 emotions: anger, excitement, fear, sad, surprised, frustrated, happy, disappointed and neutral, data are in English. To obtain the dataset, send a request using the following form: https://docs.google.com/forms/d/e/1FAIpQLScBecgI2K5bFTrXi_-05IYSSwOcqL5mX7dh57xcJV1m_NoznA/viewform	This database is used in simple emotion recognition tasks [64], and also in multimodal emotion analysis tasks [65].

6 Discussion

6.1 Application

Most of the speech classification work focuses on speaker recognition, emotion recognition, dialect identification and multimodal emotion analysis. These works are based on small speech data which are mainly in English, Arabic, German, Chinese and French. Very little research has been done on speech analysis in low-resource languages, particularly African languages. Common voice, which integrates the voice collection in African languages, has not yet provided a large dataset that can be used for speech recognition. However, there are other areas of speech applications that have not yet been studied, such as :

- Opinions and sentiments analysis through audio-visual media. This would be important for identifying people's viewpoints and their expectations when he express themselves ;
- The identification of expressions related to a given phenomenon (e.g: terrorism, war, theft, call to violence etc) through speech conversations;
- Speech recognition in low-ressource languages. This would be of interest for language identification in speech;
- Automatic speech translation in low-ressources languages. This would be of great importance in removing language barriers between communities. It would allow people to follow conferences and seminars (e.g: scientific or religious) in other local languages;
- Transcription from speech to text in low-ressources languages. This would be important for discourse analysis in local languages.
- Pronunciation recognition is an interesting subject that merits study nowadays. It will allow for example to evaluate the degree of a person's expression in a given language.

These shortcomings are due to the fact that not enough researchers are interested in audio data. Also, existing speech representation techniques are not very well adapted. In deed, they allow us to process audio data of short duration and require the data to be of good quality and without noise. These representation techniques and existing machine learning models can be improved in such a way to be able to automatically eliminate noise and automatically slice long audio according to some topics encountered. This would make it possible to solve these problems with very accuracy.

6.2 Models

The existing speech analysis studies use mainly supervised methods. Thinking about self-supervised methods and unsupervise methods for low-resource languages would be much more interesting than using supervised methods. These methods do not require labelled data. However, the serious problem with supervised learning is setting up a labelled dataset. Labelling data is very costly in terms of time and resources. We have also seen through this state of the art that low-resource languages do not yet have labelled databases. Also, self-supervised and unsupervise methods have given interessing results in others domains like images recognition and text classification.

6.3 Data preparation

Researchers are much more focused on finding the best classification models. Yet speech analysis involves three issues: data quality, speech features extraction and classification. A model that classifies best depends on the quality of the data and the feature extractor. It would be more advantageous to look for ways of obtaining excellent data quality. For example, when labelling data, use mathematical theories such as graphs to make the choice and assign a good label to a speech. We should also include experts in the domain of science of language to take account of all the aspects that make it possible to understand a language, instead of relying on majority votes. In terms of data quality, research should also focus on automatic speech cleaning solutions to remove noise, interactions and other things that can compromise data quality. Data slicing and labelling is a crucial step, and very costly in terms of financial, human and time resources. Setting up a system for collecting, slicing and automatically annotating data is a challenge that remains to be solved.

6.4 lingustic characteristics

In order to carry out speech classification work on some languages such as African languages, which are tone languages, the extraction of speech signal characteristics using the fundamental frequency (F0) would give better results for these languages than the MFCC. This is because sounds can be distinguished either by their pitch or by their timbre. Pitch is the perceived note and timbre is the perceived signal shape. The F0 measures the pitch of the sound, which corresponds to its frequency of vibration, measured in hertz, while the MFCC measures the timbre. Nasalisation and vowel length are linguistic features that should be included in feature extraction. Vowel length is the doubling of certain vowels which a different meanings. For example, in "Moore", which is the main language spoken in Burkina Faso, **nwã** and **nwa** are semantically distinct. Also **Zaabre** means evening, whereas **Zabre** means a fight. It is the same for **n peege**, which means to accompany opposed to **n pege**, which means to wash.

7 Conclusion

This paper presented a literature review on speech classification, focusing on the types of works, audio characteristic extractors, algorithms and databases used for speech classification. The results of our research show that the majority of studies in this domain have focused on speaker recognition, speech emotion recognition, dialect identification in languages. Recent studies has been increasingly interested in the multimodal analysis approach. MFCC is the most widely used method for speech processing. neurone networks are most commonly used to classify speech. Several speech databases have been established by researchers with specific purposes in order to facilitate these kinds of studies. But the dominant languages in these databases are English, Arabic, German, Chinese and French. We have discussed some topics that may be of interest to research in the field of speech analysis. We have also discussed the limits of existing work.

It would be interesting to analyse other phenomena in speech besides emotions, dialects or speakers. The establishment of usable speech databases in low-resource languages could considerably develop researches on these languages.

References

[1] Issa D, Fatih Demirci M, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control. 2020 May;59:101894. Available from: https://www.sciencedirect.com/science/article/pii/S1746809420300501.

[2] Abdullah NSD, Zolkepli IA. Sentiment Analysis of Online Crowd Input towards Brand Provocation in Facebook, Twitter, and Instagram. In: Proceedings of the International Conference on Big Data and Internet of Thing. ACM;. p. 67-74. Available from: https://dl.acm.org/doi/10.1145/3175684.3175689.

[3] Patodkar VN, I R S. Twitter as a Corpus for Sentiment Analysis and Opinion Mining;5(12):320-2. Available from: http://ijarcce.com/upload/2016/december-16/IJARCCE%2074.pdf.

[4] Vepsäläinen T, Li H, Suomi R. Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections;34(3):524-32. Available from: https://www.sciencedirect.com/science/article/pii/S0740624X16301411.

[5] Gokulakrishnan B, Priyanthan P, Ragavan T, Prasath N, Perera A. Opinion mining and sentiment analysis on a Twitter data stream. In: International Conference on Advances in ICT for Emerging Regions (ICTer2012). IEEE;. p. 182-8. Available from: http://ieeexplore.ieee.org/document/6423033/.

[6] Mountassir A, Benbrahim H, Berrada I. Sentiment classification on arabic corpora;16(1):73-96. Available from: https://www.cairn.info/revue-document-numerique-2013-1-page-73.htm.

[7] Tan H, Wang L, Zhang H, Zhang J, Shafiq M, Gu Z. Adversarial attack and defense strategies of speaker recognition systems: A survey. Electronics. 2022;11(14):2183.

[8] Brown A, Huh J, Chung JS, Nagrani A, Garcia-Romero D, Zisserman A. Voxsrc 2021: The third voxceleb speaker recognition challenge. arXiv preprint arXiv:220104583. 2022.

[9] Lataifeh M, Elnagar A, Shahin I, Nassif AB. Arabic audio clips: Identification and discrimination of authentic Cantillations from imitations. Neurocomputing. 2020 Dec;418:162-77. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0925231220312881.

[10] Kabir MM, Mridha MF, Shin J, Jahan I, Ohi AQ. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. IEEE Access. 2021;9:79236-63. Conference Name: IEEE Access.

[11] Muljono, Prasetya MR, Harjoko A, Supriyanto C. Speech Emotion Recognition of Indonesian Movie Audio Tracks based on MFCC and SVM. In: 2019 International Conference on contemporary Computing and Informatics (IC3I); 2019. p. 22-5.

[12] Badshah AM, Ahmad J, Rahim N, Baik SW. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In: 2017 International Conference on Platform Technology and Service (PlatCon); 2017. p. 1-5.

[13] Al-Azani S, El-Alfy ESM. Audio-Textual Arabic Dialect Identification for Opinion Mining Videos. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI); 2019. p. 2470-5.

[14] Ismail T, Deka GK, Dutta SK, Singh LJ. Kamrupi Dialect Identification Using GMM. In: International Conference on Signal Processing (ICSP 2016). Vidisha, India: Institution of Engineering and Technology; 2016. p. 3 (4 .)-3 (4 .). Available from: https://digital-library.theiet.org/content/conferences/10.1049/cp.2016.1442.

[15] Gu Mingliang, Xia Yuguo, Yang Yiming. Semi-supervised learning based Chinese dialect identification. In: 2008 9th International Conference on Signal Processing. IEEE;. p. 1608-11. Available from: http://ieeexplore.ieee.org/document/4697443/.

[16] Ferre G. Analyse multimodale de la parole. Rééducation orthophonique. 2011;246:73. Available from: https://hal.science/hal-00609124.

[17] Zhang L, Ruan L, Hu A, Jin Q. Multimodal Pretraining from Monolingual to Multilingual. Machine Intelligence Research. 2023;20(2):220-32.

[18] Sulubacak U, Caglayan O, Grönroos SA, Rouhe A, Elliott D, Specia L, et al. Multimodal machine translation through visuals and speech. Machine Translation. 2020;34:97-147.

[19] Shenoy A, Sardana A. Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. In: Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML); 2020. p. 19-28. ArXiv:2002.08267 [cs, eess]. Available from: http://arxiv.org/abs/2002.08267.

[20] Delbrouck JB, Tits N, Brousmiche M, Dupont S. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In: Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). Association for Computational Linguistics;. p. 1-7. Available from: https://aclanthology.org/2020.challengehml-1.1.

[21] 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); 2021. p. 449-53.

[22] Mahmood A, Utku K. Speech recognition based on convolutional neural networks and MFCC algorithm. Advances in Artificial Intelligence Research. 2021;1(1):6-12.

[23] Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc.; 2020. p. 12449-60. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[24] Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:3451-60.

[25] Shah VH, Chandra M. Speech recognition using spectrogram-based visual features. In: Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019. Springer; 2021. p. 695-704.

[26] Han W, Zhang Z, Zhang Y, Yu J, Chiu CC, Qin J, et al. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:200503191. 2020.

[27] Torres-Boza D, Oveneke MC, Wang F, Jiang D, Verhelst W, Sahli H. Hierarchical sparse coding framework for speech emotion recognition. Speech Communication. 2018 May;99:80-9. Available from: https://www.sciencedirect.com/science/article/pii/S0167639317303412.

[28] Garcia-Ordas MT, Alaiz-Moreton H, Benitez-Andrades JA, Garcia-Rodriguez I, Garcia-Olalla O, Benavides C. Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network. Biomedical Signal Processing and Control. 2021 Aug;69:102946. Available from: https://www.sciencedirect.com/science/article/pii/S1746809421005437.

[29] Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and Recurrent Neural Networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA); 2016. p. 1-4.

[30] Wagner S. Intralingual speech-to-text-conversion in real-time: Challenges and Opportunities; 2005.

[31] Flores AC, Icoy RI, Peña CF, Gorro KD. An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set. In: 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST); 2018. p. 1-4.

[32] Hermansky H, Sharma S. Temporal patterns (TRAPs) in ASR of noisy speech. In: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). vol. 1;. p. 289-92 vol.1. ISSN: 1520-6149.

[33] Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P. Front-End Factor Analysis for Speaker Verification;19(4):788-98. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[34] Anguera X, Bonastre JF. Fast speaker diarization based on binary keys. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP);. p. 4428-31. ISSN: 2379-190X.

[35] Patino J, Delgado H, Yin R, Bredin H, Barras C, Evans NW. ODESSA at Albayzin Speaker Diarization Challenge 2018. In: IberSPEECH;. p. 211-5.

[36] Etienne C. Apprentissage profond appliqué à la reconnaissance des émotions dans la voix [phdthesis];. Available from: https://tel.archives-ouvertes.fr/tel-02479126.

[37] Stevens SS, Volkmann J, Newman EB. A Scale for the Measurement of the Psychological Magnitude Pitch;8(3):185. Publisher: Acoustical Society of AmericaASA. Available from: https://asa.scitation.org/doi/abs/10.1121/1.1915893.

[38] Furui S. Cepstral analysis technique for automatic speaker verification;29(2):254-72. Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

[39] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016. p. 5200-4. ISSN: 2379-190X.

[40] Caihua C. Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM. In: 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS); 2019. p. 173-6.

[41] Saad F, Mahmud H, Shaheen M, Hasan MK, Farastu P. Is Speech Emotion Recognition Language-Independent? Analysis of English and Bangla Languages using Language-Independent Vocal Features; 2021.

[42] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE;. p. 770-8. Available from: http://ieeexplore.ieee.org/document/7780459/.

[43] Bae HS, Lee HJ, Lee SG. Voice recognition based on adaptive MFCC and deep learning. In: 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA); 2016. p. 1542-6. ISSN: 2158-2297.

[44] Mittal S, Agarwal S, Nigam MJ. Real Time Multiple Face Recognition: A Deep Learning Approach. In: Proceedings of the 2018 International Conference on Digital Medicine and Image Processing. DMIP '18. New York, NY, USA: Association for Computing Machinery; 2018. p. 70-6. Available from: https://doi.org/10.1145/3299852.3299853.

[45] Huang KY, Wu CH, Hong QB, Su MH, Chen YH. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. p. 5866-70. ISSN: 2379-190X.

[46] Litovitz T. The TESS Database. Drug Safety. 1998 Jan;18(1):9-19. Available from: https://doi.org/10.2165/00002018-199818010-00002.

[47] Toliupa S, Tereikovskyi I, Tereikovska L, Mussiraliyeva S, Bagitova K. Deep Neural Network Model for Recognition of Speaker's Emotion. In: 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T); 2020. p. 172-6.

[48] Gupta M, Patel T, Mankad SH, Vyas T. Detecting emotions from human speech: role of gender information. In: 2022 IEEE Region 10 Symposium (TENSYMP); 2022. p. 1-6. ISSN: 2642-6102.

[49] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A database of German emotional speech. vol. 5; 2005. p. 1517-20.

[50] Wang K, An N, Li BN, Zhang Y, Li L. Speech Emotion Recognition Using Fourier Parameters. IEEE Transactions on Affective Computing. 2015 Jan;6(1):69-75. Conference Name: IEEE Transactions on Affective Computing.

[51] Pham MH, Noori FM, Torresen J. Emotion Recognition using Speech Data with Convolutional Neural Network. In: 2021 IEEE 2nd International Conference on Signal, Control and Communication (SCC); 2021. p. 182-7.

[52] Livingstone SR, Russo FA. 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. 2018 May;13(5):e0196391. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391.

[53] Sowmya G, Naresh K, Sri JD, Sai KP, Indira DNVSLS. Speech2Emotion: Intensifying Emotion Detection Using MLP through RAVDESS Dataset. In: 2022 International Conference on Electronics and Renewable Systems (ICEARS); 2022. p. 1-3.

[54] Anusha R, Subhashini P, Jyothi D, Harshitha P, Sushma J, Mukesh N. Speech Emotion Recognition using Machine Learning. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI); 2021. p. 1608-12.

[55] 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); 2015. p. 5206-10. ISSN: 2379-190X.

[56] Hong QB, Wu CH, Su MH, Wang HM. Sequential Speaker Embedding and Transfer Learning for Text-Independent Speaker Identification. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2019. p. 827-32. ISSN: 2640-0103.

[57] Laptev A, Korostik R, Svischev A, Andrusenko A, Medennikov I, Rybin S. You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation. In: 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2020. p. 439-44.

[58] Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. 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). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 2236-46. Available from: https://aclanthology.org/P18-1208.

[59] Hu G, Lin TE, Zhao Y, Lu G, Wu Y, Li Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. arXiv; 2022. ArXiv:2211.11256 [cs] version: 1. Available from: http://arxiv.org/abs/2211.11256.

[60] Chen M, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency LP. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow UK: ACM; 2017. p. 163-71. Available from: https://dl.acm.org/doi/10.1145/3136755.3136801.

[61] Kumar A, Vepa J. Gated Mechanism for Attention Based Multimodal Sentiment Analysis. arXiv; 2020. ArXiv:2003.01043 [cs, stat] version: 1. Available from: http://arxiv.org/abs/2003.01043.

[62] Ghosal D, Akhtar MS, Chauhan D, Poria S, Ekbal A, Bhattacharyya P. Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 3454-66. Available from: https://aclanthology.org/D18-1382.

[63] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation. 2008 Dec;42(4):335-59. Available from: https://doi.org/10.1007/s10579-008-9076-6.

[64] Li Z, Tang F, Zhao M, Zhu Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. arXiv; 2022. ArXiv:2203.13504 [cs, eess] version: 1. Available from: http://arxiv.org/abs/2203.13504.

[65] Joshi A, Bhat A, Jain A, Singh AV, Modi A. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. arXiv; 2022. ArXiv:2205.02455 [cs] version: 1. Available from: http://arxiv.org/abs/2205.02455.