# Machine Learning applications in Speech processing: a review

Borlli Michel Jonas SOME, Go Issa Traore

# Speech Processing : a literature review

**Borlli Michel Jonas SOME**[1] **and Go Issa TRAORE**[1]

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

*E-mail : goissatraore@yahoo.fr

**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 extractors; speech database; MFCC.

## I 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 [46]. 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 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.

## II 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.) [13, 22, 24, 27] 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 [14] 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.

## III 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].

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

### 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 table 1.

| Keywords | Objectif |
|---|---|
| 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<br>Machine Learning and speech classification | |
| Speech database | know the existing speech data sources and their characteristics |

Table 1: Keywords used

### 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.

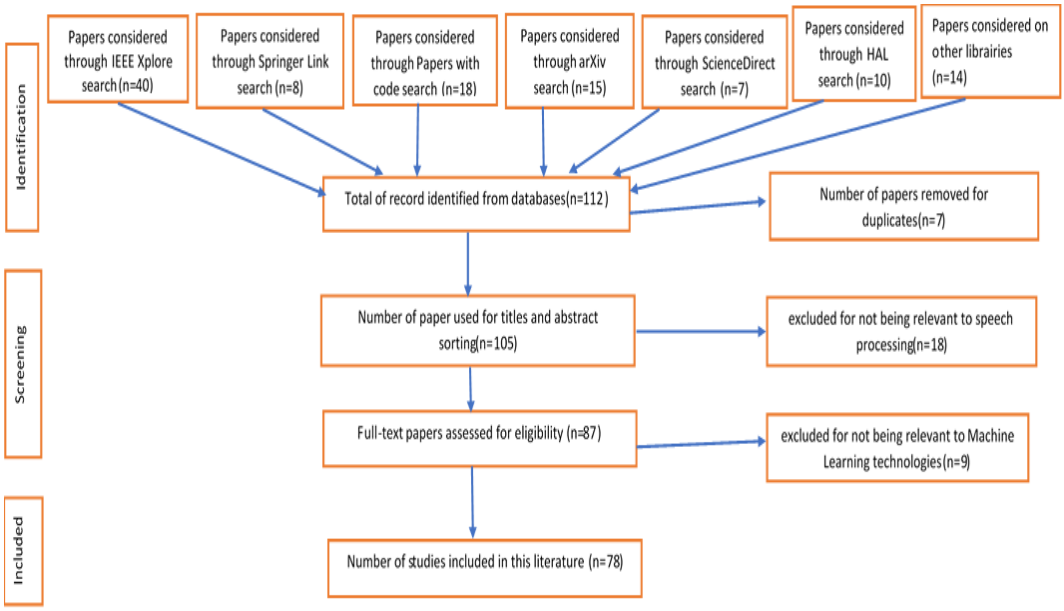


Figure 1: Studies selection process workflow diagram

### 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.

## IV DEFINITION AND EXPLANATION OF SOME CONCEPTS

### 4.1 Speech

Speech can be difine as the articulated human language used to convey thought. It is distinct from various oral communications, such as crying,alarms, warnings, and wail.
Speech is a transmission with multiple layers of information that varies over time. Both the temporal (seconds) and frequency (Hz) components of speech information can be observed.

### 4.2 Speech analysis

Speech analysis: Speech analysis is often referred to as interaction analysis. This technology uses artificial intelligence (AI) to extract meaning from recordings of audio calls. Its main function is to analyse data gathered from spoken words to obtain useful information about customers and patients. It also transcribes the audio calls so that they can be consulted.

### 4.3 Multimodal analysis

The goal of multimodal analysis is to relate various linguistic data that together enable the development and interpretation of a particular message. Therefore, we can distinguish between the oral modality, which includes prosody and voice quality, the visual modality, which includes gestures and facial expressions, and the verbal modality, which includes syntax, lexical choice, phonèmes, and discursive organization. The linguists are focusing on multimodal analysis since the early 2000s. This kind of analysis is effectively used in a wide range of fields, including psychiatry, the development of animated agents and therapeutic remédiation[4].

### 4.4 Speaker recognition

The automatic recognition of a speaker is the use of a machine to identify a speaker based on spoken words. These systems can be used to either identify a specific person or to confirm a person's verified identity.

- Identification of the speaker (language): Establish whether the unknown speaker (or language) is identifiable as one of the known speakers (or languages). In such cases, it is frequently assumed that the unknown voice must originate from a group of known speakers. A one-to-many mapping is done in these kinds of projects;
- Speaker (language) verification/Authentication/Detection/confirmation: Determining whether an unknown speaker (language) corresponds to a specific speaker (language). In other words, does a given voice (or language) correspond to the voice of a person who is already known (or a language that is already known)? In this type of work, one-to-one mapping is used.

### 4.5 Emotions and automatique speech emotions recognition

- An emotion is a response to a trigger, which may come from reality or from our thoughts (worry, memory, etc.). The classification of emotions on which all specialists agree is based on the distinction between primary emotions (or basic or fundamental emotions) and secondary emotions (or complex emotions) [15]. Primary emotions are innate, universal, such as joy, sadness, anger, fear, disgust and surprise. Secondary emotions are complex, acquired emotions that result from a mixture of several primary or related emotions. For example, love, guilt and pessimism are secondary emotions.
- Automatique speech emotions recognition is the process of identifying the emotions in a speech regardless of its semantic content using computer science.

### 4.6 Spectrogram

It is a representation of speech energy distribution according to time and frequency. There are three dimensions to the acoustic signal. These dimensions can be represented in a reference frame with three axes: the x-axis representing time, the y-axis representing frequency and the luminosity of the point representing power. To obtain a spectrogram, the signal must be split into windows and the spectrum of each of these windows calculated [42]. Using a Fast Fourier Transform, the signal is converted into a spectrogram. The spectrogram is widely used to study speech signals analysis. A spectrogram is presented by the figure 2 For speech signal analysis, the spectrogram is used to:

- **Time-frequency visualisation:** The spectrogram shows the distribution of speech energy over time and frequencies. It offers a more comprehensible and complete view than simple time or frequency representations;
- **Component identification:** The various components of a signal, such as the formants of the human voice, harmonics or transient sound events, can be identified by examining the spectrogram;
- **Variations analysis:** The spectrogram can be used for frequency identification and time variations identification in a signal. For example, it can show frequency changes, modulations or interruptions;
- **Speech studies:** Spectrograms are widely used in phonetics and linguistics for speech features studies such as vowels, consonants and transitions between sounds;
- **Anomaly detection:** a Spectrograms can be used to detect anomalies or unexpected events in a sound signal, such as sudden interruptions or variations, or extraneous noise.

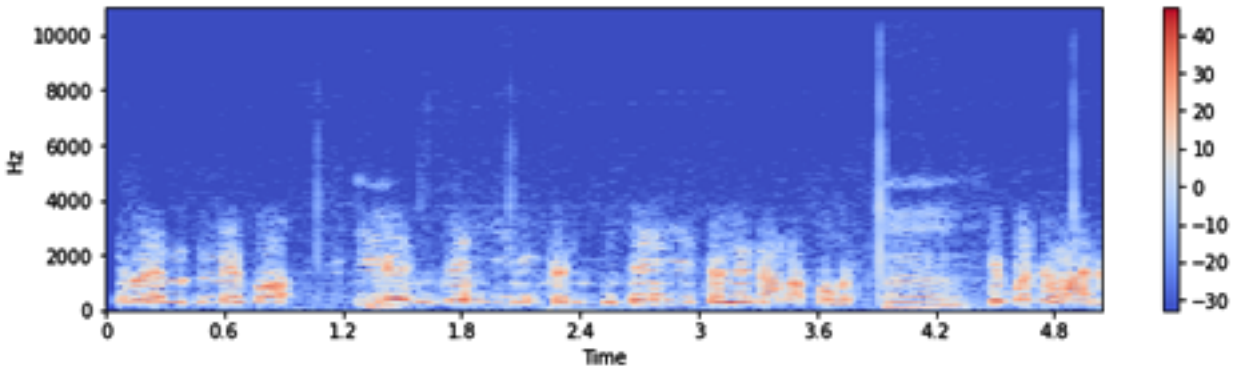


Figure 2: Spectrogram representation

## V RESULTS

The papers considered for this study include scientific articles, doctoral theses and a book. The results we obtained are divided into five categories: the approaches and techniques for speech feature extraction, 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.

### 5.1 Approaches and techniques for speech feature extraction

There are three approaches to extract speech features: The temporal approach, the frequential approach and the cepstral approach.

#### *5.1.1 Temporal approach*

This approach studies the speech signal in order to observe the temporal aspect of the signal. It is also known as an audiogram. A number of characteristics can be deduced from this temporal shape, which can be used for speech processing. Various techniques can be used to analyze the temporal aspect of the speech signal in order to deduce its parameters. Amount these techniques, we have: Linear Predictive Coding (LPC), the zero crossing rate (PPZ), fundamental frequency (F0).

- **Linear Predictive Coding (LPC):** One technique for representing and coding speech is called linear predictive coding (LPC). Its principal is that speech can be represented by a linear process. Speech, however, is not a completely linear process. An error is introduced by the moving average, which is made up of the weighted sum of the signal throughout the time steps. This error is fixed by adding the term e(n). Thus, LPC entails figuring out which coefficients minimize the error e (n). This method is not perfect, since the prediction error may be large and cannot be corrected by this method.
  This mistake is minimized by the Residual Excited Linear Prediction (RELP) technique. The idea is to compare the signal that is obtained via linear prediction with the original signal.
- **Zero Crossing Rate:** ZCR is an interesting descriptor. It is obviously correlated to the fundamental frequency, but also provides information about the noise present in the signal. It formula is presented in 1.

$$\sum_{2 \leq t \leq N} \frac{|sign(i_t) - sign(i_{t-1})|}{2(N-1)} \tag{1}$$

  This method is simple, and very effective when combined with other descriptors. The zero crossing rate can be calculated directly on the signal, but also provides additional information if it is calculated on the modified signal.
- **Fundamental frequency(F0):** the fundamental frequency F0 determines the pitch of the sound. The pitch of a sound corresponds to the frequency at which it vibrates. The human ear can hear frequencies emitted between approximately 16 and 16,000 Hz. In speech, if the frequency is high, it is high-pitched and if the frequency is low, it is considered low-pitched. The pitch of a pure tone corresponds to its frequency of vibration, measured in hertz (number of periodic vibrations per second). The faster the vibration, the pitch is high; the lower the vibration, the pitch is lowe. Only periodic sounds have a fundamental frequency.

### *5.1.2 Frequential or spectral approach*

The spectral representation is the second approach to characterise and represent the speech. These methods are based on a frequency decomposition of the signal without any knowledge a priori of its fine structure. This involves transforming the original signal from a temporal representation to a frequency representation using the Fourier transform given in formula 9.

### *5.1.3 Cepstral approach*

A posteriori deconvolution of the signal is required in order to isolate the two pieces of information contained in the speech signal, namely the fundamental frequency and the transformation, presumed to be linear, performed by the vocal tract. The cepstrum can be used to carry out this deconvolution. This corresponding time-frequency representation is now referred to be Cepstral rather than frequential. The cepstrum method is based on the Fourier transform, but it allows the original frequency of the fundamental to be separated from the transformation that the conduit has carried out, thanks to an effective technique. A possible extension of cepstral coefficients is to pass them through a non-linear frequency space close to human hearing, so as to obtain coefficients distributed according to a Mel scale. This procedure produces Mel Frequency Cepstral Coefficients (MFCC). The calculation of the MFCC is detailed in section (5.3.3).

## 5.2 Performance evaluation of a speech processing model

A crucial part of speech processing is determining a model's reliability. A speech recognition model's quality is assessed using evaluation measures. Evaluation metrics allow us to assess a model's performance and exert control over it to customize it to a company's requirements. Different metrics can be used to assess a model[69]. A metric in speech processing that best permits assessing a model based on the data's structure (balanced or unbalanced data).
We present some common classification metrics used to evaluate models. In the detection of a theme "th" in speech, the nomenclature of predictions associated with an algorithm is represented in the table 2.

| | **Speech contening th** | **Speech doesn't contening th** |
|---|---|---|
| **Speech detected as belonging to the theme th** | True Positive (TP) | False Positive (FP) |
| **Speech detected as not belonging to the theme th** | False Negative (FN) | True Negative (TN) |

Table 2: Nomenclature of predictions

### *5.2.1 Accuracy*

Accuracy is the simplest metric for evaluating a model. It is the ratio between the number of correct predictions and the total number of predictions made for a set of data. It is shown by formula 2.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

The accuracy measure ranges from 0 to 1. The more accurate the model, the closer it is to 1. The accuracy helps when the target class is well-balanced, but it is not a good idea to use it with unbalanced classes since the model will always tend to predict the dominant classes. This method is referred to be blind since it does not distinguish between error types and classes. This is the reason it is insufficient for unbalanced data sets.

#### 5.2.2 *Logarithmic loss or log loss :*

Loss is applied when the classifier's output is a numerical probability rather than a class étiquette. The loss measures the sum of the differences between the probability that an item will belong to a theme and the annotation of that word.
For a binary classification, the formula is 3.

$$Loss = -(ylog(p) + (1-y)log(1-p)) \tag{3}$$

For a multiclasses classification, the formula is 4.

$$Loss = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=1}^{M} Y_{ij} * log(P_{ij}) \tag{4}$$

Where:
**N** : numbers of observation,
**M** : numbers of possible speech class labels (theme1, theme2, theme3 etc.),
**log**: natural logarithm,
**y**: binary indicator,
**p**: model's predicted probability.

#### 5.2.3 *Precision*

Precision for theme "th" is the number of speech correctly associated with theme "th" relative to all the pieces associated with that theme. It is important for the asymmetric and unbalanced dataset. The precision is better when the false positives are few. The formula 5 calculate the precision.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

#### 5.2.4 *Recall*

The recall for the theme th corresponds to the number of speeches correctly associated with the theme th in relation to the total number of speeches in this theme. It is calculate by formula 6.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

It is comprised between 0 and 1 and a value close to 1 indicates that the method has correctly identified the majority of the speeches of the theme th that were present in the database.

#### 5.2.5 *F-score*

The F-score is the harmonic mean of precision and recall. The classifier obtains a high F-score if precision and recall are high.
Between two models, the best performer is the one with the f-score value closest to 1. It is calculate by the formula 7.

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

### 5.3 The existing work on speech analysis

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

#### *5.3.1 Speaker recognition*

Speaker recognition is a task that identifies the speaker from multiple speech using machine [68]. These systems have two uses[62]:speaker verification and speaker diarisation. There is a lot of studies in this area. Mohammed Lataifeh et al. [49] 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. [57] 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.

#### *5.3.2 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. [41] 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. [25] 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. [35] 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).

#### *5.3.3 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. [36] 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. [20] 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. [9] 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.

#### *5.3.4 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)[12]. Many recent studies have been conducted on this topic[51, 70]. Aman Shenoy et al. [50] 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. [44] 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.

### 5.4 Speech features extraction methods

According to Jiang [56], 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.

#### *5.4.1 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[58], wav2vec 2.0[43], HuBERT[55], spectrogram[61], Neural Networks[45]; etc. The MFCC (Mel-frequency cepstral coefficients) is the most popular representation of an speech signal [34]. MFCC components are the most representative feature of audio description [54]. It allows to performs better than experts in some case. It is the case in the study of Mohammed Lataifeh et al.[49]. we will present this method in more detail in the section 4.2.3. Other feature extraction methods exist: Abdul Malik Badshah et al. [25] 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. [21] 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) [54].The wav2vec 2.0[43] 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.

### *5.4.2 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 [7].

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 [29].

### *5.4.3 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 [3, 10, 11, 33], 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 [38]. 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 8 for Hamming window process.

$$w = 0.5 + 0.46 * \cos\left(\frac{2\pi * n}{N-1}\right) \tag{8}$$

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 9.

$$F(n) = \sum_{k=0}^{N-1} U(n) * \exp\left(-jn\frac{2\pi}{N}k\right) \tag{9}$$

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 [6]. This filter bank simulates the perception of frequencies by the human ear.The following formula 10 is used to obtain Mel-scale

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

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 11.

$$C[k] = \log 10 \left(mel * spectrogram[k]\right) \tag{11}$$

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 [1]. 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 12.

$$S[j] = \sum_{n=0}^{N-1} s[n] * \cos \frac{\pi}{N} \left(n + \frac{1}{2}\right) * j \tag{12}$$

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 3.

### 5.5 Machine Learning algorithms used

Most of the studies on speech analysis use machine learning algorithms. According to Mohammed Lataifeh et al [49], 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 [21, 23, 25]. There are also many studies that use SVM [37, 41].

In the paper of Wootaek Lim et al [21], they proposed a method for analyzing sequential speech data based on the concatenation between CNN and RNN. By applying their architecture on a

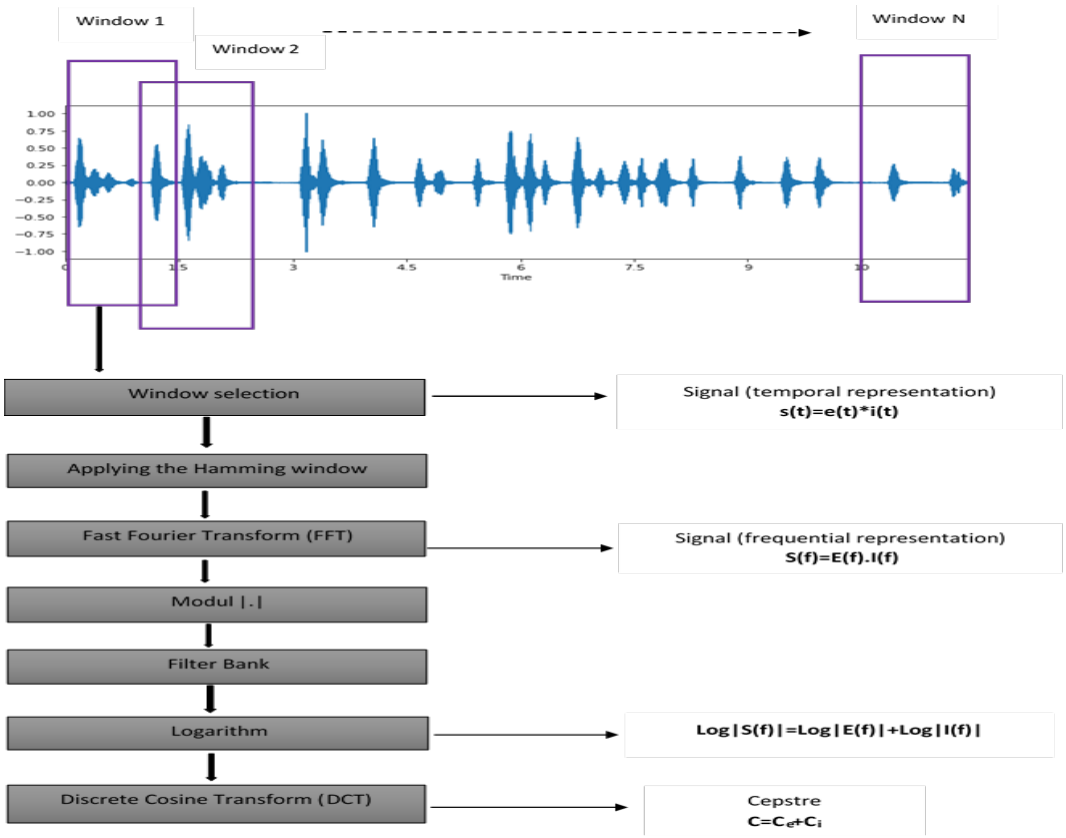


Figure 3: MFCC extraction steps

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 [60] 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 [19], voice recognition [18] , face recognition [32], speech emotion recognition [40]. 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.

## VI 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 3.

| Database name | characteristics | use |
|---|---|---|
| **Some simple speech databases** | | |
| TESS(Toronto Emotional Speech Set)[2] | 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 [52, 63]. |
| EMODB (Berlin Database of Emotional Speech)[5] | 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 [17, 59]. |
| RAVDESS( Ryerson Audio-Visual Database of Emotional Speech and Song)[31] | 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 [67], for speech emotion recognition [53]. |
| LibriSpeech[16] | 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 [39], speech recognition [48] |
| **Some multimodal speech databases** | | |

| | | |
|---|---|---|
| CMU-MOSEI(CMU Multimodal Opinion Sentiment and Emotion Intensity)[28] | 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 [50, 64] |
| CMU-MOSI (Multimodal Corpus of Sentiment Intensity)[26] | 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. [47][30] |
| Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)[8] | 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 [66], and also in multimodal emotion analysis tasks [65]. |

Table 3: Presentation of some data sources used for speech analysis

## VII DISCUSSION

### 7.1 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.

### 7.2 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.

### 7.3 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.

### 7.4 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.

## VIII 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] S. Furui. "Cepstral analysis technique for automatic speaker verification". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.2 (Apr. 1981). Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing, pages 254–272. ISSN: 0096-3518.

[2] T. Litovitz. "The TESS Database". en. In: *Drug Safety* 18.1 (Jan. 1998), pages 9–19. ISSN: 1179-1942.

[3] H. Hermansky and S. Sharma. "Temporal patterns (TRAPs) in ASR of noisy speech". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). Volume 1. ISSN: 1520-6149. Mar. 1999, 289–292 vol.1.

[4] I. Poggi and C. Pelachaud. "Emotional Meaning and Expression in Animated Faces". In: *International Workshop on Affective Interactions*. 1999.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. "A database of German emotional speech". In: volume 5. Sept. 2005, pages 1517–1520.

[6] S. S. Stevens, J. Volkmann, and E. B. Newman. "A Scale for the Measurement of the Psychological Magnitude Pitch". In: *The Journal of the Acoustical Society of America* 8.3 (June 15, 2005). Publisher: Acoustical Society of AmericaASA, page 185. ISSN: 0001-4966.

[7] S. Wagner. *Intralingual speech-to-text-conversion in real-time: Challenges and Opportunities*. Aug. 2005.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. "IEMOCAP: interactive emotional dyadic motion capture database". en. In: *Language Resources and Evaluation* 42.4 (Dec. 2008), pages 335–359. ISSN: 1574-0218.

[9] Gu Mingliang, Xia Yuguo, and Yang Yiming. "Semi-supervised learning based Chinese dialect identification". In: *2008 9th International Conference on Signal Processing*. 2008 9th International Conference on Signal Processing (ICSP 2008). Beijing, China: IEEE, Oct. 2008, pages 1608–1611. ISBN: 978-1-4244-2178-7.

[10] X. Anguera and J.-F. Bonastre. "Fast speaker diarization based on binary keys". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ISSN: 2379-190X. May 2011, pages 4428–4431.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. "Front-End Factor Analysis for Speaker Verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (May 2011). Conference Name: IEEE Transactions on Audio, Speech, and Language Processing, pages 788–798. ISSN: 1558-7924.

[12] G. Ferre. "Analyse multimodale de la parole". fr. In: *Rééducation orthophonique* 246 (2011), page 73.

[13] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera. "Opinion mining and sentiment analysis on a Twitter data stream". In: *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. 2012 International Conference on Advances in ICT for Emerging Regions (ICTer). Colombo, Western, Sri Lanka: IEEE, Dec. 2012, pages 182–188. ISBN: 978-1-4673-5530-8 978-1-4673-5529-2 978-1-4673-5528-5.

[14] A. Mountassir, H. Benbrahim, and I. Berrada. "Sentiment classification on arabic corpora". In: *Document numerique* 16.1 (May 15, 2013), pages 73–96. ISSN: 1279-5127.

[15] Y.-A. Thalmann. *Petit livre de-Décodeur des émotions*. First, 2013.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. Apr. 2015, pages 5206–5210.

[17] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. "Speech Emotion Recognition Using Fourier Parameters". In: *IEEE Transactions on Affective Computing* 6.1 (Jan. 2015). Conference Name: IEEE Transactions on Affective Computing, pages 69–75. ISSN: 1949-3045.

[18] H.-S. Bae, H.-J. Lee, and S.-G. Lee. "Voice recognition based on adaptive MFCC and deep learning". In: *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*. ISSN: 2158-2297. June 2016, pages 1542–1546.

[19] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, June 2016, pages 770–778. ISBN: 978-1-4673-8851-1.

[20] T. Ismail, G. Deka, S. Dutta, and L. Singh. "Kamrupi Dialect Identification Using GMM". en. In: *International Conference on Signal Processing (ICSP 2016)*. Vidisha, India: Institution of Engineering and Technology, 2016, 3 (4 .)–3 (4 .) ISBN: 978-1-78561-783-6.

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

[22] V. N. Patodkar and S. I.R. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: *IJARCCE* 5.12 (Dec. 30, 2016), pages 320–322. ISSN: 22781021.

[23] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. "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)*. ISSN: 2379-190X. Mar. 2016, pages 5200–5204.

[24] N. S. D. Abdullah and I. A. Zolkepli. "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*. BDIOT2017: International Conference on Big Data and Internet of Thing. London United Kingdom: ACM, Dec. 20, 2017, pages 67–74. ISBN: 978-1-4503-5430-1.

[25] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network". In: *2017 International Conference on Platform Technology and Service (PlatCon)*. Feb. 2017, pages 1–5.

[26] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency. "Multimodal sentiment analysis with word-level fusion and reinforcement learning". en. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow UK: ACM, Nov. 2017, pages 163–171. ISBN: 978-1-4503-5543-8.

[27] T. Vepsäläinen, H. Li, and R. Suomi. "Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections". In: *Government Information Quarterly* 34.3 (Sept. 1, 2017), pages 524–532. ISSN: 0740-624X.

[28] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. "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, July 2018, pages 2236–2246.

[29] A. C. Flores, R. I. Icoy, C. F. Peña, and K. D. Gorro. "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)*. July 2018, pages 1–4.

[30] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya. "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, Oct. 2018, pages 3454–3466.

[31] S. R. Livingstone and F. A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". en. In: *PLOS ONE* 13.5 (May 2018). Publisher: Public Library of Science, e0196391. ISSN: 1932-6203.

[32] S. Mittal, S. Agarwal, and M. J. Nigam. "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, Nov. 2018, pages 70–76. ISBN: 978-1-4503-6578-9.

[33] J. Patino, H. Delgado, R. Yin, H. Bredin, C. Barras, and N. W. Evans. "ODESSA at Albayzin Speaker Diarization Challenge 2018." In: *IberSPEECH*. 2018, pages 211–215.

[34] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli. "Hierarchical sparse coding framework for speech emotion recognition". en. In: *Speech Communication* 99 (May 2018), pages 80–89. ISSN: 0167-6393.

[35] A. B. Abdul Qayyum, A. Arefeen, and C. Shahnaz. "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition". In: *2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*. 2019, pages 122–125.

[36] S. Al-Azani and E.-S. M. El-Alfy. "Audio-Textual Arabic Dialect Identification for Opinion Mining Videos". In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. Dec. 2019, pages 2470–2475.

[37] C. Caihua. "Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM". In: *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*. July 2019, pages 173–176.

[38] C. Etienne. "Apprentissage profond appliqué à la reconnaissance des émotions dans la voix". PhD thesis. Université Paris Saclay (COmUE), Dec. 18, 2019.

[39] Q.-B. Hong, C.-H. Wu, M.-H. Su, and H.-M. Wang. "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)*. ISSN: 2640-0103. Nov. 2019, pages 827–832.

[40] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen. "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)*. ISSN: 2379-190X. May 2019, pages 5866–5870.

[41] Muljono, M. R. Prasetya, A. Harjoko, and C. Supriyanto. "Speech Emotion Recognition of Indonesian Movie Audio Tracks based on MFCC and SVM". In: *2019 International Conference on contemporary Computing and Informatics (IC3I)*. Dec. 2019, pages 22–25.

[42] J. C. Taboada-Echave, M. Medina-Melendrez, and L. N. Gaxiola-Sánchez. "Spectrum Sample Calculation of Discrete, Aperiodic and Finite Signals Using the Discrete Time Fourier Transform (DTFT)". In: *Supercomputing*. Edited by M. Torres and J. Klapp. Cham: Springer International Publishing, 2019, pages 18–26. ISBN: 978-3-030-38043-4.

[43] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems*. Edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Volume 33. Curran Associates, Inc., 2020, pages 12449–12460.

[44] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont. "A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis". In: *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Challenge-HML 2020. Seattle, USA: Association for Computational Linguistics, July 2020, pages 1–7.

[45] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu. "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context". In: *arXiv preprint arXiv:2005.03191* (2020).

[46] D. Issa, M. Fatih Demirci, and A. Yazici. "Speech emotion recognition with deep convolutional neural networks". en. In: *Biomedical Signal Processing and Control* 59 (May 2020), page 101894. ISSN: 1746-8094.

[47] A. Kumar and J. Vepa. *Gated Mechanism for Attention Based Multimodal Sentiment Analysis*. arXiv:2003.01043 [cs, stat] version: 1. Feb. 2020.

[48] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin. "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)*. Oct. 2020, pages 439–444.

[49] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif. "Arabic audio clips: Identification and discrimination of authentic Cantillations from imitations". en. In: *Neurocomputing* 418 (Dec. 2020), pages 162–177. ISSN: 09252312.

[50] A. Shenoy and A. Sardana. "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)*. arXiv:2002.08267 [cs, eess]. 2020, pages 19–28.

[51] U. Sulubacak, O. Caglayan, S.-A. Grönroos, A. Rouhe, D. Elliott, L. Specia, and J. Tiedemann. "Multimodal machine translation through visuals and speech". In: *Machine Translation* 34 (2020), pages 97–147.

[52] S. Toliupa, I. Tereikovskyi, L. Tereikovska, S. Mussiraliyeva, and K. Bagitova. "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)*. Oct. 2020, pages 172–176.

[53] R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma, and N. Mukesh. "Speech Emotion Recognition using Machine Learning". In: *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. June 2021, pages 1608–1612.

[54] M. T. Garcia-Ordas, H. Alaiz-Moreton, J. A. Benitez-Andrades, I. Garcia-Rodriguez, O. Garcia-Olalla, and C. Benavides. "Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network". en. In: *Biomedical Signal Processing and Control* 69 (Aug. 2021), page 102946. ISSN: 1746-8094.

[55] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pages 3451–3460.

[56] H. Jiang, X. Wu, X. Xie, and J. Wu. "Audio Public opinion Analysis Model based on heterogeneous Neural Network". In: *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. Jan. 2021, pages 449–453.

[57] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi. "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities". In: *IEEE Access* 9 (2021). Conference Name: IEEE Access, pages 79236–79263. ISSN: 2169-3536.

[58] A. Mahmood and K. Utku. "Speech recognition based on convolutional neural networks and MFCC algorithm". In: *Advances in Artificial Intelligence Research* 1.1 (2021), pages 6–12.

[59] M. H. Pham, F. M. Noori, and J. Torresen. "Emotion Recognition using Speech Data with Convolutional Neural Network". In: *2021 IEEE 2nd International Conference on Signal, Control and Communication (SCC)*. Dec. 2021, pages 182–187.

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

[61] V. H. Shah and M. Chandra. "Speech recognition using spectrogram-based visual features". In: *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*. Springer. 2021, pages 695–704.

[62] A. Brown, J. Huh, J. S. Chung, A. Nagrani, D. Garcia-Romero, and A. Zisserman. "Voxsrc 2021: The third voxceleb speaker recognition challenge". In: *arXiv preprint arXiv:2201.04583* (2022).

[63] M. Gupta, T. Patel, S. H. Mankad, and T. Vyas. "Detecting emotions from human speech: role of gender information". In: *2022 IEEE Region 10 Symposium (TENSYMP)*. ISSN: 2642-6102. July 2022, pages 1–6.

[64] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li. *UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition*. arXiv:2211.11256 [cs] version: 1. Nov. 2022.

[65] A. Joshi, A. Bhat, A. Jain, A. V. Singh, and A. Modi. *COGMEN: COntextualized GNN based Multimodal Emotion recognitioN*. arXiv:2205.02455 [cs] version: 1. May 2022.

[66] Z. Li, F. Tang, M. Zhao, and Y. Zhu. *EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition*. arXiv:2203.13504 [cs, eess] version: 1. Mar. 2022.

[67] G. Sowmya, K. Naresh, J. D. Sri, K. P. Sai, and D. V. Indira. "Speech2Emotion: Intensifying Emotion Detection Using MLP through RAVDESS Dataset". In: *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. Mar. 2022, pages 1–3.

[68] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, and Z. Gu. "Adversarial attack and defense strategies of speaker recognition systems: A survey". In: *Electronics* 11.14 (2022), page 2183.

[69] G. Varoquaux and O. Colliot. "Evaluating Machine Learning Models and Their Diagnostic Value". In: *Machine Learning for Brain Disorders*. Edited by O. Colliot. New York, NY: Springer US, 2023, pages 601–630. ISBN: 978-1-0716-3195-9.

[70] L. Zhang, L. Ruan, A. Hu, and Q. Jin. "Multimodal Pretraining from Monolingual to Multilingual". In: *Machine Intelligence Research* 20.2 (2023), pages 220–232.