MSc in NLP Supervised Project

Université de Lorraine

IDMC

Speaker diarization with overlapped speech
Bibliographical report

Authors:
Justine Diliberto
Cindy Pereira
Anna Nikiforovskaja

Supervisor:
Md Sahidullah

November 20, 2023

Abstract

In this project, we report the work done in the first phase of our project, aiming at improving speaker diarization with overlapped speech. We first discuss the speaker diarization in general and speaker diarization with the overlapped speech in particular. We present different important related work as a part of a bibliographical investigation, and analyze the acoustic characteristics of overlapped speech and the impact of overlapped speech on speaker diarization. Then, we perform experiments on the dataset provided for the second DIHARD Diarization Challenge. The main cause for speaker diarization errors has been found to be overlapped speech, especially in situations with background noise and when people are away from the microphone. Another issue is that a system performs poorly if there is too much overlapped speech in a recording, even for the non-overlap segments. The most recent diarization methods involve using neural networks, which are even more effective if combined with some signal processing techniques. These findings will be useful for the next phase of this project, whose purpose is to suggest an innovative method to solve the issues raised in this report.

Contents

List of Abbreviations

ANN	Artificial Neural Network
BIC	Bayesian Information Criterion
BLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DER	Diarization Error Rate
DNN	Deep Neural Network
EHMM	Ergotic Hidden Markov Model
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HSLN	Human Speech-Like Noise
JER	Jaccard Error Rate
LP	Linear Prediction
LPC	Linear Predictive Coding
MFB	MelFilter-Banks
MFCC	Mel-Frequency Cepstral Coefficients
ModSE	Modulation Spectrum Energy
NMF	Non-negative Matrix Factorization
NMS	Non-Maximum Suppression
PCA	Principal Component Analysis
PLDA	Probabilistic Linear Discriminant Analysis
RMSE	Root Mean Square Energy
RPNSD	Region Proposal Network based Speaker Diarization
SAD	Speech Activity Detection
SD	Speaker diarization
STLP	Sum of Ten Largest Peaks
VAD	Voice Activity Detector
VB-HMM	Variational Bayes Hidden Markov Model

Chapter 1

Introduction

The aim of this report is to understand the topic of *speaker diarization*, to give some analysis of the overlapped speech issue, and also to discuss prior works on overlapped speech detection.

This first chapter will be introducing the subject of this report by shortly explaining speech signal and Speaker Diarization (SD). This is followed by a discussion on the components, applications, and issues of SD. Finally, the scope and organization of this report will be presented.

1.1 Speech signal and speech technology

When a speech sound is produced, it is, in fact, a sequence of waves of energy that are created and start traveling the air (Quatieri, 2006). Uttered words are sequences of sound waves coming from our *phonatory system*. The breath flow is altered to fit the needs of the speech production as the air coming from the lungs is used for speech production. Then, vocal folds can be vibrating, opening or closing, and narrowing the gap of the larynx to allow different volumes of air to pass or not. After that, different oral cavity elements, and sometimes also nasal cavity, have a role in altering this flow of air even more by creating resonance. The speech signal is stored in digital storage as a sequence of samples encoded in different formats. The number of samples per second is known as *sampling rate*.

Speech technology involves the processing of speech signals by a machine. Speech sounds are analyzed by computing short-term characteristics representing acoustic and prosodic information. These components are then compared to stored patterns to recognize spoken words, speaker, emotion, and language.

1.2 What is speaker diarization?

Speaker diarization designates the act of dividing an audio input into different segments corresponding to different speakers, as described by Friedland & Leeuwen (2010). In other words, it is the task of finding *who spoke when* in an audio recording containing several speakers' voices. This involves the unsupervised identification of each speaker within an audio stream and of the durations during which each speaker is speaking (Anguera et al., 2012).

Speaker diarization is a relatively new field and thus is still in need of research and improvements. Some competitions such as DIHARD (Ryant et al., 2019b), the Rich Transcription Evaluation by the American National Institute of Standards and Technologies (Sadjadi et al., 2017) are organized to promote research in this field.

1.3 Components of speaker diarization system

As explained by Anguera et al. (2012), the general architecture of speaker diarization systems can be summarized with different steps as shown in Fig. 1.1. First, the audio data given as input is preprocessed. The *preprocessing* tasks aim at improving the quality of the input, and they can vary greatly according to the domain. They often consist in a *voice activity detector*, among other tasks. Next, *segmentation* is achieved and *speaker embeddings*, which are speaker representations, are extracted. Then, *cluster initialization* is performed. It depends on which type of approach is used by the system, which can be *bottom-up* or *top-down*. If the system has a bottom-up approach, a set of clusters will be selected at this step. On the contrary, if this is a top-down approach system a unique segment will be chosen. After that, the distances between clusters are calculated, in addition to applying splitting or merging tools several times, either to merge clusters or add new ones, as explained by D. Reynolds, Kenny & Castaldo (2009). The final step is called *stopping criterion* as it determines when the iteration of the two previous steps should stop.

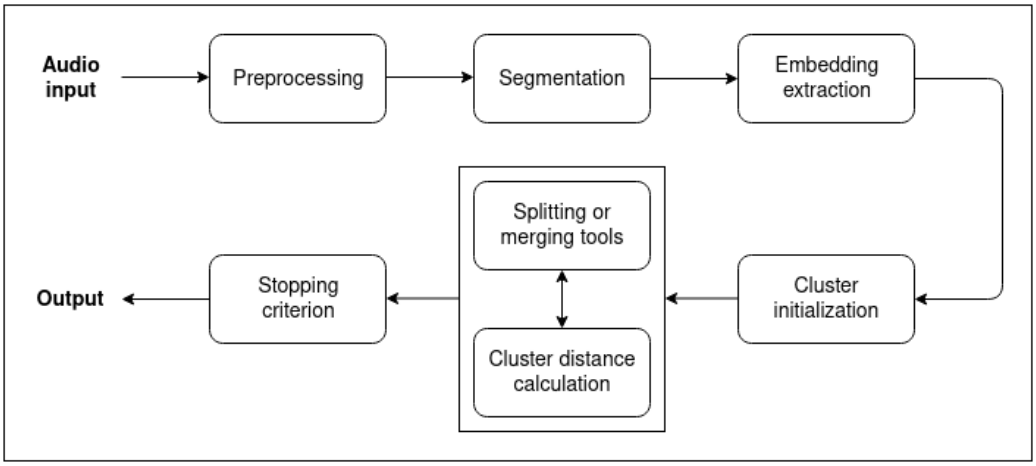

Figure 1.1: Components of a typical speaker diarization system.

The uniform segmentation for state-of-the-art speaker diarization systems is followed by speaker embedding extractions. Commonly, x-vector embeddings are extracted and they are used with clustering technique called Agglomerative Hierarchical Clustering. In addition, often, re-segmentation is also applied for frame-level refinements of results.

The majority of the clustering approach used in diarization systems belongs to one of the two following types: top-down or bottom-up approach. The most common one is the bottom-up approach and consists of the generation of several clusters that will be merged until one remains for each speaker. The top-down approach is the opposite, as one cluster is examined at the beginning and split into several ones. A bottom-up strategy called *agglomerative hierarchical clustering* (AHC) technique is predominantly used in state-of-the-art speaker diarization systems (Ryant et al., 2019b).

1.4 Application of speaker diarization technology

Speaker diarization is a useful tool and has many applications as evoked by Tranter & D. A. Reynolds, 2006, for instance:

- enabling automatic speaker-attributed speech-to-text transcription for interviews, meetings, conferences or courtroom audiences;
- ameliorating the task of searching and indexing audio archives;

- improving accuracy and reducing computational cost of automatic speech recognition, when used as a pre-processing step;
- speaker spotting in voice assistant technology.

1.5 Issues and Challenges

The state-of-art-speaker diarization systems show reasonably well performance in controlled conditions. However, the performance is degraded in realistic conditions due to the following reasons:

- overlapping speech where speech signals of two or more speakers are overlapped;
- background noise where speech signal is degraded by environmental sounds;
- distance variations between speakers and microphone.

1.6 Scope of the project

This project will focus on the particular issue of speaker diarization with overlapped speech, by providing performance analysis to determine the effects of overlaps, as well as acoustic analysis to understand causes for poor diarization results. It will also be aiming at presenting a list of bibliographical references of methods for detecting overlapping areas. These insights will be gathered with the objective of developing a new effective method for detecting overlapped speech during the second phase of our project.

1.7 Organization of the project

The second chapter, Chapter 2, will be focusing on the method, aiming at describing the dataset, the metrics, and the system used for our research. Then, in Chapter 3, several speech detection methods will be reviewed, classified according to their category. The acoustic and performance analyses of overlapped speech will be presented in the Chapter 4. Finally, a summary followed by our objectives for the next phase of this work will be presented in the Chapter 5.

Chapter 2

Experimental setup

2.1 Dataset description

2.1.1 Source of the dataset

The dataset used for our experimentation is the Second DIHARD Diarization Challenge dataset, as explained by Ryant et al. (2019a) and Sahidullah et al. (2019). The DIHARD Speech Diarization Challenges are a series of yearly challenges on speaker diarization. To be more precise, the task is to automatically determine *who spoke when* in a multi-speaker environment and using only audio recordings. These challenges are aiming at improving the field by suggesting datasets deemed to yield poor results in the current state-of-the-art. Indeed, development and evaluation datasets are provided by the organizers of the challenge, their goal being to support research and measure performance.

2.1.2 Types of track conditions

The tracks used as input can be single channels or multichannels. For the former, the channel can be coming from a single distant microphone, a distant microphone array, a combination of head-mounted and array microphones, or a combination of binaural microphones. Concerning multichannel tracks, each audio track is composed of the output from one or several distant microphone arrays, having multiple channels. In this multichannel condition, each array has to be computed separately.

Two different Speech Activity Detection (SAD) are included in the dataset: reference SAD and system SAD. The reference SAD condition characterizes systems that are supplied with a reference speech segmentation. This segmentation has been obtained through human annotation, by merging overlapping speech segments and removing speaker identification resulting from this annotation. On the contrary, systems SAD are supplied with the unprocessed audio input, thus the speech segmentation has to be generated.

These four conditions result in four different evaluation tracks (single channel using reference SAD; single channel using system SAD; multichannel using reference SAD; multichannel using system SAD).

2.1.3 Origins of the tracks

Both the training and evaluation data for single channel tracks are taken from eleven different domains such as audiobooks, broadcast interviews, child language, clinical, courtroom, map task, meeting, restaurant, socio-linguistic field and lab, or web videos. The combination of the tracks belonging to each domain is approximately as long as two hours.

The multichannel data comes from the CHiME-5 dinner party corpus. This corpus is composed

of real conversational speech, recorded in the homes of the participants during dinner parties. Twenty parties were organized, each lasting 2 to 3 hours and to which attended 2 hosts and 2 guests. The recordings were performed by Microsoft Kinect devices (producing 4 channel linear arrays). The locations were divided in three areas, and each had two of these devices, which produces 24 channels in total.

Every segment containing personal identifying information was removed before the publishing of the dataset. The files are 16 bit FLAC type for single channel and WAV type for multichannel, sampled at 16kHz. Concerning the reference SAD files for the development set, they are given as Rich Transcription Time Marked files.

2.2 Evaluation metrics

The results of the diarization are compared to those of a human segmentation, which is called *ground truth*. When the results are different from the ground truth, an error is identified. Three kinds of error can occur: speaker error, false alarm, and missed speech.

Speaker error refers to the assignment of a segment to the wrong speaker. A false alarm occurs when a segment has been assigned to a speaker but actually contains no speech. Missed speech is the term for a segment of speech that has not been assigned any speaker.

Two kinds of error rates are usually computed to consider the results of a diarization task. Diarization Error Rate (DER) is the most famous one and is used to determine the proportion of reference speaker time that is not correctly attributed to a speaker. It is obtained by adding the segments having one of the three kinds of errors (false alarm, missed speech, and speaker error) and dividing their result by the total speaker time.

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

Jaccard Error Rate (JER) is based on the Jaccard Index, aiming at computing the optimal mapping between a reference and system speaker pair. For each reference speaker, a specific JER can be drawn by dividing the sum of false alarms and missed speeches by the union of reference and system speaker segments. The JER is simply the average of every specific JERs.

$$\text{JER}_{ref} = \frac{\text{FA} + \text{MISS}}{\text{TOTAL}} \qquad \text{JER} = \frac{1}{N}\sum_{ref} \text{JER}_{ref}$$

2.3 Description of the speaker diarization system

2.3.1 State-of-the-art systems

The current state of the art for speaker diarization systems, as explained by Snyder et al. (2017), is turning away from previously used i-vectors to obtain speaker characteristics for the embedding extraction step. This new kind of system is focusing on the use Deep Neural Network embeddings to distinguish speaker differences, by mapping variable-length utterances to fixed-dimensional embeddings called x-vectors, however the challenge is to gather enough training data.

Snyder et al. (2018) introduce a method to expand the dataset by adding noise and reverberation to an existent dataset, and this method made the system become powerful enough. After that a Probabilistic Linear Discriminant Analysis classifier is used to compare the newly created embeddings.

The results of this new method are further discussed by Snyder et al. (2019), where it is annouced that the error rate for multiple speakers dropped and the performance for single speaker audios stayed the same. A successful method to remove domain sensitive threshold during the clustering stage is also presented.

2.3.2 Baseline system

The system we used is the baseline system supplied by the Second DIHARD Diarization Challenge, as defined by Ryant et al. (2019b). Four different tasks are performed, that is to say speech enhancement, beamforming, speech activity detection and diarization.

Firstly, a model is trained to forecast the ideal ratio masks from log-power spectra features using a densely-connected long short-term memory architecture, which is a kind of Deep Neural Network model particularly useful to make predictions.

Then, weighted delay-and-sum beamforming, a mathematical technique to identify the distance and orientation of sound waves caught by a microphone, is carried out.

After that, speech activity detection for tracks 2 and 4 is completed thanks to WebRTC's SAD, as found in the *py-webrtc* Python package (see 2.1.2).

Finally, the diarization is achieved by isolating each recording into small overlapping segments, extracting x-vectors, scoring using probabilistic linear discriminant analysis, and clustering with agglomerative hierarchical clustering (see 1.3).

Chapter 3

Review of overlapped speech detection methods

3.1 Overview

This chapter aims at providing a bibliographical review of some overlapped speech detection methods, belonging to three main categories that are either signal processing, statistics, or neural networks methods. The signal processing techniques are the first to have been invented, but some methods are still employed and improved nowadays. The statistical methods belong to the second generation of speaker diarization techniques, they were most utilized and developed during the period between 2000 and 2013. The neural network-based methods are the most recent as they appeared in the early 2010s and belong to the last generation of speaker diarization techniques.

3.2 Signal processing techniques

Kobayashi et al. (1996) explore the domain of Human Speech-Like Noise (HSLN) hoping to find a physical measurement inherent feature of a speech signal that would help the problem of overlapped speech in automatic speech detection. To generate HSLN, they normalize fifty sentences of phonetically balanced speech, they fold speech segments and superimpose those signals more than a thousand times. They found that by combining static parameters along with dynamic parameters, they could discriminate speech from background bubble noise more efficiently.

In their paper, Boakye, Vinyals & Friedland (2011) want to improve the previous work they had done with overlapped speech in speaker diarization, using feature analysis. To do so, they analyze the following speech features in order to select the essential ones for overlapped speech to incorporate them into their segmentation system: 12th-order Mel-Frequency Cepstral Coefficients (MFCCs), Root Mean Square energy (RMSE), Linear Predictive Coding (LPC) residual energy, diarization posterior entropy, spectral flatness, harmonic energy ratio, modulation spectrogram features, kurtosis, zero-crossing rate, and harmonicity. Compared to their previous work, they obtained a significant improvement in DER through the feature analysis technique.

The article by Zelenák & Hernando (2011) investigates a complementary method to a system based on short-term spectral parameters to handle the detection of overlapped speech regions, as measuring prosody features can bring some knowledge about the speakers and should help to differentiate them. The pitch, intensity, and four formant measures are selected by an algorithm with minimal-redundancy-maximal-relevance criterion and used to build a feature-set model for the system. The analysis shows that this method has slightly better results in terms of overlapped speech detection error, precision, and recall.

In their article, Heittola et al. (2013) explain how they tackle the problem of overlapping acoustic sound event detection by preprocessing the signal with unsupervised source separation. They use a Non-negative Matrix Factorization (NMF)-based method to separate the given audio, they

apply a continuous-density Hidden Markov Model (HMM) to model sound-event-conditional feature distributions, and finally they apply Viterbi algorithm to detect the sound event. This method allows a significant increase in performance compared to previous work using sound source separation.

Charlet, Barras & Liénard (2013) present a way to deal with overlapping speech segments for the diarization of broadcast news and debate videos. The general idea is to detect the overlapping segments to postpone their analysis to the moment after the labeling of the single speaker segments is done, the overlapping segments are then assigned the same speaker labels as their surrounding single speaker segments. Two different systems have been developed using cepstral features or a multi-pitch analysis, and their results were approximately similar with a 26% decrease of the DER in the best situation when compared to methods involving no overlapping speech detection.

Shokouhi & Hansen (2017) present a novel method to detect speech overlaps in monophonic recordings. They base their method on pyknograms, which are harmonically enhanced spectrograms first introduced in the paper by Potamianos & Maragos (1996). To detect the speech overlaps they count the euclidean distance between consequent pyknogram frames and afterward they introduce a segment-based score which is simply a mean distance between consequent frames on a specific segment of the recording. Finally, they separate classes based on the segment-based score, and classes with a higher score were considered to be overlap speeches. Evaluation showed promising results and they also made a few experiments to show that this method for overlap detection helps in a speaker verification task.

The method suggested by Baghel, Prasanna & Guhal (2020) aims at detecting transition points between overlapped and non-overlapped speech segments by using bag-of-audio-words. More precisely, the characteristics of three distributions are computed: the Sum of Ten Largest Peaks (STLP) of the spectrum and Mel-Frequency Cepstral Coefficients (MFCC) are used for estimating the vocal tract, the excitation source is evaluated through the Hilbert envelope of Linear Prediction (LP) residual signal, and the modulation spectrum is assessed thanks to the modulation spectrum energy (ModSE). Three features compose this analysis (3d, 13d, 16d) and the 16d feature scored the best result with an identification rate close to 75%.

3.3 Statistical methods

In their paper, Wrigley et al. (2005) offer a method to achieve reliable detection of speakers in multichannel audio. They use an Ergodic Hidden Markov Model (EHMM) with four states: speaker alone, speaker plus crosstalk, crosstalk, and silence to increase the flexibility of their system in comparison to previous work. They obtained a system that can distinguish between the four states of a recording, and which is particularly reliable for finding speaker alone activity.

Hu, Chieh-Cheng & Wei-Han (2007) propose an enhancement to previous work using the Gaussian Mixture Model (GMM) to find a suitable speaker's location detection algorithm that can detect multiple speech sources. They use the Gaussian Mixture location model and a location detection method to obtain a threshold of the probability of speaker for each location. This system has been proved to detect properly speaker's location and to reduce the average error rates.

Boakye et al. (2008) present a Hidden Markov Model (HMM)-based method to create an overlap detection system along with a diarization segment post-processing procedure. They model state emission probabilities with a multivariate GMM and they compute the frame-level speaker posterior probability which they sum to obtain a score for each speaker, the highest one being the one assigned to the segment. The proposed method provided key directions to follow for future work on overlapped detection systems.

Huijbregts, Van Leeuwen & Jong (2009) develop an overlapped speech detection model and use it in two separate ways. They assume that the overlapping speech can be represented in GMM and that at each speaker change there is a higher probability of having an overlap. So they train their model to predict 500ms before and after the speaker change. Afterwards they use their model with HMM and Viterbi iterations and run diarization with and without overlap model, choosing the results with the best likelihood. They also use this overlapping model as a final run to detect overlapping regions. This overlapping speech detection model improved all of their results in terms

of DER, even though the number of false alarms of this score increased, and also this approach is considered to be domain-independent.

In their paper, Shum et al. (2011) explore the low-dimensional Total Variability subspace approach to speaker clustering. First, they segment the speech with a Modulation Frequency-based Voice Activity Detector (VAD). Then they apply Principal Component Analysis (PCA)-based projections in the Total Variability space, which adds to a speaker- and session-dependent supervector a rectangular matrix of low rank along with a total factor vector, in order to better represent speaker variabilities and compensate for channel inconsistencies. This way, they simplified the previous systems and still achieved state-of-the-art performance.

Silovsky et al. (2011) submit a Probabilistic Linear Discriminant Analysis (PLDA)-based method to improve the clustering module for broadcast streams speaker diarization. The common baseline is composed of the following steps: feature extraction, SAD using both an energy-based detector with adaptive threshold and a Gaussian Mixture Model (GMM) based detector, speaker segmentation using Bayesian Information Criterion (BIC) technique, and segment clustering involving BIC. The proposed system brought innovative segment clustering modules, that are multifold- and onefold-PLDA based, and the latter obtained a higher improvement when compared to the baseline system, with 42% less speaker error rate for the case of 2-stage clustering.

Yella & Bourlard (2013) introduce the interesting idea that it may be better to find overlap segments by using not only short-term features of the small segments but also long-term ones. To add long-term features they train also a Poisson distribution model to predict the number of overlaps based on the number of speaker changes in a given bigger time segment. This Poisson distribution model is incorporated into the baseline HMM/GMM overlap detection model, which leads to a baseline diarization model quality increase in terms of DER.

3.4 Neural network based methods

Snyder et al. (2017) examine a Deep Neural Network (DNN) method consisting of replacing i-vectors with embeddings generated by a feedforward DNN to differentiate speakers from segments with variable lengths. A temporal pooling layer is added in the DNN and gathers long-term speaker characteristics before the speaker and speech segment pairs are put together using a PLDA-based feature. This method greatly enhances results for short speech segments, and a lesser improvement is noted for long speech segments.

A new system using Convolutional Neural Network (CNN) is given by Zajíc, Hrúz & Müller (2017) to enable statistics accumulation refinement. The purpose is for the CNN to output a probability value for a speaker change in a given segment, and to this end, each input is cut into small segments represented by i-vectors. This technique allows a better speaker representation in the last i-vector, as a notable improvement is acknowledged in the article proving that the DER decreased by 16% when compared to the baseline.

Yoshioka et al. (2018) deal with the overlapping speech introducing an *unmixing transducer*, which separates a multichannel recording into a fixed number of time-synchronous audio streams. They base their unmixing transducer on windowed Bi-directional Long Short-Term Memory (BLSTM) recurrent neural layers, this model can be trained on short recordings. They tested this model on the meeting transcription task and on real-life meetings. The model showed better results than other meeting transcription models.

The article by Hogg, Evers & Naylor (2019) offers a method to perform multiple hypothesis tracking to segment overlapping areas. The main goal is to use the harmonic structure in relation to the pitch. This method involves steps such as harmonic subset generation, tracking multiple hypotheses with maximum weighted clique, and multiple Kalman filters for pitch tracking. To conclude, it is indeed possible to detect the presence of overlapped speech and the results are comparable to recent machine learning methods.

Kunešová et al. (2019) use generated data to train a CNN for overlap detection. They took several corpora with recordings of single phrases and combined them into recordings with overlap with different volume levels and some background noise. The model performed well on clean and noise-free data, however it did not perform that well on noised data. Overall an approach using

generated data seems promising.

Andrei, Cucu & Burileanu (2019) want to train several neural networks to detect overlapped speech and estimate the number of competing speakers at a given time on English language based on human capacity. They use their previous work to detect overlapped segments but target to count speech sources on short signal fragments (25ms), using the single speaker periods to build voice profiles and they train a new CNN model for each targeted timeframe. Finally, they got better results than current literature and their system has a higher performance than humans.

In the paper by Bullock, Bredin & Garcia-Perera (2020) they use bidirectional Long Short-Term Memory (BLSTM) recurrent neural layers in their neural network architecture to distinguish speech segments with overlap. Afterwards they perform resegmentation and final diarization using an i-vector-based Variational Bayes Hidden Markov Model (VB-HMM). As a result, their model for overlap detection beats state-of-the-art for several datasets and sets the baseline for future experimentation on DIHARD II. Moreover, their experiments with speaker diarization showed that using oracle overlapped speech detection provided in the dataset only made minor improvements on DER, which means that their model for diarization may be more likely improved by better speaker assignment, not overlap detection.

In their paper, Huang et al. (2020) introduce a new speaker diarization method called Region Proposal Network based Speaker Diarization (RPNSD). With this method, they combine the segmentation, embedding extraction, and re-segmentation (every step of a standard diarization system) into one neural network, and the only task left after this NN application is to apply clustering and non-maximum suppression (NMS) to predict the diarization. They obtained good improvements over the actual results and still have a shorter pipeline working well with overlapped segments.

Kanda et al. (2020) developed a method for overlapped speech recognition, based on Attention-based Encoder Decoder. They use d-vectors, which are speaker profile vectors, they represent the voices of the speaker who can possibly be in the recording. Their model extracts recording features and speaker features using encoders, keeping in mind the possible speaker profiles with attention mechanism. Afterwards they apply a decoder to get a transcript of the recording. Their model beats the baseline they had.

In the article by Kinoshita, Delcroix & Tawara (2020) an innovative approach is considered and consists in bringing together two existing approaches with the aim to keep the advantages of both and to cast aside their imperfections. By mixing a clustering-based approach, which performs greatly for long audios but fails if overlap speech is present, and an end-to-end neural diarization approach, which handles overlap speech accurately but its flaws are dealing with long audios as they require a huge amount of computational memory and time. When considering results, there has been a significant improvement of the proposed method if the test data is longer than 5 minutes, and if it is shorter both the new method and regular end-to-end neural method are equal.

Málek & Žďánsky (2020) explore the idea of using x-vectors not only to differentiate speakers in a recording as it is currently done, but also to extract other features from a front-end x-vector network. The theory that front-end x-vectors enclose more information than only speaker information is tested, such as voice activity detection, overlapped speech detection, speaker identification, and absence of speech. This would considerably reduce the computational needs, as the different tasks evoked earlier are usually done separately. This method surpasses the usual features, however the accuracy is lessened if there is too much background noise.

In their article, Raj, Huang & Khudanpur (2020) use another approach for overlapped speaker diarization which uses an external overlap detector to apply the clustering of the segment-level embeddings. Their clustering rely on the following two steps: first they ignore the discrete constraints to relax the Non-deterministic Polynomial-time hardness discrete clustering into a continuous version, and generate a solution set through orthonormal transformations of eigenvectors of the normalized Laplacian. Finally they find a discrete solution close to any of the solutions obtained and modify the "sum-to-one" constraint in the discretization stage to introduce overlap awareness while self-tuning the clustering process with p-binarization and normalized maximum eigengap techniques. Their method provided an improvement over standard single-speaker clustering models and was even competitive with other overlapped diarization methods.

Raj et al. (2020) introduce a method, called DOVER-Lap, for speaker assignment to the over-

lapped speech regions which combines an output from several diarization systems. They have two stages in this method. There is a label mapping stage to match the hypotheses' speakers from each diarization system, and a label voting stage to finally decide who was speaking during a particular segment. For the label mapping stage, they build a weighted K-partite graph of hypotheses, where K is the number of hypotheses, and find a maximum matching there. These stages are processed repeatedly, considering all the weights with a greedy approximation algorithm. As for voting stage they decide that the amount of speakers in the given region is a weighted mean number of speakers in the hypotheses, and then they perform a majority voting to choose those speakers. They performed an evaluation combining different diarization system and the methods showed a consistent and significant improvement.

Yousefi & Hansen (2020) use CNN to classify speech as overlap or non-overlap, based on several different features. They explore the effects of different features such as spectral magnitude, pyknogram, MelFilter-Banks (MFB), and MFCC. It turned out that using pyknograms as features provides the best performance, giving an increase of 10% in accuracy and 15% in F1-score, comparing to the previous results. However, pyknogram based model is computationally less efficient than models using MFB and MFCC features.

3.5 Summary of the methods

Many different methods for overlapped speech detection were introduced in the last years. These overlapped speech detection methods are used not only in speech diarization but also in other speech-related tasks. Even though some of the methods were only applied for overlapped speech for non-diarization tasks, the ideas from those methods could still be used in our own methods.

Even though, fully signal processing techniques are not that popular anymore, they are still used in the most successful deep learning methods, for example they used pyknograms in the paper by Yousefi & Hansen (2020). The same goes for statistical methods. Even though they are not really developed solely nowadays, they are used as pre-processing or post-processing to boost deep learning methods, it is noticeable that HMM and GMM are used a lot.

Another interesting thing in the methods is that some used self-generated data or even no data at all. It definitely leads to domain-independent methods and it may be useful, even if those new methods may not be as efficient as domain-dependent ones.

Finally, there are also methods to combine the output of several diarization methods to increase the performance, like DOVER-Lap. This method can also be useful in our future work.

It is clear that currently BLSTM and CNN based techniques are the most effective in terms of quality of overlapped speech detection. However, there is still room both in terms of quality performance and computational efficiency.

There are several ways of improving the baseline we have mentioned in 2.3, using described methods. For example we could first perform overlap detection described by Bullock, Bredin & Garcia-Perera (2020) and then remove the overlapped regions and the resulting recordings pass to the baseline model. Finally, we could return the overlapped speech regions and provide a speaker reassignment based on x-vectors representation. Another way we could train the BLSTM on MFCC features, which are anyway extracted in our baseline method and then make an overlap-aware resegmentation from the same paper using the VB-HMM module, leaving the initial segmentation from the baseline method.

Chapter 4

Analysis of overlapped speech and its impact on speaker diarization performance

4.1 Acoustic analysis

We have conducted some analysis of various audios using Praat[1] software so that we could understand what are the possible variables that lead to an unclear signal. We have chosen to use recordings of ourselves because we were able to control these variables.

First, we have tried to distinguish the differences between a recording of two people speaking at the same time and an audio file with only one person speaking. We encountered many problems in the analysis because more than one variable was evaluated there: the pronounced sentences weren't the same and the overlapped speech was produced with people from different genders.

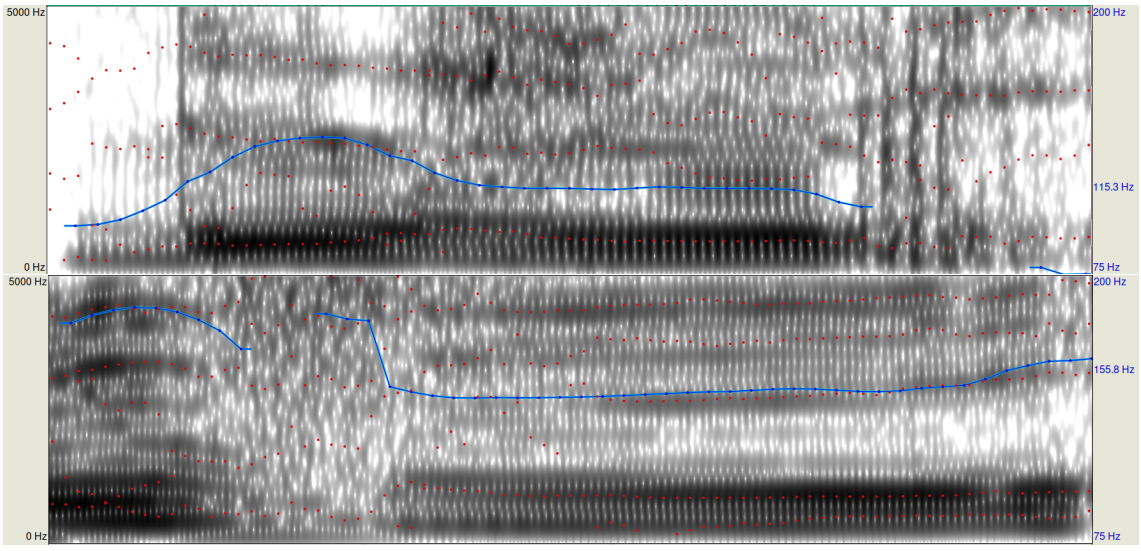

Figure 4.1: The first figure shows the spectrogram of a male voice. The second figure shows the spectogram of a female voice.

The first variable we could control and detect was the gender of the speaker. We recorded the same French word "*Bonjour*" told by a man and a woman in a quiet environment. Both of them spoke near the microphone and conditions were optimal. One can notice in fig. 4.1 that the male recording has a lower pitch (115.3 Hz) than the female one (155.8 Hz). Knowing that, we used

[1] https://fr.wikipedia.org/wiki/Praat

only female recordings for the developments of analysis.

We have generated an overlapped speech composed of two single recordings (fig. 4.2: 122.4 Hz and fig. 4.3: 121.1 Hz) using computer tools to add them, such as Audacity *Mix* option or Python's *pydub* library[2]. In each single audio, a different woman was telling the same English sentence: "*Nevermind how long*". The environment was quiet, both of them were speaking near their microphone, and the conditions were optimal. We have noticed that when there are two really clear audios with no noise and people speaking near their microphone saying the same sentence, the computer-generated overlapped speech (fig. 4.4: 117.3 Hz) is quite clear and looks like the mean of the two audios. The only difference is the fact that the pitch seems to be slightly lower than in both single recordings. Unfortunately, the conditions are never that great in real life, and two people never say the same sentence at the same time.

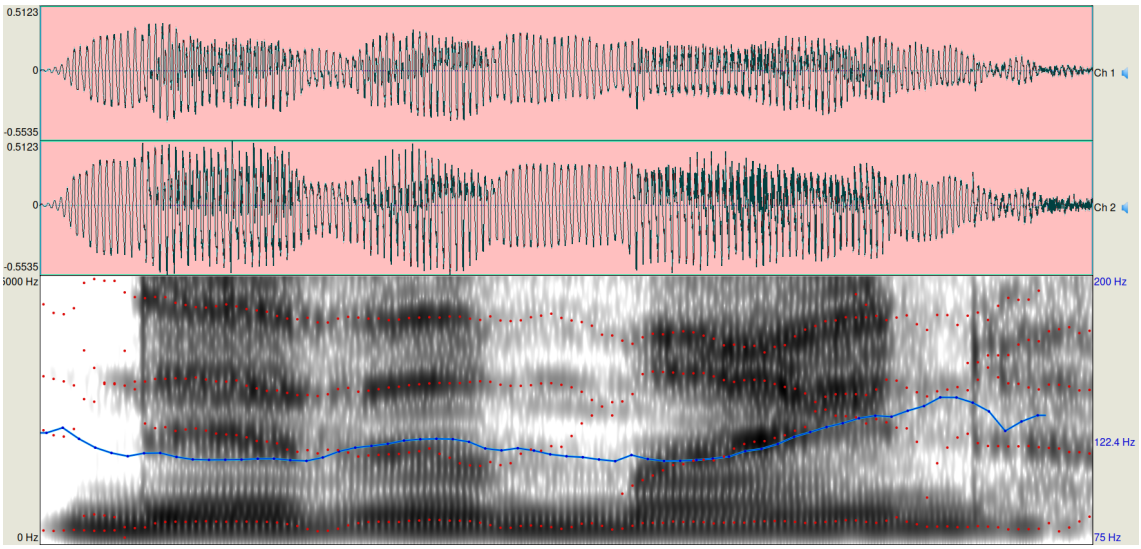

Figure 4.2: A spectrogram of the first speaker.

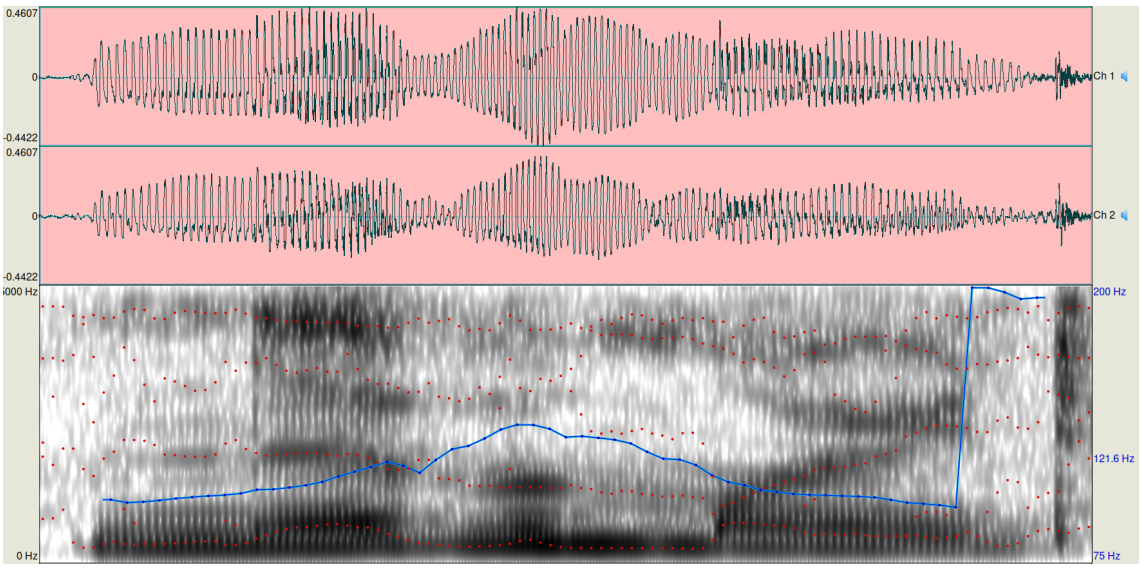

Figure 4.3: A spectrogram of the second speaker.

However, when we record someone speaking over other people chatting or in a noisy environment, the signal becomes really unclear and it is way more difficult to detect clear information, as one can see in fig. 4.5, which is the spectogram of an English conversation with more than three people speaking and laughing at the same time.

[2]https://github.com/jiaaro/pydub

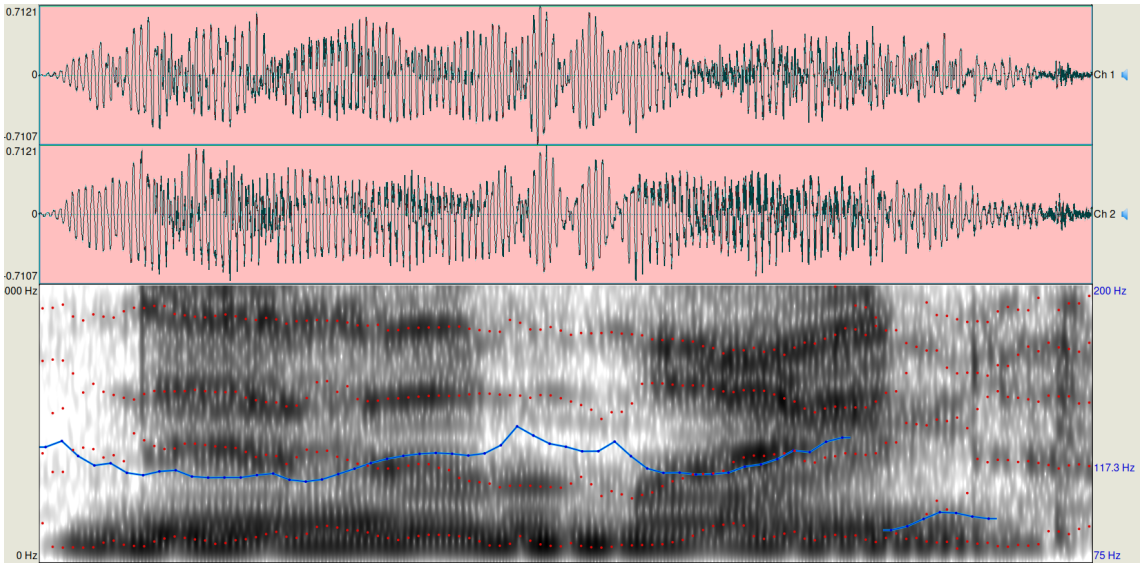

Figure 4.4: A spectrogram of a computer-generated overlapped speech.

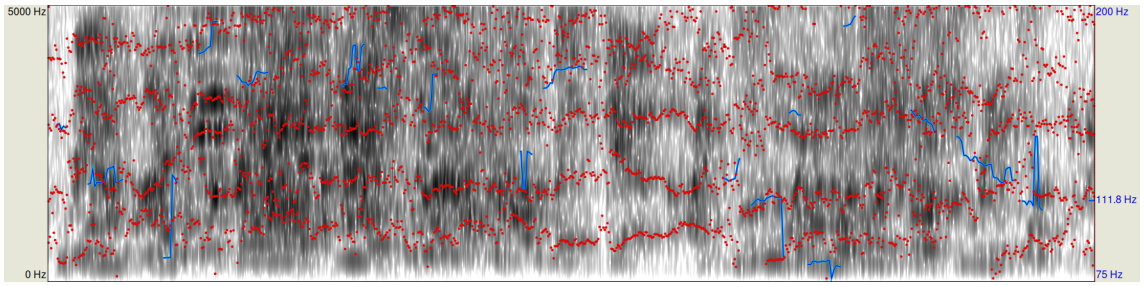

Figure 4.5: A spectogram of an overlapping speech with many people talking at the same time.

Thus, speaker diarization becomes really difficult when handling overlapped speech because signals undergo huge changes in comparison to single-speaker audio.

4.2 Performance analysis

We have performed an analysis of the quality of the speaker diarization and what possibly led to a lack of quality on some recordings. This performance analysis was held on the DIHARD II dataset using the baseline for the year 2019 [3], described in 2.1.

Particularly we have taken the web video group of recordings from the dataset for the analysis because this group contains diverse recordings both with a small and huge amount of speakers, and both with little to a huge amount of distortion.

After running track 1 baseline solutions we studied the resulting DER values. Let's first look into the properties of the recording with a big DER value (particularly recording DH_0156 with DER value equal to 70.22). One can see the visualization of when speakers were talking according to the original recording compared to the diarization results in fig. 4.6.

It is noticeable that in the taken recording file there were a lot of speech overlaps, which makes it harder for the diarization algorithm to perform well. As a result of this experiment, we have decided to check how important the amount of overlaps is to the performance of the model.

To see how important the overlaps for the performance are, we have calculated the amount of the overlaps and then the quality of the diarization without overlaps. We present the results of these calculations in fig. 4.7. Full results are in the table 4.1. Percent of overlaps is calculated as

[3] https://github.com/iiscleap/DIHARD_2019_baseline_alltracks

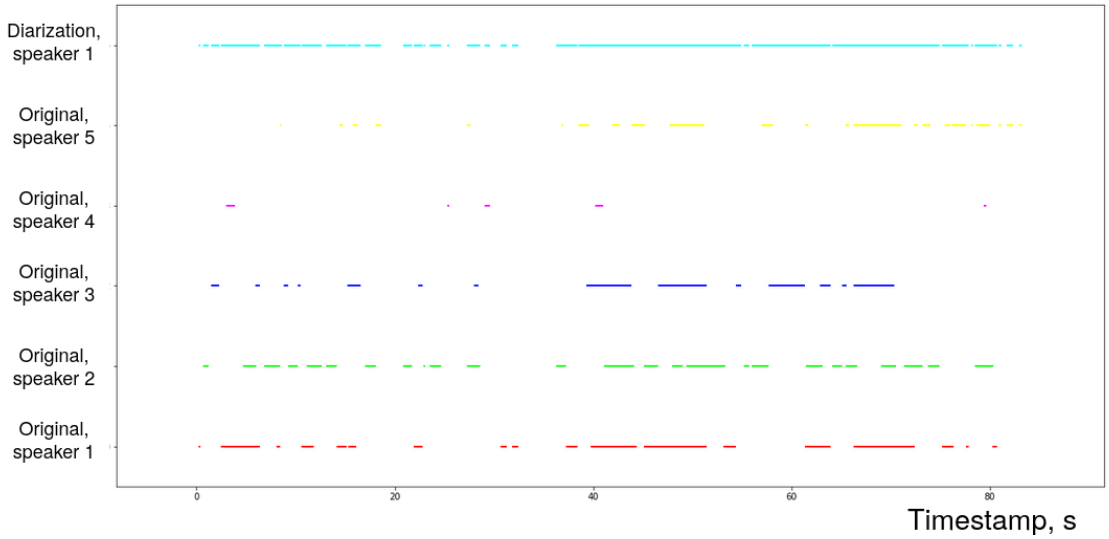

Figure 4.6: A graph showing when each speaker was speaking in the original recording and the resulting speaker diarization.

a ratio of seconds of overlapped speech to number of seconds in the whole recording (including silence, if there is any).

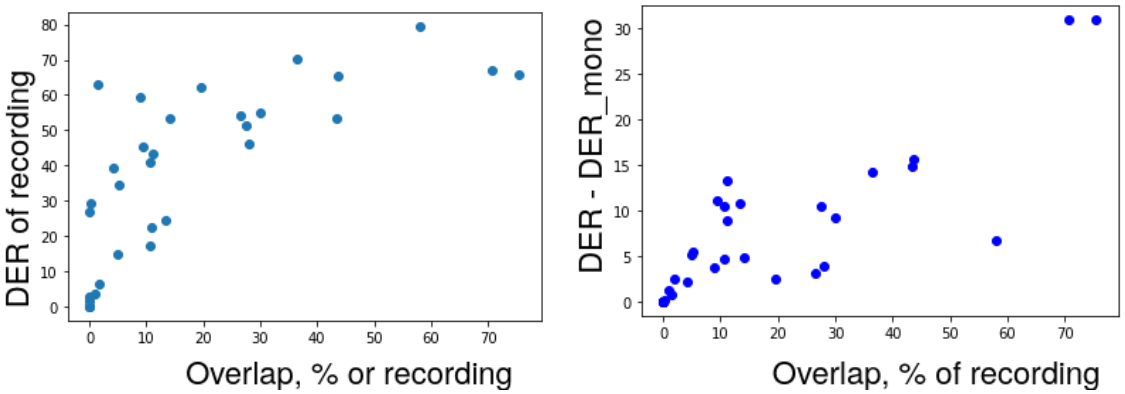

Figure 4.7: The first figure shows the dependence of DER value on the percent of the overlap. The second figure shows the dependence of the improvement of DER score on non-overlap regions of the recording.

As can be seen from the figures, the DER value tends to increase when the amount of overlap is bigger. Moreover, the DER value on the non-overlap regions is strictly less than on overlap regions and also is dependent on the amount of overlap. It means, that when an amount of overlap is big, it makes it harder for the model even to perform diarization on the non-overlap regions, even though those regions are usually easier for the model.

It seems like an issue, which we should deal with while developing our own diarization methods. To do this we can implement modern overlap-detection techniques based on BLSTM and CNN we described in 3. Moreover, we can combine different diarization systems with DOVER-Lap to increase the quality.

File	% of overlap	DER, full	JER, full	DER, non overlap	JER, non overlap
DH_0149	1.49	63.02	90.39	62.33	90.57
DH_0150	0.19	29.07	58.39	28.92	58.42
DH_0151	14.03	53.41	90.51	48.59	91.43
DH_0152	10.69	40.77	76.74	36.16	68.08
DH_0153	27.57	51.28	78.50	40.77	77.59
DH_0154	75.52	65.61	83.64	34.70	84.61
DH_0155	0.94	3.69	41.35	2.47	39.75
DH_0156	36.46	70.22	89.64	56.03	91.21
DH_0157	0.0	0.00	0.00	0.00	0.00
DH_0158	9.44	45.20	84.34	34.09	83.52
DH_0159	11.07	43.25	82.99	30.04	82.94
DH_0160	13.42	24.25	40.96	13.42	36.73
DH_0161	4.97	14.78	75.13	9.57	67.57
DH_0162	70.66	66.81	72.33	35.91	79.12
DH_0163	4.18	39.47	77.92	37.27	79.10
DH_0164	8.81	59.56	84.51	55.86	84.55
DH_0165	26.51	53.96	78.24	50.79	83.58
DH_0166	19.49	62.14	82.48	59.72	85.72
DH_0167	43.47	53.47	91.21	38.68	91.23
DH_0168	5.17	34.44	69.41	28.97	68.54
DH_0169	0.0	0.00	0.00	0.00	0.00
DH_0170	58.06	79.34	90.10	72.71	94.35
DH_0171	43.6	65.57	90.83	50.02	90.00
DH_0172	30.01	54.85	87.40	45.68	86.43
DH_0173	27.94	46.27	72.19	42.39	74.76
DH_0174	1.83	6.30	68.89	3.87	69.27
DH_0175	0.0	0.00	0.00	0.00	0.00
DH_0176	0.0	1.67	9.53	1.67	9.53
DH_0177	10.65	17.41	21.16	6.90	12.60
DH_0178	11.0	22.39	31.66	13.46	27.49
DH_0179	0.0	26.99	26.99	26.99	26.99
DH_0180	0.0	2.59	51.29	2.59	51.29

Table 4.1: All the statistics for files from *webvideo* group, including percent of overlap and DER and JER calculated both on full recordings and non-overlap regions.

4.3 Impact of speech overlap in full dataset

In a separate experiment, we have computed SD performance in terms of DER with and without overlap on the full development set of the DIHARD II corpus. The experiment was done with all 192 speech files and the same baseline system as used in the previous section.

The results are shown in Table 4.2. The DER is reduced by more than 40% compared to the condition that includes overlap. This confirms that SD performance can be substantially improved if SD system is capable to accurately handle the overlapped speech.

Overlap	DER (%)
Yes (Baseline)	23.74
No	14.08

Table 4.2: Speaker diarization performance on full DIHARD II dataset (development) for with and without overlapped speech.

Chapter 5

Conclusion

5.1 Summary

This project has presented the outcome of the preliminary part of a project intended to improve speaker diarization with overlapped speech.

Speech formation comes from the phonatory system, and consists of sound waves. Speech technology has been defined as the analysis of components from these sound waves. A description of speaker diarization has been given, which is the task of defining "who spoke when", before illustrating it with its five most common components: preprocessing, cluster initialization, splitting or merging tools, cluster distance calculation, and stopping criterion. Some applications and issues for speaker diarization have then been exposed.

The dataset used for our performance analysis has been taken from the Second DIHARD Diarization Challenge, which is composed of four types of tracks coming from various domains, and that can be single or multichannel and with reference or system SAD. The DER and JER evaluation metrics have been explained. Lastly, the baseline for the Second DIHARD Diarization Challenge is the system we used for our analyses.

We have studied several articles to give an outline of the method for speaker diarization they are offering. These methods have been classified into three chronologically ordered categories: signal processing techniques, statistical methods, and neural network methods. There has been an evolution in the use of these categories, which appeared at different times. Thus the most recent one is the most commonly used, however the oldest ones were not abandoned, as they are still applied in addition to neural network methods and in pre-processing or post-processing steps.

Finally, the acoustic analysis focused on perceiving the physical causes for speaker diarization errors, which are the presence of background noise and a variation in the distance from the microphone between two speakers. The performance analysis proved that the DER values increase when there is overlapped speech in the audio, and even the non-overlapped speech segments from this same audio are affected. However it needs to be indicated that this analysis was not performed on the full dataset, as we focused on web videos taken from track 1.

5.2 Future work

In the second phase of our work, an innovative approach will be developed based on the knowledge we gathered during this phase. This method will be based on deep learning techniques, which have become more prominent these last few years, and possibly by combining them with other methods to obtain an improvement of performances, as we have seen that combined methods are more effective. To be precise, we describe our plan for the next phase in the following paragraph.

First, we should try modern speech overlap detection methods to improve the baseline method we have and compare the results. We think that this would take about two months in total.

Afterwards it would be interesting to improve the quality of speaker assignment by using such methods as DOVER-Lap, described by Raj et al. (2020), and compare the results. We think, this part will take about one month of work. Then we leave a month to find any possible problems with the solutions we have and to probably improve the diarization methods itself.

Finally, a month will be left for the project finalization and running the results on a different dataset. For example, during the last phase we can also compare out new method to the baseline on DIHARD III, introduced by Ryant et al. (2020). The challenge for DIHARD III contains only two tracks, one with SAD and one for diarization from scratch. As we are mostly focused on the track with SAD in DIHARD II, we will evaluate our resulting method on the first track of DIHARD III dataset.

Bibliography

Andrei, V., Cucu, H. & Burileanu, C. (2019) Overlapped Speech Detection and Competing Speaker Counting – Humans Versus Deep Learning. *IEEE Journal of Selected Topics in Signal Processing*. 13, 850–862.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. & Vinyals, O. (2012) Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*. 20 (2), 356–370.

Baghel, S., Prasanna, S. M. & Guhal, P. (2020) Overlapped/Non-Overlapped Speech Transition Point Detection Using Bag-of-Audio-Words. *2020 IEEE International Conference on Signal Processing and Communications*, 1–5.

Boakye, K., Trueba-Hornero, B., Vinyals, O. & Friedland, G. (2008) Overlapped speech detection for improved speaker diarization in multiparty meetings. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4353–4356.

Boakye, K., Vinyals, O. & Friedland, G. (2011) Improved Overlapped Speech Handling for Speaker Diarization. *12th Annual Conference of the International Speech Communication Association*, 941–944.

Bullock, L., Bredin, H. & Garcia-Perera, L. P. (2020) Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. Barcelona, Spain, 7114–7118.

Charlet, D., Barras, C. & Liénard, J.-S. (2013) Impact of overlapping speech detection on speaker diarization for broadcast news and debates. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada, 7707–7711.

Friedland, G. & Leeuwen, D. van (2010) Speaker recognition and diarization. *Semantic Computing*, 115–129.

Heittola, T., Mesaros, A., Virtanen, T. & Gabbouj, M. (2013) Supervised model training for overlapping sound events based on unsupervised source separation. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8677–8681.

Hogg, A. O., Evers, C. & Naylor, P. A. (2019) Multiple hypothesis tracking for overlapping speaker segmentation. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 195–199.

Hu, J.-S., Chieh-Cheng, C. & Wei-Han, L. (2007) A Robust Statistical-Based Speaker's Location Detection Algorithm in a Vehicular Environment. *EURASIP Journal on Advances in Signal Processing*. 2007.

Huang, Z., Watanabe, S., Fujita, Y., García, P., Shao, Y., Povey, D. & Khudanpur, S. (2020) Speaker Diarization with Region Proposal Network. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6514–6518.

Huijbregts, M., Van Leeuwen, D. A. & Jong, F. (2009) Speech overlap detection in a two-pass speaker diarization system, 1063–1066.

Kanda, N., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Zhou, T. & Yoshioka, T. (2020) Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. *arXiv preprint arXiv:2006.10930*.

Kinoshita, K., Delcroix, M. & Tawara, N. (2020) Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. *arXiv preprint arXiv:2010.13366*. abs/2010.13366.

Kobayashi, D., Kajita, S., Takeda, K. & Itakura, F. (1996) Extracting speech features from human speech like noise. *Proceeding of 4th International Conference on Spoken Language Processing*. Vol. 1, 418–421 vol.1.

Kunešová, M., Hrúz, M., Zajíc, Z. & Radová, V. (2019) Detection of overlapping speech for the purposes of speaker diarization. *International Conference on Speech and Computer*, 247–257.

Málek, J. & Žďánsky, J. (2020) Voice-Activity and Overlapped Speech Detection Using x-Vectors. *International Conference on Text, Speech, and Dialogue*, 366–376.

Potamianos, A. & Maragos, P. (1996) Speech formant frequency and bandwidth tracking using multiband energy demodulation. *The Journal of the Acoustical Society of America*. 99 (6), 3795–3806.

Quatieri, T. F. (2006) Discrete-time speech signal processing: principles and practice. Pearson Education, 781.

Raj, D., Garcia-Perera, L. P., Huang, Z., Watanabe, S., Povey, D., Stolcke, A. & Khudanpur, S. (2020) DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs. *arXiv preprint arXiv:2011.01997*. abs/2011.01997.

Raj, D., Huang, Z. & Khudanpur, S. (2020) Multi-class Spectral Clustering with Overlaps for Speaker Diarization. *arXiv preprint arXiv:2011.02900*.

Reynolds, D., Kenny, P. & Castaldo, F. (2009) A study of new approaches to speaker diarization. *10th Annual Conference of the International Speech Communication Association*, 1047–1050.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S. & Liberman, M. (2019a) Second dihard challenge evaluation plan. *Linguistic Data Consortium, Tech. Rep*.

— (2019b) The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines. *20th Annual Conference of the International Speech Communication Association*, 978–982.

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S. & Liberman, M. (2020) The Third DIHARD Diarization Challenge. *arXiv preprint arXiv:2012.01477*.

Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C. S., Reynolds, D. A., Singer, E., Mason, L. P. & Hernandez-Cordero, J. (2017) The 2016 NIST Speaker Recognition Evaluation. *18th Annual Conference of the International Speech Communication Association*, 1353–1357.

Sahidullah, M. et al. (2019) The Speed Submission to DIHARD II: Contributions & Lessons Learned. *arXiv preprint arXiv:1911.02388*.

Shokouhi, N. & Hansen, J. H. L. (2017) Teager–Kaiser Energy Operators for Overlapped Speech Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 25 (5), 1035–1047.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. & Glass, J. R. (2011) Exploiting Intra-Conversation Variability for Speaker Diarization. *12th Annual Conference of the International Speech Communication Association*.

Silovsky, J., Prazak, J., Cerva, P., Zdansky, J. & Nouza, J. (2011) PLDA-based clustering for speaker diarization of broadcast streams. *12th Annual Conference of the International Speech Communication Association*.

Snyder, D., Garcia-Romero, D., Povey, D. & Khudanpur, S. (2017) Deep Neural Network Embeddings for Text-Independent Speaker Verification. *18th Annual Conference of the International Speech Communication Association*, 999–1003.

Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D. & Khudanpur, S. (2019) Speaker recognition for multi-speaker conversations using x-vectors. *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5796–5800.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. (2018) X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5329–5333.

Tranter, S. E. & Reynolds, D. A. (2006) An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*. 14 (5), 1557–1565.

Wrigley, S. N., Brown, G. J., Wan, V. & Renals, S. (2005) Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*. 13 (1), 84–91.

Yella, S. H. & Bourlard, H. (2013) Improved overlap speech diarization of meeting recordings using long-term conversational features. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7746–7750.

Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X. & Alleva, F. (2018) Recognizing Overlapped Speech in Meetings: A Multichannel Separation Approach Using Neural Networks. *19th Annual Conference of the International Speech Communication Association*, 3038–3042.

Yousefi, M. & Hansen, J. H. L. (2020) Frame-Based Overlapping Speech Detection Using Convolutional Neural Networks. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6744–6748.

Zajíc, Z., Hrúz, M. & Müller, L. (2017) Speaker Diarization Using Convolutional Neural Network for Statistics Accumulation Refinement. *18th Annual Conference of the International Speech Communication Association*, 3562–3566.

Zelenák, M. & Hernando, J. (2011) The detection of overlapping speech with prosodic features for speaker diarization. *12th Annual Conference of the International Speech Communication Association*, 1041–1044.