(a) UNIVERSITY OF LORRAINE (b) IDMC (c) LORIA

MSc Natural Language Processing - 2019/2020
UE 705 - Supervised Project

Acoustic scene classification for speaker diarization: a preliminary study

Students :
Tahani FENNIR
Fatima HABIB
Cécile MACAIRE

Supervisors:
Md Sahidullah
Romain Serizel

December 2019

Abstract

Speaker diarization is a task of labelling segments of a long audio recording according to the speaker information. This work investigates the speaker diarization in the presence of different acoustic environments. In this report, we have used DIHARD II dataset to analyze speaker diarization performance for eleven different acoustic environments or scenes. In this part of the report, we introduce the basic concepts of speaker diarization and we report preliminary results in terms of *diarization error rate* (DER), *Jaccard error rate* (JER). Our preliminary results indicate that the speaker diarization performance can be substantially improved for several acoustic classes given that the acoustic condition is known apriori.

Contents

List of Figures

Chapter 1

Introduction

This work is focusing on speaker diarization. In order to fully understand what is it and what are the challenges, we first give you some introductory elements about the speech basics, how the speech is collected and stored, and analyzed. Also, an overview of speech processing applications and different features and machine learning methods. Then, after set a definition and the main applications of speaker diarization, we explain, thanks to the DIHARD II [2] challenge and manipulations on specific data (calculations of the Diarization Error Rate (DER) and the Jaccard Error rate (JER), classification (t-SNE plots)), the problem statement regarding this speech recognition domain. Finally, we conclude our first step of work and we present our future steps to improve the results we got.

1.1 Speech basics

The human has the ability to produce hundred of sounds but not all of them are speech sounds, only the ones that we use in the spoken language are considered as such. Since sound is a wave, we can relate the properties of sound to the properties of a wave. We can therefore represent sound by a waveform, the vertical axes are the **Amplitude** which is the amplitude of sound pressure variations, measured from 0. The basic properties of sound are: **pitch**, **loudness** and **tone**.

- **Pitch**: is a perceptual property of sounds that allows their ordering on a frequency-related scale extending from low to high [3].The higher the pitch the higher the sound frequency.
- **Loudness**: The amplitude of a sound wave determines its loudness or volume. A larger amplitude means a louder sound, and a smaller amplitude means a softer sound.
- **Tone**: Tone, in acoustics, is a sound that can be recognized by its regularity of vibration. A simple tone has only one frequency, although its intensity may vary. A complex tone consists of two or more simple tones, called overtones. The tone of lowest frequency is called the fundamental; the others, overtones [4].

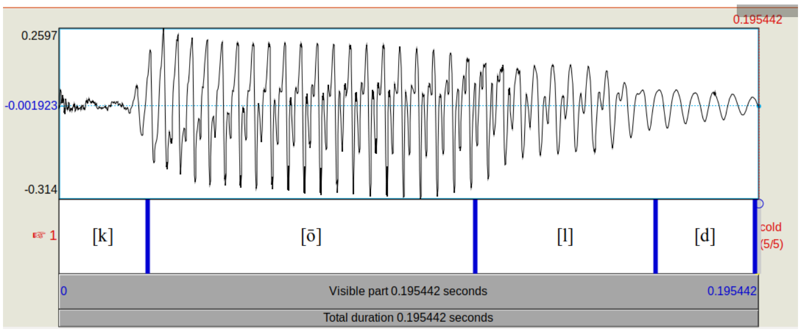
0.195442
0.2597
-0.001923
-0.314
1
[k]
[ō]
[l]
[d]
cold
(5/5)
0
Visible part 0.195442 seconds
0.195442
Total duration 0.195442 seconds

Figure 1.1: The waveform of the English word "cold"

1.2 Speech production & perception

Speech production is the process of producing speech sounds and occurs when air passes through the vocal system. The sounds vary according to the position of the vocal parts during the air flow. We call these parts the **articulators**, and their study is called the **articular phonetics**. On the other hand, **voice perception** refers to how sounds are heard, understood and interpreted.

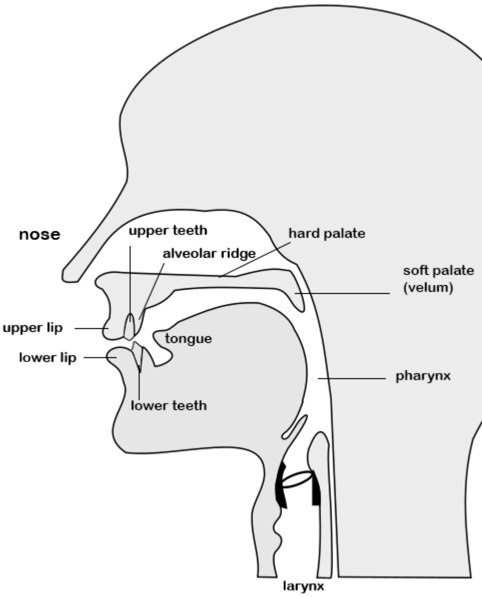

Figure 1.2: The articulators

1.3 How the speech signal is collected & stored (sampling rate, formant, compression, etc.)

- **Sampling**
  The signals that we use in the real world are analog signals like the voice. In order to process these signals, we need to convert them into digital signals so they can be understood by the computer. This converting process is called sampling.

- **Sampling rate** is the number of samples per second. The rate of an analog signal is taken in order to be converted into digital form.

- **Formants**
  A formant is an acoustic energy concentration in the speech wave around a specific frequency. There are several formants in every 1000Hz band, each at a different frequency. In other words, formants occur at intervals of approximately 1000Hz. Every formant corresponds to a vocal tract resonance.

1.4 Analysis of speech (time-domain & spectrogram visualization)

The analysis of speech is conducted by **spectral analysis**. Because the speech signals are time-varying, the analysis has to be a time-frequency analysis.

The analysis can be done thanks to spectrogram using the Fourier transform (correspond to the distribution study of energy along frequency). Indeed, spectrogram visualization shows a three-dimensional image of the evolution of the speech signal with time, frequency and intensity. In a spectrogram, the energy is shown with black color. The energy and the characteristics of speech signal are parameters that are used to identify vowels or consonants in speech. Vowels are represented by a periodic signal and a significant amount of energy while the consonants are represented by a random signal and less energy.

But before using the Fourier transform, the signal is splitted into windows. Small windows correspond to a wide range spectrograms and long windows to narrow band spectrograms. Wideband spectrogram is used to observe the formant structure (dark bands which correspond to the peaks in the spectrum) while narrowband spectrogram informs us of the harmonic structure. Pitch can be determined by finding inverse of the time duration after which the waveform repeats itself. Pitch is the difference in hertz between the harmonics. It is the fundamental frequency of vibration of the vocal folds, which are present at the top of one's trachea.

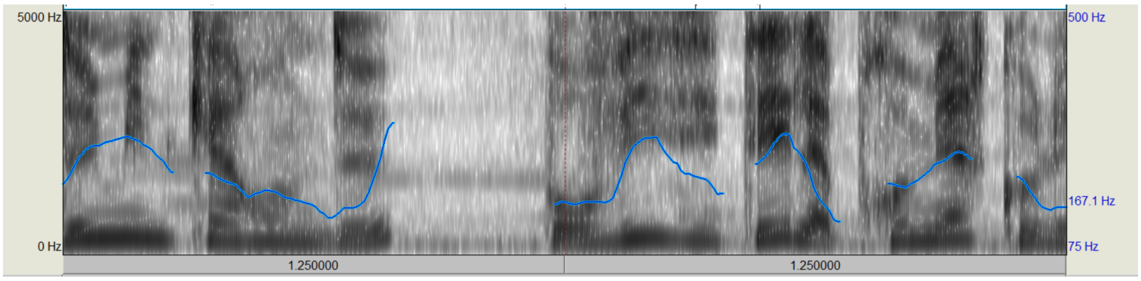

Figure 1.3: The visual representation of speech

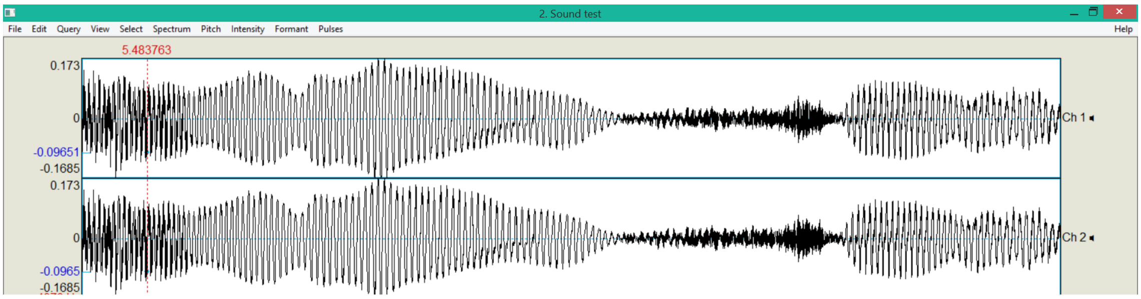

Figure 1.4: Speech signal representation

1.5 Overview of speech processing applications

This section introduces you an overview of speech processing applications. These applications are used in a day-to-day basis.

We can divide these speech processing applications in 4 categories:

- **speech recognition**
- **language recognition**
- **speaker recognition**
- **speaker diarization**

1.5.1 Speech Recognition

The capability of a machine or program to recognize words and phrases in spoken language and convert them to a machine-readable format is known as Speech Recognition.

The predominant use has been in the area of recognition and understanding of speech (voice dictation, natural language voice dialogues, voice dialing for smartphones, . . .). An example of application is Voice-to-Text Apps. Google Assistant is one of them. In more details, the user exploits his voice in order to look up information and tell to Google Assistant what to do (send messages, write an email, add events to calendar, . . .). Another interesting example is real-time translation. The principle is as follows: the user talks to his phone. After setting up two languages, the app will automatically translate the speech into the desired language. But again, this technology is going through several issues. It includes the difficulty of identifying certain words due to variations in pronunciation, the lack of languages support and the inability to override background noises.

1.5.2 Language recognition

If we talk about spoken language recognition, it refers to the ability of determining which is the language used in a speech sample [5]. This specific area has multiple application domains. The first domain concerns language translation which translated the speech to the target language (iTranslate in iOS, Google Translate, ...). Another area is in spoken document retrieval which is the process of indexing and then retrieving relevant items from a huge collection of registered speech audios when a user has a specific natural language query [6].

1.5.3 Speaker recognition

In an audio file, identify who is speaking is called speaker recognition. It is possible by extracting specific information about the speaker in speech waves [7]. What is the use of it ? We can take the example of security voice applications to unlock your smartphone, or control access (banking transactions, voice mail, ...). Also, this technology is promising for criminal and investigations purposes which are based on voice samples records. Identifying who is speaking at a conference, a meeting or during a dialogue is also an important area of application.

Speaker diarization, which will be explain in the next section, is an aspect of speaker recognition.

1.5.4 Speaker diarization

An important task in audio retrieval and processing is speaker diarization or indexing which is the procedure to automatically divide a conversation involving numerous speakers into homogeneous segments and to group together all segments corresponding to the same speaker. The first part of the procedure is known as speaker segmentation, whereas the second part is known as speaker clustering. For this reason, speaker segmentation followed by speaker clustering is recognized as the speaker diarization [8] (the section 1.7 will include more details).

1.5.5 Other application areas

In addition to these four main areas, we can point out:

- **speech verification** which is a subfield of speech recognition system which verify the correctness of the pronounced speech. It provides secure access to information, and builds systems with moderate security.

- **speech enhancement**: aims to improve speech quality (intelligibility, clarity, . . .) which is reducing background noises, eliminate echo, improve voice quality, . . . and make the speech more natural.

1.6 Overview of different features & machine learning methods for speech processing

Before giving an overview of the different features and machine learning methods for speech processing, let's set some elements of definition.

What is machine learning ?
Machine learning is the science of automatic data pattern detection [9]. It is the science of getting computers to learn and act like humans do and improve their learning over time.

How does machine learning work ?
Machine learning uses two techniques:

- **supervised learning** (training a model on data to predict the future) based on classification (into specific classes) and regression tasks.
- **unsupervised learning** (find patterns and structures in input data) based on clustering tasks. It will identify clusters in data based on similar characteristics.

Regarding speech technologies, machine learning paradigms have appeared in this specific field thanks to important development of computer [10].

Especially, in the field of acoustic speech recognition, we can describe the development step by step of different approaches chronologically:

- **Vector Quantization (VQ)** to classify audios. For example, several sounds are recorded in different places. This corresponds to the vector x to be classified. After determining the vectors that correspond to a typical sound in a specific environment, vector quantization will find the vector closest to the sound to be classified by calculating the distance between them [11].
- **Gaussian Mixture Model** contains gaussians, represented by $\kappa \in \{1, ..., K\}, K$ is the number of clusters from the dataset. Each κ is represented by the mean μ which is a dimensional vector and the covariance matrix, σ which define the width of the cluster. This model is efficient if there is one class in one classifier.
- **Support Vector Machine (SVM)** uses a high-dimension space. In more details, the aim is to find a hyper-plane in N dimension spaces, thanks to binary data to, at the end, classify them. The hyper-plane dimension will be determined by the number of features.
- **Artificial Neural Network (ANN)** is based on our human neuronal system. It consists of a set of neurons splitted into layers (input, hidden and output layers). Each neurons are connected to each other through weighted connections. A neuron value is the multiplication of the value of a connected neuron with a weight (set with the stochastic Gradient-descent algorithm). The weight can be computed with a bias. Thanks to an activation function $f(x)$, the bias value is transformed and attached to the neuron in the adjacent layer.
- **Deep Neural Networks (DNN)** consists of multiple machine learning algorithms in the form of multiple models [12]. Deep learning algorithms are used to enhance performance of computers in order to understand in a better way human capabilities. When the ANN uses 2 or 3 layers, Deep Neural Network can use more than 1000 layers.
- **Convolutional Neural Network (CNN)** is a deep learning approach for images and spectrogram images. The algorithm will attribute weight and bias in distinct features inside the image and will output a model in order to differentiate between images.
- **Long Short-Time Memory (LSTM** is a Recurrent Neural Network (RNN). This approach takes into account, not only the current input, but also the previous one, hence the name recurrent.

Speech processing features

Machine learning methods extract features (energy, frequency, source, ...) from speech data to identify who is speaking. Some interesting audio features extraction approaches for speaker recognition are:

- **Mel-Frequency Cepstral Coefficients (MFCC)** which extract features to represent the short-term power spectrum of a spectral envelop.
- **Principle Component Analysis (PCA)** which try to create the best data distribution representation by finding the most relevant combination of features.

Chapter 2

Speaker diarization

2.1 Definition & application

Speaker diarization is the operation of labeling a speech signal with labels corresponding to the identity of speakers [1]. It is the job of deciding "who spoke when?" in an audio or video recording involving an estimated number of speakers and an unspecified number of speakers. It has become a key technology for many tasks, including navigation, retrieval, or higher-level audio data inference.

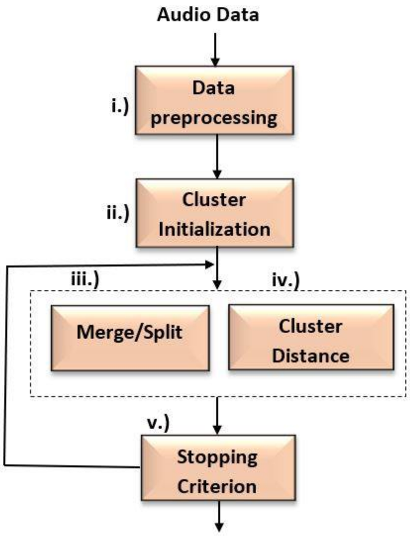

Figure 2.1: General speaker diarization architecture [1]

Fig. 2.1 represent a block diagram of the generic modules that build the most speaker diarization systems. The data preprocessing step (i) tends to be comparatively domain specific, for the data of meeting. Preprocessing typically involves noise reduction (e.g. Wiener filtering), multichannel acoustic beamforming, the parameterization of speech data into acoustic features (such as MFCC, PLP, etc.) and detect speech segments using the speech activity detection algorithm. Cluster initialization (ii) depends on the approach to diarization, i.e. the choice of an initial set of clusters in bottom-up clustering or a single segment in top-down clustering. Next, in Fig. 2.1 iii/iv, a distance between clusters and a split/merging mechanism is used to iteratively merge clusters or to introduce new ones. Optionally, data purification algorithms can be used to make clusters more discriminant. Finally, as illustrated in Fig. 2.1 (v), stopping criteria are used to determine when the optimum number of clusters has been reached.

Some of the applications for speaker diarization are [13]:

- Rich transcription: rich transcription (RT) (Meignier et al., 2006; Gales et al., 2006)[14] many metadata adds in spoken file, for example speaker identity and sentence boundaries.

- Movie Analysis: the analysis of movies involves different assignments. For instance, the detection dialog determines whether or not an audio recording dialog occurs.
- Automatic speech recognition (ASR) systems: segmentation algorithms could be used to divide audio into small segments for ASR processing systems.
- Audio archiving and monitoring: having archived meetings or conferences, they can be easily reached and monitored by interested persons who were unable to join such meetings.
- Audio indexing and retrieval: the speaker diarization system provides the automatic indexing of spoken audio files, enabling the end user to search the audio document by the identity of the speakers or their number.
- Speaker count: this application includes determining the number of speakers taking part in a conversation (most likely without having any previous information on either of the speakers).
- Call routing: a further application for speaker detection and tracking is automatic call routing depending on the identity of the caller.

There are three main application[14] domains for speaker diarization, as shown by Reynolds and Torres-Carrasquillo (2004):

- Broadcast news (BN): radio and TV programs with a variety of content, usually containing commercial breaks and music, on a single channel.
- Meetings: meetings or lectures in which multiple individuals communicate in the same room. Normally, recordings are made with a few microphones.
- Conversational telephone speech (CTS): single-channel recording of telephone conversations between two or more individuals.

Speech and speech indexing, document content structuring, speaker recognition (in the presence of multiple or competing speakers), speech-to-text transcription (i.e. speech-to-text-assigned speakers), are obvious examples of applications for speaker diarization algorithms.

2.2 Challenges & problem statement

A principal restriction of most current speaker diarization frameworks is that just one speaker is assigned to each segment. The existence of overlapped speech, though, is popular in multiparty meetings and, consequently, presents a significant challenge to automatic systems. Specifically, in regions where more than one speaker is active, missed speech errors will be incurred and given the high performance of some state-of the-art systems, this can be a substantial fraction of the overall diarization error. The biggest single challenge is the handling of overlapping speech, which needs to be attributed to multiple speakers.

The fields of application, from broadcast news, to lectures and meetings, vary widely and pose different problems, such as access to multiple microphones, multimedia information, recognize the location and acoustic sign or overlapping speech. The detection and treatment of overlapping speech remains an unresolved problem.

In the period of the spring 2018, the initial DIHARD challenge [15] ran and 20 teams registrations have been attracted of which 13 submitted systems. DIHARD I [15], contains a one channel input condition using wide band speech sample from 11 demanding fields, ranging from clean recordings of read audio books to very noisy high interactive recording of speech. And also offer multichannel input needs participants to perform diarization from farfield microphone. The DIHARD II challenge [2] is the second in a series of speaker diarization challenges and was designed to enhance the robustness of diarization systems to variation in recording equipment, noise levels and conversational environments. Like its previous challenge, it examines the performance of the diarization system under two SAD conditions: diarization from the reference SAD supplied and diarization from scratch.

Showing in further details the problem statement of speaker diarization, we ran the experiment (which uses the Kaldi project, a toolkit dedicated to speech recognition) and computed several indicators. For the next part, we are only going to talk about the DER (Diarization Error Rate) and the JER (Jaccard Error Rate) for each 11 categories which constitute the DIHARD II challenge dataset [2]. The 11 categories are as follow :

- audiobooks,
- broadcast interviews,
- child,
- clinical,
- court,
- map tasks,
- meeting,
- restaurant,
- socio-field (everyday conversations),
- socio-lab,
- webvideos.

DER is the overall percentage of the reference speaker time that is not accurately attributed to the speaker, that accurately attributed is described in terms of the optimal mapping between the reference speakers and the system speakers. More specifically, the DER is equal to

$$DER = \frac{\text{FA} + \text{MISS} + \text{ERROR}}{\text{TOTAL}}$$

and that each element refer to as following:

- TOTAL is the complete reference speaker time, therefore, the count of the periods of all the reference speaker segments.
- FA is the total time of the device speaker not related to the reference speaker.
- MISS is the overall reference time of the speaker not related to the device speaker.
- ERROR is the full reference time of the speaker assigned to the incorrect speaker. The higher the percentage, the more incorrect the diarization is.

The JER corresponds to the Jaccard Error Rate, a metric based on Jaccard Index and specially computed for this challenge. The Jaccard Index is used to calculate the accuracy of a segmentation thanks to the ratio between two segmentation. To compute the Jaccard Error rate, they determine the best mapping score between reference and system speakers and for each, the Jaccard index is calculated. It will finally result of a percentage which equals to 1 - average scores.

$$JER = \frac{\text{FA} + \text{MISS}}{\text{TOTAL}}$$

- TOTAL is the length of the combination of the reference and the system speaker segments; if the reference speaker was not combined with the system speaker, it is the duration of all the reference speaker segments.
- FA is the overall system speaker time not assigned to the reference speaker; if the reference speaker is not matched with the system speaker, it is 0.
- MISS is the overall reference speaker period not assigned to the system speaker; if the reference speaker has not been combined with the system speaker, it is equivalent to TOTAL.

Categories	**Average DER**	**Average JER**	**Number of files**
audiobooks	2.8	2.82	12
broadcast interviews	5.68	41.08	12
child	31.79	61.74	23
clinical	20.83	36.23	23
court	15.19	56.37	12
maptask	6.59	11.85	23
meeting	34.56	61.09	14
restaurant	49.77	79.03	12
socio field	14.84	40.19	12
socio lab	10.94	16.32	16
webvideo	37.85	62.15	32
TOTAL	**23.16**	**55.75**	**191**

Table 2.1: Table of average DER and JER for each category for global threshold (-0.3).

The table 2.1 shows, for each category, the average value of DER and JER, obtained thanks to the DER and JER values of all the files that make up this category. The threshold is set at - 0.3. The file number 13 is not categorized. For (all) the categories you can notice that the best result is in the audiobooks category and the worst one is in the category restaurant.

Categories	**DER**	**JER**	**Threshold**
audiobooks	0.00	0.00	- 2.0
broadcast interviews	7.07	47.47	- 0.8
child	36.10	67.05	- 0.4
clinical	17.80	28.76	- 0.3
court	13.23	47,54	0.0
maptask	8.58	15.20	- 0.4
meeting	33.11	60.53	- 0.4
restaurant	52.00	77.82	- 0.3
socio field	24.48	55.66	- 0.5
socio lab	11.26	18.75	- 0.6
webvideo	39.57	77.74	- 0.5
TOTAL	**22.10**	**45.13**	-

Table 2.2: Table which contains DER and JER for each category.

In the table above, the DER and JER results are displayed for each category that was mentioned previously. We ran the experiment for them separately, and we set their threshold on a scale between -2.0 to 0.5.

The lower the DER, the clearer the sounds will be. You can notice from the table that the DER of the restaurant category or child is high compared to the audio books, which is very logic sense because the audio books are recorded in a very clear environment. Other categories have good DER results such as broadcast interview and maptasks (results bellow 10).

When we compare the results with the previous table (2.1), the results when setting the same threshold for each category are similar but are less relevant. For example, a noticeable difference is observed for the category socio field (DER table 2.1 value is 14.84 and DER table 2.2 value is 24.48).

2.3 t-SNE visualization

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction algorithm used for exploring high-dimensional datasets. It maps multi-dimensional data to two or more dimensions suitable for human observation. With help of the t-SNE algorithms, you may have to plot fewer exploratory data analysis plots next time you work with high dimensional data. To represent the data from the DIHARD II challenge [2], we performed two t-SNE plots. These two visualizations uses two types of data : x-vectors and i-vectors. X-vectors and i-vectors are extracted thanks to a deep neural network and represent, in a compact form, speech utterances. In more details, deep neural network maps sequences of features of each speech to fixed-dimensional embeddings.

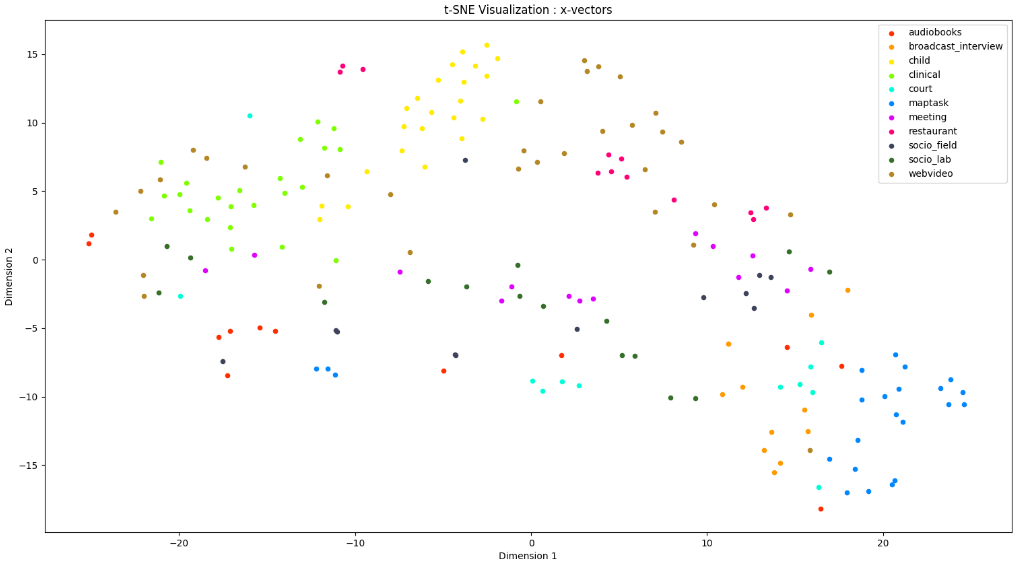

Figure 2.2: t-sne visualization of audio recordings of DIHARD II dataset [2] with x-vectors. We have chosen the development set consisting 192 files from 11 different categories.

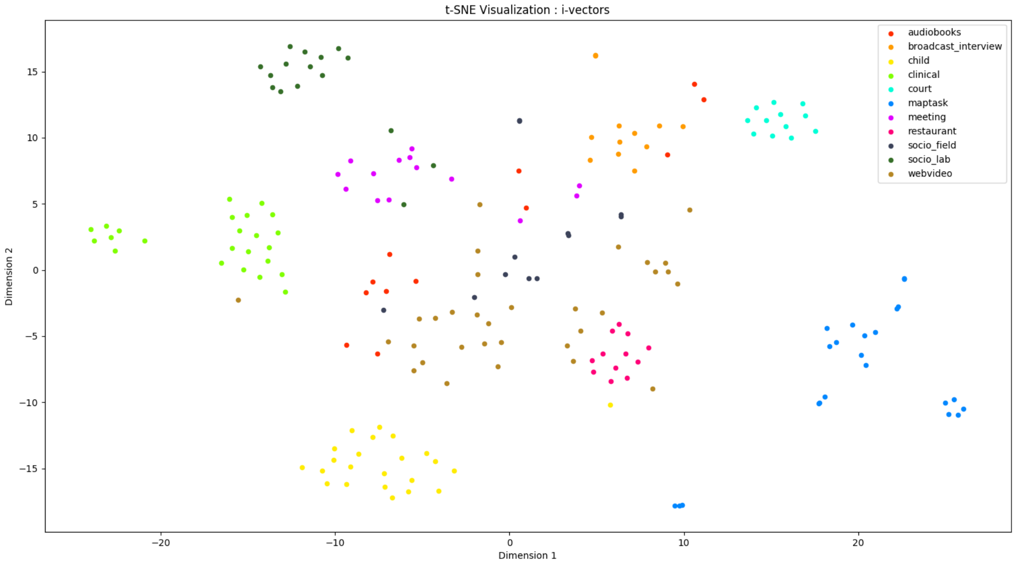

Figure 2.3: t-sne visualization of audio recordings of DIHARD II dataset [2] with i-vectors. We have chosen the development set consisting 192 files from 11 different categories.

In the figures above, the t-SNE displays the data on two dimensions. First of all, on the i-vector t-SNE plot, the files from the categories clinical, child, maptask and court are well clustered. The others have their data very scattered, such as the categories web video or audiobooks and their data are mixed between them. The webvideo data are not well represented in contract with child category. It means that some files initially labeled in a category present dimension similarities to files which belongs to another category. For example, a web video can contain speech of a child, therefore this file is not well categorized.

Chapter 3

Conclusion & Future Work

This report represents the first step of speech basics and analysis and moves on to the specificity of speaker diarization. The aim of the project is to improve the performance of the identification of a speaker in challenging acoustic scenes as restaurant or meetings by defining the acoustic conditions of the record environment towards improving speaker diarization performance.

In experiments with the DIHARD II challenge [2] dataset which contains audio files which are categorized in 11 classes (from very clear environment to very noisy one), we identified several issues. First of all, the diarization performance is different for different categories of audio files when computed with a global threshold. In experiments with category-specific speaker diarization where the thresholds are separately optimized, we have found that the optimized thresholds are not necessarily same over all the categories. This motivates us to further optimize the speaker recognition component.

In t-SNE visualization of the speech embeddings, we have observed that i-vector based embeddings are better than x-vector based embeddings. We also notice that some of the audio categories are nicely clustered and some of those clusters are somewhat closer. This indicates that some of the categories can be grouped together. On the other hand, the audio recordings of highly scattered audio categories (i.e., web videos) can be assigned to a different category indicated by the cluster closest to that file in terms of embedding similarity.

In the next phase of the project:

- We will focus on the improvement of speaker diarization by better classified the data (grouping the acoustic scene into fewer classes when it is possible) and improve the set of some parameters (threshold) for each category in order to optimize the DER and JER values.
- We will explore different classification techniques : *unsupervised* and *supervised* methods.
- We are limited to the dataset that we work on. We would like to record new data in different recording environments to see what results we can get, and how we can improve them.

Bibliography

[1] Xavier Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, February 2012.

[2] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. Second dihard challenge evaluation plan. *Linguistic Data Consortium, Tech. Rep*, 2019.

[3] Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription.* Springer science Business media, 2009.

[4] Encyclopædia Britannica, Tone. `https://www.britannica.com/science/tone-sound`, Apr 2019.

[5] Haizhou Li, Bin Ma, and Kong Aik Lee. Spoken language recognition: from fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159, 2013.

[6] Kenney Ng and Victor W Zue. Subword-based approaches for spoken document retrieval. *Speech Communication*, 32(3):157–186, 2000.

[7] Sadaoki Furui. *Chapter 7 - Speaker Recognition in Smart Environments.* Academic Press, Oxford, 2010.

[8] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-François Bonastre, and Laurent Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language*, 20(2-3):303–330, April 2006.

[9] Shalev-Shwartz Shai and Ben-David Shai. *Understanding machine learning: from theory to algorithms.* Cambridge University Press, 2016.

[10] Milan Sečujski Nikša Jakovljević Jelena Nikolić Dragiša Mišković Nikola Simić Siniša Suzić Vlado Delić, Zoran Perić and Tijana Delić. Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience*, 2019:1–19, June 2019.

[11] Aalto University Wiki, Vector Quantization (VQ). `https://wiki.aalto.fi/pages/viewpage.action?pageId=149883153`, 2019.

[12] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165, 2019.

[13] M.h. Moattar and M.m. Homayounpour. A review on speaker diarization systems and approaches. *Speech Communication*, 54(10):1065–1103, 2012.

[14] C. Barras, Xuan Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505–1512, 2006.

[15] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. First dihard challenge evaluation plan. 2018.