Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Identification of systematic differences in acoustic properties in categories of laughter in "MULTISIMO" dataset

Mingwei Shi

Supervisor: Dr.Carl Vogel

April 2024

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MCS(Integrated Computer Science)

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

This work is conducted without using generative AI in any element .

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: _____ Date: _____

Permission to Lend and/or Copy

I agree that the Trinity College Library may lend or copy this dissertation upon request.

Signed: _____ Date: _____

Abstract

Research into laughter classification is a compelling field that captivates scientists and sociologists seeking to unravel the enigmatic nature of this social signal. This paralinguistic cue possesses a notably intricate acoustic structure. Unveiling its discriminating properties could shed light on the internal acoustic structure of laughter. Previous studies have undertaken experiments to identify these discriminating acoustic properties, presenting a comprehensive pipeline that spans machine learning selection, identification of discriminating properties, and exploring factors influencing them. However, previous research has not released its dataset publicly, and some procedures require enhancement. To construct a more rigorous pipeline and comprehensively analyse discriminating acoustic properties, we compiled our dataset tailored to our research objectives from the "MULTISIMO" raw corpus(Multimodal and Multiparty Social Interactions Modelling), followed by identifying discriminating properties in mirthful and discourse laughter within our constructed dataset, performing regression analysis on the datasets to identify significant features that could explain discourse and mirthful laughter ,and exploring factors influencing discriminative acoustic properties.

The main findings in our work highlight that through a synthesis of the results from the machine learning experiments and regression analysis, we identified five shared discriminating acoustic properties across both experiments and laughter types: fundamental frequency, mel-frequency cepstral coefficient, auditory spectrum, spectral features, and jitter. The first four properties gauge energy-related information in acoustic laughter, while the last describes temporal characteristics. Our findings exhibit both concurrence and disparity with the findings from Tanaka and Campbell(2014),our replicated work, attributable to differences in the acoustic feature set quantity and the total number of utterance instances. Notably, fundamental frequency and spectral features emerge as common discriminating properties in both studies.

This work makes significant contributions both in theory and practice. Theoretically, this research has established a comprehensive pipeline encompassing dataset construction, verification, machine learning design and implementation, identification of acoustic properties, and examination of factors that may influence discriminating properties. This pipeline presents a novel avenue for researchers in audio processing and artificial intelligence. In terms of practical applications, although the study emphasises lies in theory, the developed algorithm shows promise for integration into real-time video systems to assist in laughter classification. It enables dynamic tracking of specific acoustic properties unique to instances of laughter.

Key phrases: Human laughter, Natural language dialogue, Social signals

Acknowledgements

Thanks to Prof.Carl Vogel,and my parents.

University of Dublin,Trinity College
April 2024

Mingwei Shi

Contents

List of Figures

List of Tables

List of Algorithms

Code Listings

Glossary

CSV	Comma-separated values
MUTISIMO	Multimodal and Multiparty Social Interactions Modelling
EAF	ELAN Annotation Format
Varied duration dataset	varied sampling for Opensimile dataset to merge with ELAN dataset
Fixed duration dataset	fixed sampling for Opensimile dataset to merge with ELAN dataset
ELAN CSV	CSV is generated from the EAF file
Opensimile CSV	CSV is generated from an audio file, and this CSV consists acoustic properties extracted by Opensimile.
Utterance type/event	Seven categorical values are present in the “concise merge type” column, including silence, mixture with spoken words and non-laughter vocalisation, spoken words, non-laughter vocalisation, discourse laughter, mirthful laughter and ambiguous type. We defined a term to refer to the value in this column.

The explanation of discriminating acoustic properties

Table 1: A Look-up table before the maIn text: The explanation of discriminating acoustic properties in this work

Acoustic property name	Simple explanation
F0	Fundamental frequency measures the lowest formant and preserves a specific person's voice footprint.
Jitter	The frequency oscillation in a circle is measured, and this temporal aspect property is quantified in seconds.
MFCC	Mel-frequency cepstral coefficients preserves the specific emphasised frequency in each frame of audio.
Auditory spectrum	This property measures the range of specific audio in the time domain.
Spectral Features	This property describes the general features of the OpenSimile in the time domain, and this temporal property is then transformed into the frequency domain.

1 Introduction

1.1 Background and motivation

Laughter is a social signal that has captivated the interest of philosophers and researchers for millennia and continues to be a subject of fascination(Ginzburg et al., 2020). Especially in today's competitive society and the economic downturn, laughter could somehow melt down misunderstandings and conflicts among people to some degree, as laughter, to some degree, implicitly conveys empathy and acknowledges others' opinions. Moreover, it can embolden timid individuals to forge ahead. Although laughter often occurs within interpersonal interactions, individuals may also laugh at themselves(Ludusan & Schuppler, 2022).

Conversation represents the primary context for laughter, wherein it assumes various roles. For instance, in clinical settings like laughter therapies, doctors' laughter can bolster the morale of cancer patients (Morishima et al., 2019). Similarly, in international conferences—such as those convened by the United Nations—laughter among prime ministers or presidents can foster diplomatic ties between nations. Beyond expressing joy, laughter also serves as a means of conveying and interpreting information. For example, when a student asks an embarrassing question during a lecture, a professor might respond with a chuckle to gently signal the inappropriate nature of the query, thus maintaining decorum.

In laughter research, laughter has many categorisations. One category is defined by the sound of laughter, encompassing giggles (Pietrowicz et al., 2019). Another category delves into the emotional aspects, distinguishing between mirthful and discourse laughter (Tanaka & Campbell, 2011, 2014). As the main purpose of laughter is to convey emotion (Gilmartin et al., 2013; Koutsombogera & Vogel, 2022), the latter category directly relates to peoples' emotions. Mirthful laughter emerges from authentic excitement from the heart, while discourse laughter could break down embarrassment or disguise authentic feelings to maintain etiquette to some extent.

As emotional state can be inferred from this laughter categorisation, distinguishing between these two types of laughter can sometimes be accomplished by the human auditory system or inferred from other elements such as facial expression. However, in environments with considerable noise, identifying the type of laughter becomes challenging for human listeners. In such cases, automated classification of these laughter types becomes necessary to analyse conversations and discern laughter with and without underlying emotions. Although laughter can be expressed through facial expressions(Sherman et al., 2012), facial cues may sometimes be concealed, such as with a smirk. On the other hand, listening to the sound of laughter is often more reliable, as the voice carries a unique signature that can help locate a person and discern their emotional state, which may not be easily disguised(Alluri & Vuppala, 2020). Thus, in this context, the primary focus lies on classifying acoustic laughter, even though other factors may also impact and correlate with acoustic laughter. Given this rationale, investigating the classification of acoustic laughter is

well-motivated.

Within this necessity of automatic acoustic laughter classification, the application of acoustic laughter classification between mirthful and discourse laughter has a potential market in industries. Distinguishing between mirthful and discourse laughter through acoustic classification holds potential applications, particularly in analysing the correlation between laughter in cinema and box office success across genres such as comedies and political thrillers. Filmmakers could strategically incorporate plot elements to evoke specific emotional responses from audiences based on laughter patterns, thus discerning which narrative elements are associated with different types of laughter. In international conferences, where cameras typically capture real-time discourse, laughter can be identified through video analysis, given the continuous stream of visual and audio data. By combining facial expressions and gestures captured in images with corresponding audio signals during laughter, specialised software and algorithms can be employed to differentiate between types of laughter emitted by presidents or officials from different states. For instance, if a leader's laughter aligns with unreasonable conditions or statements of other states, it may indicate collusion or deceit.

Diverse applications of automatic autistic acoustic laughter classification attract researchers to know about the internal structure of laughter (Ludusan & Wagner, 2019, 2022b; Tanaka & Campbell, 2011, 2014). Just as an amateur can distinguish between different musical instruments based on their sounds, some instruments, like the saxophone and clarinet, produce similar sounds that necessitate professional training and specialised equipment for accurate identification. Similarly, while human hearing can discern to some extent between different types of laughter, such as mirthful and discourse laughter, describing the precise distinctions between them proves challenging. The intricacies of the internal structure of these laughter types are so subtle that they often elude detection by the auditory system. For instance, the fundamental frequency, which measures voice pitch, is associated with the sound of laughter, as supported by several studies(Kipper & Todt, 2003; Mittal & Yegnanarayana, 2015; Szameitat et al., 2011; Tanaka & Campbell, 2011, 2014; Vettin & Todt, 2004).

Acoustic properties, similar like a "special timbre" in musical instruments, also need some device to extract and detect. Previous research utilised the Snack tool to extract acoustic properties from laughter, even though this feature set is relatively small(Tanaka & Campbell, 2014). A large feature set will have more acoustic properties, such as "Opensimile"(Eyben et al., 2010), a state-of-the-art feature extraction tool. "Opensimile" is an audio extract tool incorporating a 6,373 acoustic parameter set in "ComParE_2016 " collection(Eyben et al., 2015). Besides, none of the current work utilised "Opensimile" to extract acoustic properties and select some appropriate classifiers for the acoustic laughter classification task. To address this research gap, this project aims to utilise "Opensimile" as a feature extraction tool and employ an appropriate model on the relevant dataset.

Another challenge in implementing such a system is that the dataset contains relatively fewer laughter types, such as mirthful and discourse laughter. Additionally, the author of this research has opted not to release these types as annotations, recognising the significant craftsmanship required in accurately tagging laughter instances. The dataset requires continuous moments featuring various utterance events, including laughter, as well as other types such as silence and spoken words. Fortunately,Koutsombogera & Vogel (2018a) released an open dataset publicly, the MULTISIMO corpus. This dataset contains an EAF, an XML format, and an audio file. EAF contains a laughter tag and spoken word text, such as "OK, thanks for coming today [eh]," whereas the audio contains the recording related to each session. Based on this raw dataset, we could conduct further analysis to extract useful information from two types of files in the "MUTISIMO"corpus and

construct a dataset that aligns with our research question.

Based on this motivation, we proposed our research question as: "**Are there systematic differences in the acoustic properties of different acoustic laughter?**".

The semantics of each contextual term in our research question are as follows. The "acoustic laughter" is explained above, including mirthful and discourse laughter. Additionally, the phrase,"acoustic properties" is also interpreted in the above explanation. The only contextual term that has not been explained is "Systematic differences". Systematic differences in acoustic properties indicate that some properties have distinct values and contribute to predicting specific laughter; besides, systems different from the classifier perspective feature some properties in predicting specific laughter. Based on this fact, a decision tree could present the importance of features and visualise the branch selection to show the disclaiming feature. The discriminating feature in multinomial regression is also present in the coefficient in each response variable

Fortunately, our work is based on previous work's shoulders. As our project intends to identify systematic differences in acoustic properties in categories of laughter in the "MULTISIMO" dataset, previous work, Tanaka & Campbell (2014) have done similar work on their dataset towards mirthful and discourse laughter and proposes a complete framework from feature extraction and machine learning model selection to feature analysis. However, since they have not released their dataset and there are more appealing to explore the factors that impact laughter, this project intends to replicate their work and explore some phenomena they have not identified.

The significance of this work could be discussed in terms of theoretical contribution and potential practical contribution. As our work is based on algorithmic and machine learning models, particularly theoretical work, the main contribution of this work is to identify the discriminating properties in discourse and mirthful laughter. The findings of this work could inspire researchers in speech recognition and natural language processing to explore more interesting phenomena in conversation. In terms of practical implications, the methodologies developed in this study could be integrated into real-time video systems for various applications. For example, discriminating properties could be tracked in real-time, along with patient emotion variation and different laughs in psychological clinical treatment. Besides, it is worth noting that this study is based solely on the MUTLSIMO dataset. However, the algorithms developed here could be adapted to different laughter annotation systems, thereby enabling the exploration of discriminating properties across a broader range of acoustic laughter types.

Even though the work in this area sounds exciting and has many potential applications, implementing this system takes work. Hence, we split this huge amount of work into different components. The next section, "Research Objectives", will illustrate the road map to achieving this project's aim.

1.2 Research Objectives

The main objective of this research is to construct a comprehensive acoustic laughter classification pipeline that includes dataset construction, dataset verification, machine learning selection, identification of discriminating properties, and investigation of factors impacting discriminating properties.

Expect dataset construction and verification; the rest of the components in our dataset are similar to Tanaka and Campbell's (2014) work. For dataset construction, we used the "MULTSIMO" dataset to construct a dataset that fits our research question.

Overall, the objectives of the dissertation are as follows:

• Develop an algorithm to classify annotation in each session EAF to make a simple version of the symbol annotated in the original EAF(cf. Figure.1.1)) serve as the categorical value of the target variable, such as transforming “OK, thanks for coming today.[eh] we’re going to play a quiz” to simpler representation, such as “M”, a simple represents mixture type with speaking and non-laughter vocalisation(“[eh]”).

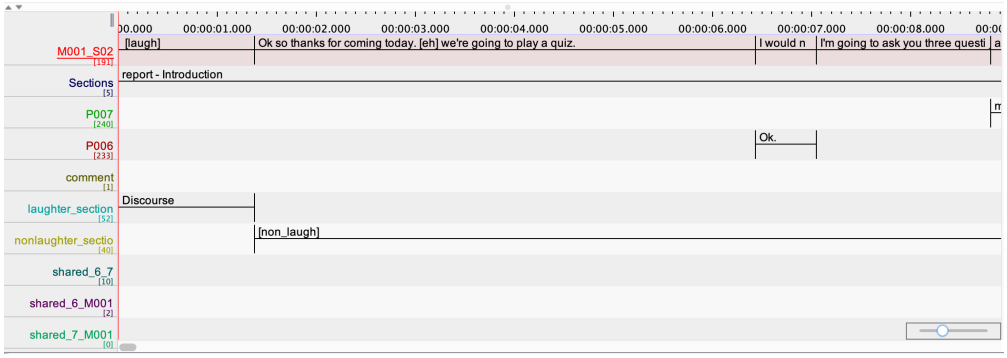


Figure 1.1: Session 2 information in EAF file visualisation by ELAN

• Develop a string symbolisation algorithm to detect the silence moment not annotated in EAF and mark it as “Silence” (see the sixth row in Figure 1.1, showing that after the “Discourse” annotation, there is a blank gap, and this gap is a silence moment).

• Construct continuous time moments from EAF from the ANNOTATION tag in the EAF file to streamline each session’s start-to-end moment. Use the above algorithm to classify raw annotation, detect silence simultaneously, and then store them in a CSV called “ELAN CSV.”

• Construct a CSV containing acoustic properties called “Opensimile CSV”.

To reach this objective, we need to utilise the Python segmentation library to cut each session audio file, ranging from 5 min to 10 min, into multiple audio pieces based on a fixed duration (200 ms) and a variable duration (the same duration as each moment in EAF per session). Then, we need to use the Python Opensimile API to extract acoustic properties from each audio file and store each moment with acoustic properties in the same row.

• Merge two types of CSVs, including ELAN CSV and Opensimile CSV, to construct the final version of CSV, which will include one fixed-duration version and a varied-duration version.

For the varied duration version final dataset, we need to align the Opensimile CSV each moment to align the moment in the EAF one by one as each two CSV sunrise start time and end time in each moment; for the fixed duration dataset, these two CSVs does not synchronise in start and end times. Designing an algorithm to align both CSV systems is necessary for this case.

• Apply various data verification techniques to assess the constructed CSV.

• Employ machine learning classifiers to identify the discriminating properties in each version of the dataset (fixed and varied duration dataset) in different laughter (mirthful and discourse laughter) in the diverse models (decision tree and multinomial logistic regression).

• Conduct regression analysis to identify the significant properties to explain the variance of the target variable and adopt non-parametric hypothesis testing to identify the discriminating property

correlation.

• Adopt quantitative and qualitative assessments to evaluate our work, including quantitative and qualitative approaches.

The quantitative approach uses classification accuracy and the Cohen kappa coefficient to measure machine learning model performance to assess the selected models, while the qualitative approach evaluates factors that impact discriminating property to support or falsify the claim from related literature.

1.3 Dissertation structure

The dissertation is organised in the following manner:

- Chapter 1—Introduction: This chapter constructs the research territory, signifies the significance of our research within the research scope, and outlines the research question and contribution to this work.
- Chapter 2—Research Background and Literature Review:This chapter briefly introduces the key techniques in the acoustic laughter classification and elaborates on Tanaka & Campbell (2014)'s analysis of this work. Additionally, retrospect another laughter system in the "MULTSIMO"dataset to identify research direction in this area.
- Chapter 3—Methodology:This chapter describes the dataset construction and verification, as well as the consideration of machine learning models and the design of experiments.
- Chapter 4—Results:This chapter describes the parameter setting in each machine learning model and presents the results from the designed research methodology.
- Chapter 5—Evaluation:This chapter utilises quantitative and qualitative assessment to evaluate our results from the experiment chapter.
- Chapter 6—Conclusion:This chapter summarises the dissertation, discusses the potential drawbacks and lists future directions.

2 Research Background and Literature Review

This chapter furnishes the foundational knowledge necessary for comprehending the research question addressed in this dissertation by reviewing related research in the field. Accordingly, it begins by introducing techniques in acoustic laughter classification and provides a detailed analysis of our replicated work, drawing from Tanaka & Campbell (2014). Following this, it presents previous laughter systems outlined in undergraduate final-year reports or master's dissertations, as these systems typically entail a more extensive amount of work compared to peer-reviewed articles. Finally, resources utilised in this dissertation, such as corpora and feature extraction tools, are introduced.

2.1 Acoustic Laughter classification and our replicated work

Laughter serves as a non-verbal language and social signal, attracting researchers to inspect this factor to identify the key factors impacting laughter. Acoustic laughter classification is one of the areas to explore in the task. This research focuses on identifying acoustic laughter and exploring existing work on the classification of acoustic features can provide valuable insights for our study. Tanaka & Campbell (2014) present a comprehensive framework, spanning from feature extraction to machine learning model selection to feature analysis. Emulating their methodology not only guides the research but also empowers us to develop a new dataset tailored to our specific objectives, potentially uncovering insights beyond those elucidated by Tanaka & Campbell (2014). This replication effort is motivated by the desire to extend the existing pipeline in this specialised research area and to identify and address any limitations encountered, thereby enhancing the robustness of future endeavours.

2.1.1 Key techniques in laughter classification

Acoustic laughter classification involves several key steps: feature extraction, model selection, identification of determinant factors, and feature analysis. The cornerstone of this pipeline is determining the crucial factors that impact laughter classification, which could be acoustic features or internal structures of laughter like bouts and duration. Among the tools available, the Snack toolkit and Opensmile(Eyben et al., 2010) have gained widespread acceptance for feature extraction in prior research. For model selection, various approaches are utilised to classify different laughter types, including decision tree models (Tanaka & Campbell, 2014), principal component analysis (Tanaka & Campbell, 2014), neural networks(Knox & Mirghafori, 2007), and others. In the feature analysis phase, both parametric statistical tests such as the Student's t-test (Tanaka & Campbell, 2011, 2014) and non-parametric tests like the Wilcoxon rank test (Maggie, 2021) are employed in this research domain.

In terms of discriminant factors, common ones are listed in the table below (cf. Table 2.1). We elucidate the purpose, measurement, and correlation of each feature with acoustic laughter classification. The concept of formants is particularly intricate compared to the others; therefore, priority is given to explaining this concept.

Table 2.1: Common discriminant factors

Determinant factor	Paper
Formant	Tanaka & Campbell (2011, 2014)
Root mean energy	(Ludusan & Wagner, 2022a)
Bout	Tanaka & Campbell (2011, 2014)

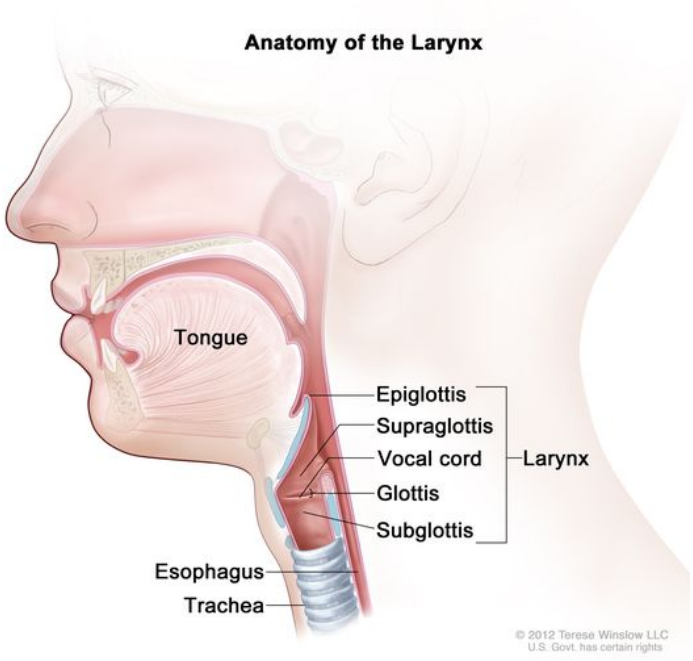


Figure 2.1: Vocal cord sample from the website of national cancer institute

The formant is the vocal fingerprint, created by the resonance of airflow and the human vocal cords at a specific speed, or more simply, it represents the pitch of voice (cf. Figure.2.1 from national cancer institute website(https://www.cancer.gov/publications/dictionaries/cancer-terms/def/vocal-cord)). The fundamental frequency is the pitch of the voice, which could be metaphorised as the basic note(the first harmonic peak in the acoustic spectrum) of a musical instrument. The other formats (f1,f2,f3,f4) delineate the vocal contour, shaped by variations in the mouth and throat, and reveal distinctive nuances within the voice that the fundamental frequency alone cannot convey.

Several research studies substantiate that formants correlate with vowels and constant laughter(Tanaka & Campbell, 2011, 2014; Trouvain & Schröder, 2004), such as “/a/”. This acoustic contour assists in eliciting different types of laughter based on formants. For example, mirthful laughter is produced by authentic emotional engagement. The vibration of this type of laughter is stronger than its discourse laughter counterpart as the voice of this type of laughter is lower pitch.

Understanding the concept of formants requires some phonetic knowledge, as described above. The remaining concepts are more straightforward. Root mean square (RMS) energy measures the intensity of voice, and different types of laughter can be differentiated based on this feature, which is recognisable by the human auditory system. On the other hand, Bout measures the duration of a specific type of continuous interval, such as laughter in a conversation. This duration of laughter could reflect the airflow length in the mouth and throat. Due to the emotional enjoyment conveyed by mirthful laughter, the mouth tends to be more open compared to discourse laughter.

2.1.2 Elaboration of Tanaka and Campbell (2014)'s work

Tanaka & Campbell (2014) aimed to develop a sensor module that could be applied in presenting different types of laughter (polite versus mirthful) in a video clip rendered by computer graphics techniques. To achieve this, the audio signal from the natural dialogue served as the primary source for conducting two experiments. These experiments aimed to investigate the correlation between ten volunteers, potentially from diverse backgrounds such as different countries or genders, and specific laughter types. Specifically, the first corpora in their Experiment were "Expressive Speech Processing"(ESP)[1] .

The dataset contained two aspects of information regarding ten participants, including nationality (6 Japanese, 2 Chinese, and 2 Americans) and gender (5 female and 5 male). Specifically, the first corpora in their experiment was the Expressive Speech Processing. The dataset contained two aspects of information regarding ten participants, including nationality(6 Japanese,2 Chinese,2 American) and gender(5 female and 5 male). To select a representative sample of this corpora, they conducted an ablation study focused on different native language populations. This included Japanese male speakers aged 20 who participated in experiment A, Japanese male speakers aged 20 who participated in experiment B, English male speakers aged 20 who participated in experiment A, English female speakers aged 20 who participated in experiment A, Chinese female speakers aged 30 who participated in experiment A, and Chinese male speakers aged 20 who participated in experiment A. Comparing it with the first dataset, which contained audio information from ten volunteers, the second dataset, FAN(ages), was contained within ESP. This dataset only focused on one volunteer, a young Japanese female.

With the explanation of these two datasets, Japanese students were the sample to verify the laughter types, and this person only presented polite social laughter instead of loud laughter. The second dataset aimed to establish the connection between speakers' acoustic features and different categories of laughter. Based on the explanation of the datasets, they conducted the first experiment and recruited 20 Japanese students to reflect the most recognised laughter by Japanese students.

The findings of this experiment revealed that "mirthful" and "polite" laughter were the predominant categories. Hence, they selected these two representative laughter to conduct the experiment 2. The second experiment aimed to extract the relevant acoustic parameters and analyse the relevance between these features and the two types of laughter they finally decided in experiment 1. Not only did they select fundamental prosodic parameters, such as F0, but they also introduced additional parameters, including spectral tilt, shape parameters and positional parameters, to boost audio quality and to encode the laughter acoustic dynamics. They also computed these parameters'mean, maximum, and minimum values as the acoustic features.

Then, they applied principal component analysis (PCA) to reduce the dimensionality of the data. Their results revealed that, among other features, fundamental frequency,spectral slope,power, and F0moveAB

[1] https://www.speech-data.jp/

were the main components for all speakers. In addition to PCA, they employed decision trees to conduct more refined filtering on three parameters: fundamental frequency, power, and two parameters derived from the PCA analysis, min-max and min.

Based on the decision tree, these features were influential, including mean,pmax, pct and dn. However, the classification accuracy of this model was relatively low in specific groups of people(FAN). High dimensional acoustic parameters might lead to overfitting. To address this, the researchers considered both contributing features from principal component analysis and the decision tree classifier. Eventually, they identified seven important acoustic features, including “fmean,ppct, pmax,h1a3,duration, No.Call and F0moveAB”.

Subsequently, they employed a support vector machine (SVM) to predict category based on these acoustic features. However, further analysis revealed no statistically significant difference between speaker-independent outcomes and results after reducing dimensionality. As a result, they conducted an error analysis using the student-t test, focusing specifically on tokens of polite laughter from two speakers: an English male speaker and a Chinese female speaker. Their findings suggested that errors in classification might be attributed to noise within the audio files.

2.1.3 Insight from Tanaka and Campbell(2014)’s work

Tanaka & Campbell (2014) work uses a decision tree to conduct binary labelling tasks, discourse, and mirthful laughter classification by showing feature selection having ten leaves. While their visualisation allows for the identification of discriminating acoustic properties, it lacks clarity regarding the importance of features for nodes or branches at the same level, as it only provides the hierarchy rank from the root to the node. Therefore, it is preferable to present a ranked list of feature importance generated by the decision tree, as it clarifies the level of importance. Furthermore, they categorised five utterance types: non-laughter vocalisation, discourse laughter, mirthful laughter, derisive laughter, and other types.

However, their presentation only focuses on binary classification tasks concerning discourse laughter and mirthful laughter. In this setup, the reduction of negative samples from prediction and classification might undermine the trustworthiness of their results to some extent. To address this limitation, expanding the consideration to other utterance types and introducing more samples would enhance the rigor of the process.

Simultaneously, the multi-labelling task must also be taken into account. Unlike binary classification, where there is no competition among target response labels, multi-classification involves such competition. Incorporating this experiment will enhance the rigor of this study.

In addition to employing a decision tree for identifying discriminating properties, they also employed principal component analysis (PCA). However, there is an issue within this model selection. Even though PCA is able to interpret the most discriminating feature, the rest of the features are not adequate. Therefore, it is better to consider another statistical model that can integrate all discerning features. In the testing phase, acoustic properties were analysed using a student t-test. However, the normality of each property was not confirmed beforehand. Given that the hypothesis of the student t-test relies on parametric assumptions, it is recommended to conduct tests for normality prior to selecting types of statistical hypotheses.

Based on the above analysis and constructive critique towards Tanaka & Campbell (2014)’s work, this research provides guidance for identifying acoustic processes, spanning from feature extraction to machine learning model selection and statistical hypothesis testing. In comparison to existing studies in this domain, it presents a more lucid framework for conducting research, thereby motivating the replication of this study

using the new MULTISIMO dataset(Koutsombogera & Vogel, 2018a).

2.2 Previous laughter system developed on MUTISIMO dataset for laughter classification

By reviewing previous system concerning laughter classification is work in the undergraduate final year project and master dissertation (cf. Table.2.2), we intend to identity the research direction from these works as they provided a relatively compete sample. Laughter systems developed on the MUTISIMO dataset mainly inspect aspects related to laughter.

Table 2.2: Previous laughter system in the final year project or dissertation

Paper	Dataset	Research object	Address issue
Mohan (2019)	MULT-SIMIO	Conversational dominance and laughter	Investigate the relationship between participant dominant and two laughter including mirthful and discourse laughter
Hegarty (2022)	MULT-SIMIO	The social laughter in the conversion and OCEAN personality	Investigation of the effects of ratified/ratyting/solo laughter and an association between the OCEAN personality model and natural/social laughter events

2.2.1 The research findings of Mohan

Mohan (2019) intended to investigate the relationship between the level of conversational dominance measured by five annotators and two types of laughter, discourse and mirthful laughter in two components of an experiment. The dominance score ranges from 1 to 4, indicating that the higher the mark, the more willing to dominate the conversation.

In first experiment, a dominance score was utilised to assess participant involvement. More concretely, Specifically, the total and average number of two types of laughter (mirthful and discourse laughter) were analysed using the Kruskal-Wallis test (for median) and the Wilcoxon Rank Sum test (for mean). This analysis aimed to identify statistical patterns, supporting the hypothesis that low dominant individuals tend to participate more in discourse laughter compared to mirthful laughter. Conversely, individuals with high dominant scores were more likely to express mirthful laughter than their low dominant score counterparts.

In the second experiment, the focus was on understanding the relationship between the two types of laughter and their frequency of occurrence in the final quarter of each dialogue session. Using the same statistical techniques as the first experiment, the researchers formulated a hypothesis that discourse laughter frequency increases in the final quarter of each session, while mirthful laughter occurs randomly throughout.

Reflecting on these findings, several insights emerge. The first experiment sheds light on the role of laughter in facilitating conversation dynamics, while the second experiment delves into the temporal aspects of laughter. Given our research interest in identifying acoustic properties in laughter, further investigation could

explore the correlation between specific acoustic processes in mirthful or discourse laughter and domain scores to understand how internal structures impact conversational dominance. Moreover, delving into time series analysis of different acoustic properties in the two types of laughter could yield valuable insights. This approach would deepen our understanding of the relationship between temporal aspects and laughter expression, enriching our exploration of the intricate dynamics of human communication.

2.2.2 The research findings of Hegarty's work

Hegarty (2022) investigated the effects of social and natural laughter and an association between the OCEAN (openness, conscientiousness, extraversion, Agreeableness and Neuroticism) personality model and specific laughter events, such as mirthful laughter on the MULTISIMO corpus (Koutsombogera & Vogel, 2018b). Social laughter in their work involves solo, ratified, and ratifying laughter, while natural laughter in this work includes mirthful and discourse laughter. Following data processing, modelling and evaluation, the author this work concluded that mirthful laughter is more likely to be found by individuals who dominate conversations, resulting in them being more likely to initiate laughter followed by other participants. Regarding the personality model findings, it was observed that a ratified laughter leader is more likely to be noticed among conscientious individuals.

As social laughter is not annotated in the MULTSIMO dataset, they devised an algorithm to identify this type of laughter and construct a dataset suitable for their research objective. Dataset construction is equally crucial in our research, as we aim to capture each continuous moment's utterance event. This algorithm could potentially offer insights to guide us in constructing our dataset.

Specifically, they utilised an additional column to record the social laughter information for each moment, employing a half-second threshold to discern ratified laughter. Ratified laughter was determined if the time difference between the end of the previous laughter and the start of the current laughter was less than or equal to this threshold. Laughter failing to meet this condition was categorised as solo laughter. Due to the complexity of laughter annotation, two cases were examined. In the first case, where solo laughter was ratified, it needed to be reclassified as ratified laughter. In the second case, where laughter ratified another ratified laughter, the previous laughter retained its original status.

Furthermore, they assessed whether the time difference between the start of the initial instance of laughter and the end of successive laughter fell within the threshold to determine the ratified status of the first laughter instance. In addition to the median dominance score derived from the MUTISIMO dataset, extra columns were introduced to include scores of OCEAN personality traits obtained from participant self-reflection parchments and collated local percentiles of the original test dominance score. These additional columns were incorporated to facilitate statistical analysis.

For dataset verification, they examined participant codes to ensure that instances where the same volunteer marked two consecutive laughter occurrences as "ratified" were flagged. In such cases, these paired laughter instances were reclassified as individual instances of solo laughter. Additionally, they accounted for another scenario where previous verification procedures might have missed cases where two successive laughter instances reciprocated the preceding laughter. To rectify this, multiple conditional statements were implemented. Besides automated verification, manual testing involved listening to the audio to cross-reference with the laughter annotations, serving as another layer of verification to ensure accuracy.

Through their meticulous dataset construction and verification process, our work must carefully choose the

appropriate dataset structure and algorithm to extract pertinent information from the MUTISIMO dataset. This involves conducting various automatic and manual verifications to guarantee the correctness of our dataset.

To conduct their experiment, the researchers proposed several statistical hypotheses regarding the influence of conversation dominance and personality traits on various instances of laughter. They utilized an R script to examine the interaction of laughter tendencies with personality and dominance scores. In the analysis of laughter within the social context, they found a higher occurrence of solo discourse and communal laughter with mirthful characteristics, along with a lower incidence of solo mirthful laughter and communal laughter than anticipated, as determined by the chi-square test.

They conducted a Wilcoxon rank sum test on Big Five OCEAN personality scores by testing the personality traits and specific laughter instances, such as mirthful or discourse laughter. This result showed that individuals with a relatively high scores in openness and neuroticism were more likely to exhibit mirthful laughter, whereas those with lower scores in these traits tended to display discourse laughter. Subsequently, they investigated the correlation between the five personality traits and the social dimension of laughter using the Wilcoxon rank sum test. However, this analysis yielded little evidence of a significant relationship between OCEAN traits and specific types of laughter. In further exploration, they conducted additional testing involving all personality traits. This revealed that individuals with higher agreeableness scores were more likely to influence the duration of laughter, with higher agreeableness scores associated with shorter laughter duration.

To examine the impact of the laughter leader (a participant who initiates laughing), they conducted several supplementary experiments. These experiments suggested that individuals with longer laughter durations within a group, relatively higher conscientiousness scores, and higher dominance scores were more likely to assume the role of a laughter leader. Although the analysis of laughter leaders did not demonstrate statistical significance regarding the association between dominance and the leadership role in laughter, an alternative hypothesis from the literature was explored. According to this hypothesis, a laughter leader should not merely follow others' laughter. Based on this premise, it was found that individuals who initiated and ratified their laughter were more likely to have higher dominance scores compared to those who only ratified others' laughter or laughed independently.

This research involves constructing continuous time intervals incorporating natural laughter, social laughter, and annotations for the Big Five personality traits during dataset construction. Prior to the experiment, various verification were undertaken to ensure dataset accuracy. During the experiment phase, a series of statistical tests were conducted to assess the impact of psychosocial factors on laughter. Overall, the study offers several insights and avenues for further exploration. The researchers utilised the Big Five inventory to gauge the correlation between personality traits and social/mirthful laughter. In contemporary contexts, the Myers-Briggs Type Indicator (MBTI) is often favored for career path assessment. This personality assessment could also be viewed through the lens of MBTI, which can be translated into Big Five personality traits. Such correlation investigations hold potential relevance in real-world scenarios. Additionally, since they conducted a continuous time moment dataset, a time series plot in terms of social laughter could be considered, as this visualisation gives a more direct sense of ratified and rating laughter.

2.2.3 Research direction from these two systems

From the two aforementioned research studies(Hegarty, 2022; Mohan, 2019), exploring another dimension of laughter influence in conversation necessitates the creation of a cross-classification annotation dataset.

However, constructing such a dataset is more labour-intensive compared to other conversation datasets due to the need for manual annotation and intricate algorithmic design. This scarcity is not unique to the MULTSIMO corpus but extends to various fine-grained dialogue datasets within dialogue systems. Examples include datasets containing multiple sentences expressing preferences towards both the user and the dialogue agent(Boyd et al., 2020; Eric et al., 2019; Ma et al., 2021; Zhang et al., 2018).

Moreover, the quality of results is significantly influenced by the adoption of appropriate statistical testing methods. In both prior systems, the Wilcoxon rank test was employed, acknowledging that the distribution of laughter is often non-normally distributed in most cases.

2.3 Corpus and feature extraction tool

In this section, we introduce the resource to assist our research :corpus and feature extraction tool.

2.3.1 MULTSIMO corpus

Laughter is a social signal and a catalyst to foster dialogue dynamics. The MUTISIMO corpus (Multimodal and Multiparty Social Interactions Modelling) was developed to analyse collaboration and task success within groups (Koutsombogera & Vogel, 2018a). Laughter instances are meticulously annotated in the MUTISIMO dataset, with laughter text tags annotated in EAF (XML format), alongside acoustic laughter information embedded within the audio files.

This dataset incorporates eighteen conversation sessions, comprising either audio or video recordings, with each session lasting approximately five to ten minutes. The structure of each session typically involves participants seated on either side of a table, responding to three questions posed by a facilitator. Each participant answers three questions and then ranks them based on popularity, addressing a question posed by the facilitator. In this process, participants could collaborate with different verbal or non-verbal signals, such as laughter.

Besides, each participant completed a Big Five personality test before the session and an experience survey after each session. This valuable data can be analyzed in conjunction with laughter occurrences to explore potential correlations between laughter and personality traits. Additionally, the dataset contains a wealth of supplementary information such as gaze behavior, hand gestures, etc., providing further insights into the factors influencing laughter dynamics.

2.3.2 OpenSimile feature extraction tool

This subsection will introduce the feature extraction tool, Opensimile in this dissertation and how this tool extracts acoustic features and what acoustic parameters this tool incorporates (cf. Table.2.3).

Table 2.3: Literature related to “Opensimile”

Paper	Address issue
Eyben et al. (2010)	How Opensimile sample acoustic features from audio file
Eyben et al. (2015)	How acoustic parameter work in the Opensimile

Eyben et al. (2010) describes how the Opensimile is being used for acoustic feature sampling. “Opensimile” is a ready-to-use and general low-level acoustic feature extraction tool, and this software utilises a single configuration feature file to perform large-scale feature extraction(Eyben et al., 2010). For the architecture of “Opensimile” , this software utilised C++ as programming laughter to perform at a fast speed when visiting memory. It employed a ring buffer memory structure to optimise space usage. Opensimile is a ready-to-use and general acoustic feature extraction tool that utilises a single configuration feature file for large-scale feature extraction. The power of Opensimile is armed with several low-level acoustic descriptors, such as Mel-Frequencey Cepstral Cofficent, pitch, and multiple functional features that relate to each acoustic feature, such as segments and percentile.

Eyben et al. (2015) identified the subset of allowable parameters crucial for acoustic parameter extraction, offering a pathway towards understanding diverse emotional states. Previous endeavors relied on machine learning models to generate expansive parameter sets, impeding generalisation. Building upon this insight,they advocated for a streamlined set easily accessible online. The selection criteria for these parameters were threefold: indexability of psychological changes, continuity with prior research, and theoretical significance. The minimalist parameter set encompasses three groups. The first group, frequency-related parameters, includes "pitch, jitter, formant frequency (formants 1, 2, and 3), and first formant bandwidth." Energy-related parameters constitute the second group, encompassing "shimmer, loudness, and harmonics-to-noise ratio." The final group, spectral parameters, integrates "alpha ratio, Hammarberg index, spectral slope, formant 1, 2, 3 relative energy, and Harmonic difference H1-H2, H1-A3."

Furthermore, low-level acoustic descriptors underwent smoothing techniques, employing arithmetic mean and coefficient of variation (e.g., standard deviation). These processed descriptors, termed "Functionals," were applied to 18 low-level descriptors, incorporating spectral and frequency parameters documented extensively in prior literature. Arithmetic mean and coefficient of variation were also applied to these extended sets. Evaluation revealed that Eyben et al.’s model for acoustic parameter extraction from audio waveforms outperformed five benchmark datasets in terms of size, which were limited to binary classification.

This research used Python programming language to implement our system and selected Opensimle python API to extract acoustic feature extraction[2].

[2] https://audeering.github.io/opensmile-python/

3 Methodology

3.1 Methodology Introduction

The research delineates the roadmap of our study along with the research methods employed, focusing primarily on innovating dataset construction, dataset correctness verification, and dataset analysis. Specifically, dataset construction stands out as pivotal within our research methodology. These methods incorporate various data structures, including linked lists and stacks, as well as diverse algorithms, such as time interval alignment and text symbolisation algorithms.

3.1.1 Research design

In this project, we meticulously examined various facets of research design to ensure the robustness and reliability of our methods. Our research approach encompasses dataset construction methods, the accuracy of dataset construction, and dataset analysis techniques. Given the central role of data in each of these methods, our approach to research design is fundamentally rooted in a data-centric perspective. The components of our research design primarily encompass research philosophy, research approach, and research methods.

Research philosophy This research adopts positivism as all design and implementation hinge upon construction and analysis rooted in empirical study. It provides a systematic framework for gathering objective data and drawing verifiable conclusions, aligning with the study's empirical focus and aim for rigorous analysis(Park et al., 2020).

Research approach The research approach contains inductive and deductive approaches. In our study, we opted for the deductive approach as it was necessary to derive knowledge from the phenomena observed in the experiment, thereby indicating that our results are derived through a bottom-up sequence.

Research methods Lastly, the research philosophy and approach significantly influence the direction of the research methods. In this study, we have opted for a quantitative research method. Building upon the aforementioned research design, we have formulated our research methods, which include both data construction and analysis techniques. These methodologies have been tailored to align with the quantitative research approach chosen for this study.

3.1.2 Structure of chapter 3

The methodology's structure comprises a cascade pipeline that includes dataset construction methods, ensuring the correctness of dataset construction, and employing data analysis methods. The correctness of

dataset construction involves employing various verification techniques to ensure the dataset fed into the machine model is 100% accurate. Furthermore, the dataset analysis methods explore the internal structure of the dataset and select suitable models for identifying discriminating acoustic properties to address our research question. The structure of this chapter is illustrated in the diagram below(cf. Figure.3.1).

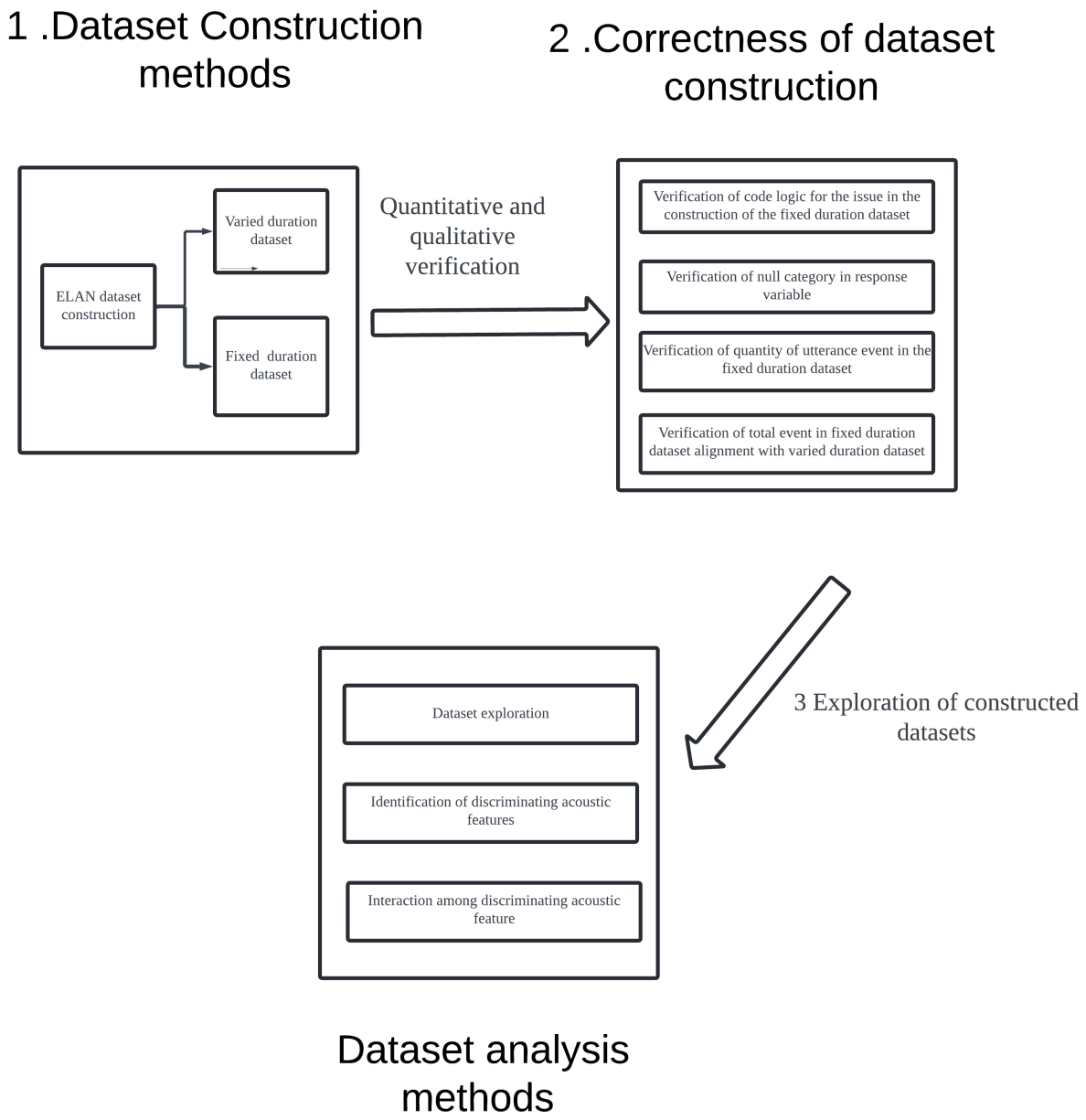


Figure 3.1: The structure of the methodology chapter

3.2 Connection between research design and research methods

As research design operates on a high-level perspective, research methods serve as instances of research design. In our study, aimed at identifying low-level acoustic features of two types of acoustic laughter, data analysis relies on algorithmic generation. We employ machine learning and statistical tools to examine the

properties of the data, aligning with the fundamental tenets of positivism in research. This approach reflects the sequential progression of research methodologies and primary methods within a quantitative research strategy.

3.3 Data construction methods

In data construction methods, we introduced how our dataset in this study was generated from the raw MULTISIMO dataset.

3.3.1 Motivation and approximate replication of Tanaka and Campbell (2014)

Tanaka & Campbell (2014) work utilised sensors to gather dialogue session recordings and categorised laughter into discourse and mirthful laughter. This dichotomy reflects the functional aspects of laughter. They employed principal component analysis and a decision tree to identify disorienting features extracted by the "Snack speech processing Toolkit." Their research has been integrated into clinical therapy and facial expression detection software, with a significant focus on identifying low-level acoustic properties.

However, they did not release their dataset, hindering replication and further exploration. To replicate and potentially build upon Tanaka & Campbell (2014)'s work, researchers require recorded conversations annotated with laughter and other dialogue elements, such as silence, spoken words, and non-laughter vocalisations. Fortunately, Koutsombogera & Vogel (2018a) addressed this gap by publishing the MULTSIMO dataset, a multimodal dialogue dataset containing various paralinguistic annotations, including laughter, publicly. Leveraging this dataset aids in replicating Tanaka & Campbell (2014)'s findings and uncovering additional phenomena.

To replicate their work effectively, it is necessary to extract the relevant paralinguistic elements from the MULTSIMO dataset. This dataset comprises multiple XML files recording continuous time and associated text information, along with several audio files capturing the voices of different participants. Given the scale and complexity of the dataset, developing an algorithm to construct the required dataset automatically is more efficient than manual extraction session by session.

An explanation of the content related to our project within the MUTISIMO dataset will determine the processing approach selection for this corpus. The raw dataset originated from 18 session EAF files and 54 audio files(three participants per session) in the MUTISIMO dataset. Each session comprised three audio files corresponding to three speakers. The EAF files were structured in XML format, necessitating the utilisation of a Python library to parse and extract the relevant tiers. Regarding the audio files, within each mono folder were three distinct files: one for the moderator, one for the left participant, and one for the right participant. Each audio file was engineered to amplify the current speaker's voice while attenuating the other speakers' voices.

The motivation behind devising an algorithm for constructing continuous moments stems from the dissatisfaction with the results obtained through manual manipulation using the export function (refer to Figure 3.2). The representation in 'CV-merge-M-L-S' cannot represent the mixed utterance between spoken words and non-laughter vocalisation. Besides, manual export might introduce more errors. For instance, we aim to avoid situations such as the incomplete utterance "[v". Ideally, both left and right brackets should be present simultaneously.

Begin Time - msec	End Time - msec	Duration - msec	CV-merge-M-L-S
0	1375	1375	[V] Discourse
1375	6441	5066	V.
6441	7057	616	V
7057	8812	1755	V
8812	9380	568	V
9380	13459	4079	V
13459	18149	4690	V.
18149	18620	471	[V]
18620	21375	2755	V
21375	23275	1900	V.
23275	29931	6656	V
29931	33993	4062	V
33993	35566	1573	V
35566	37865	2299	[V
37865	38678	813	V
38678	39850	1172	V?
39850	40514	664	V
40514	41110	596	V?
41110	42010	900	V?
42010	42745	735	V.
42745	44627	1882	V?
44627	45255	628	SILENCE
45255	50697	5442	V
50697	51572	875	V
51572	53186	1614	V?
53186	53716	530	SILENCE
53716	54280	564	V.
54280	56651	2371	V
56651	62427	5776	[V]

Figure 3.2: The continuous moments for the moderator tier in session two generated by the ELAN export function

Therefore, our primary task is to design an algorithm to generate continuous moments from the original EAF file. The author of this dissertation intended to employ two diagrams to illustrate our process in a high-level overview (refer to Figure 3.3 and Figure.3.4).

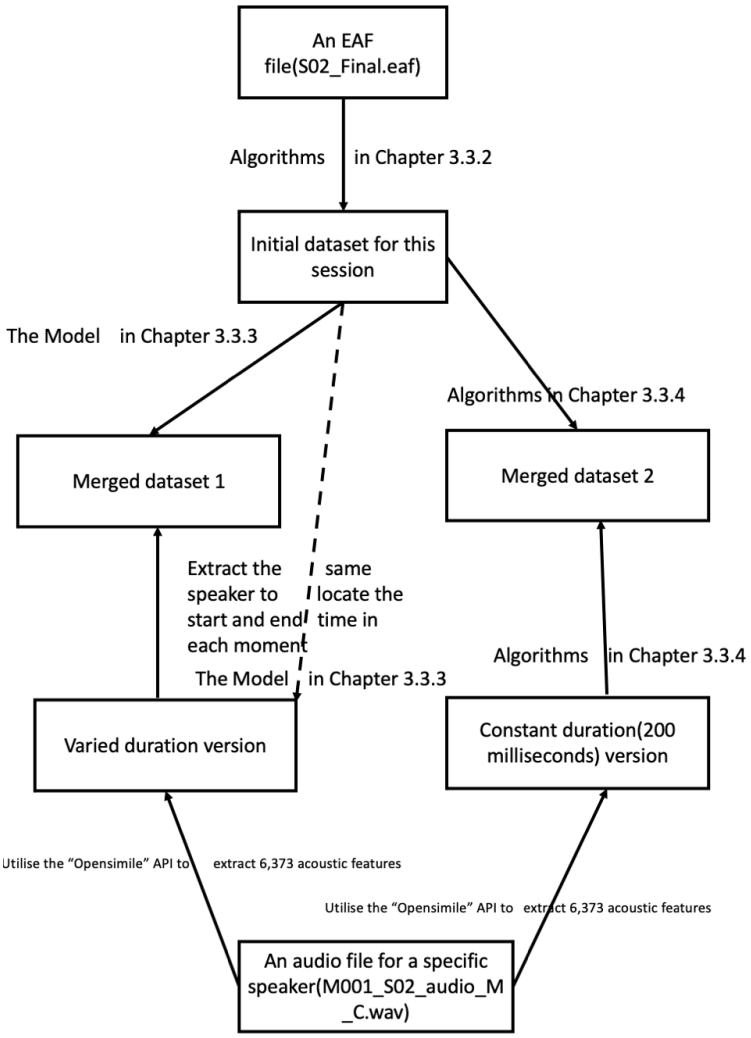


Figure 3.3: Overall Dataset construction diagram

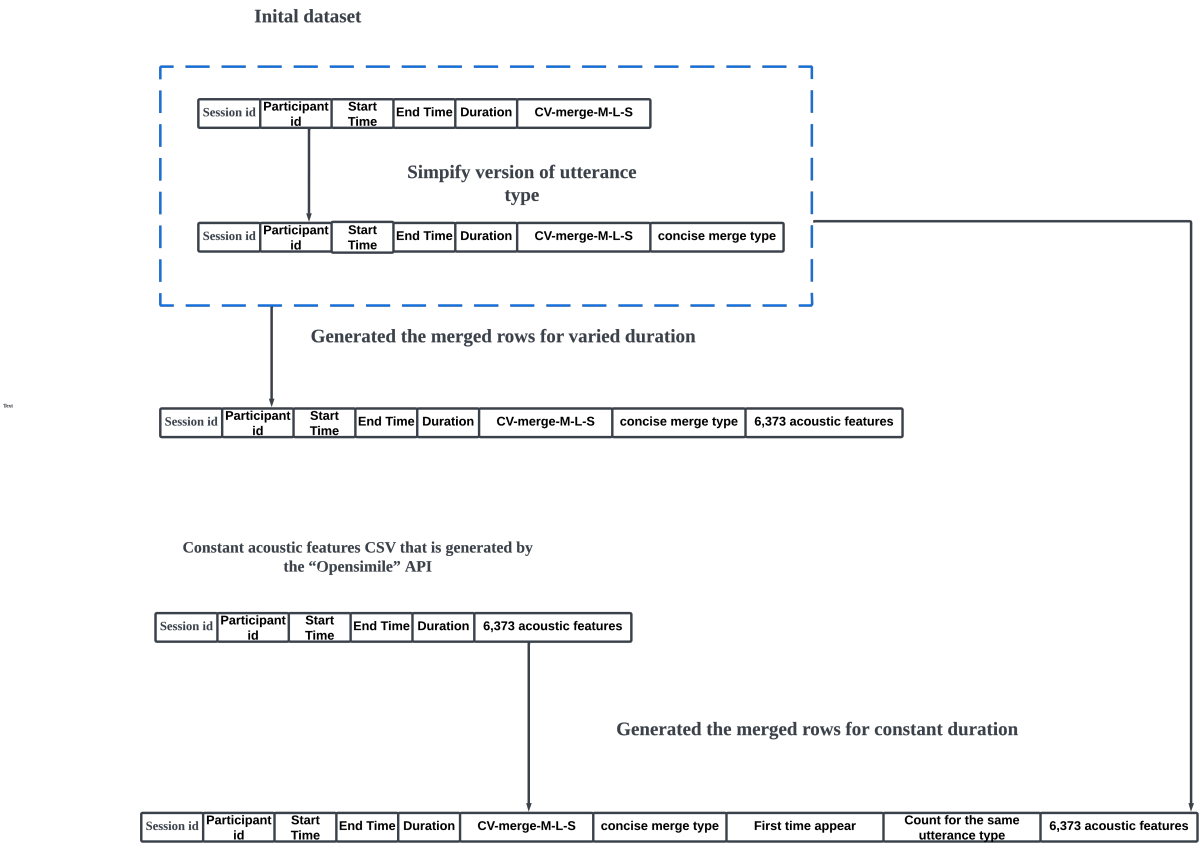


Figure 3.4: The dynamics of overhead variation diagram

Our model starts by employing our algorithms, as outlined in the section below, to construct the initial dataset. The initial dataset contains two phases: the before-simplified and simplified versions, highlighted in blue in Figure 3.4. The overhead of this dataset before simplified processing is depicted in the first rectangle within a blue contour in Figure 3.4. It's important to note that the subsequent section elucidates how the initial overhead without simplified processing transformed into the second overhead. We refer to this system, generated by our classification algorithm, as System 1. Furthermore, the format of this initial dataset is CSV, and this process simulates the ELAN export function to a certain extent. Later on, we also dub this initial dataset "ELAN CSV". Once the initial dataset is constructed, there are two methods to merge the datasets by two systems: one system generates an EAF file, while the second system is derived from an audio file and subsequently processed using the Python Opensimile API to extract 6,373 acoustic features.

In one approach, we utilise a CSV extracted by Opensimile on each audio clip with the same start and end times in each moment per session as “ELAN CSV” to align the initial dataset (see the model in Chapter 3.3.2 in the figure.3.3); the overhead of this merged version is represented by the third rectangle in Figure 3.4.

In another merging approach, we utilise a constant duration. Previous literature suggests that selecting a minimum duration range from 120 to 170 milliseconds is appropriate for auditory perception(Efron, 1970). Therefore, based on this suggestion, we choose 200 milliseconds as this value is not significantly different from the boundary of auditory perception that human beings can discern. Using 200 milliseconds as the constant duration, we extract the same duration for 18 sessions per speaker; the overhead of this operation is depicted in the fourth rectangle in Figure 3.4. Lastly, we combine the CSV generated by OpenSmile

(referred to as “Opensimile CSV” later on) with the “ELAN CSV” by applying the algorithms in the following section to generate the final constant merged CSV. The overhead of this operation is depicted in the last rectangle in Figure 3.4.

3.3.2 Initial dataset construction

The construction of the initial dataset involves receiving each session’s EAF and transforming it into continuous moments in the format provided below (cf.Figure.3.5).

Figure 3.5: Initial dataset overhead

In the sample provided above, the session ID represents an ordinal number assigned to sessions, while the participant ID indicates the participant number, typically formatted as ’P002’. Start time and end time denote specific moments in milliseconds, with the duration calculated as the difference between the end and start times. The ’CV-Merge-M-L-S’ column retains the original utterance labeling, such as ’[V] V [V] V [V] [V]’, necessitating processing of the initial labeling into a simpler representation. Finally, applying a straightforward algorithm processes the initial labeling and stores this new annotation in the ’consider merge type’ column as a candidate for the response variable. The following description outlines the primary output of initial dataset construction and its crucial functionality within the main product.

The below description illustrates the main product of initial dataset construction and critical functionality within the main product.

3.3.2.1 Overall continuous time algorithm mechanism

Our first step involves parsing the EAF structure, similar to the XML format, followed by storing and converting this structure, which comprises time and precise utterance content (cf. Listing A1.1), into Python DataFrame format. This process aims to establish continuous time intervals for each speaker, thereby creating an initial DataFrame, which serves as the input specification for this project. Subsequently, the core functionality is outlined in the pseudo-code below (cf. Algorithm.1).

There are four main steps for the below pseudo code (cf. Algorithm.1):

1 [*Lines* 1-5] This group of code intends to parse the continuous time in the EAF with respect to the "TIME_ORDER" tag and store it into a different data frame (cf. listing.A1.2) as this information could provide an initial timestamp and a terminated timestamp. Followed by the acquisition of the participant list by "Tier id" from the initial data frame, a regular expression is employed to query strings starting with "P" and "M" to locate this pattern to find players in specific session and store them into a participant list in the current session. The final instruction in this group filters the laughter tier as this column contains specific laughter annotations, such as discourse laughter.

2 [*Lines* 6-10] Once acquire a unique participant list, the algorithm iterate three players to extract relevant information. For the specific player in each iteration, the program selects specific rows for the designated participant name, such as “P002”. Given the demand for dynamically inserting time intervals and filling

silence gaps, the approach of constant insertion for a linked list data structure is being considered for this purpose. Line 10 starts iteration at each time interval in the specific player tier. This logic read the raw utterance annotation as input and produced labelled utterance type as the output by the algorithm.2.

3 [*Lines* 11-21] The next step determines the inclusion of laughter macro in the current time tuple. On the one hand, if it exists, this logic aligns the abstract laughter annotation, such as "[laugh]" with the precise laughter "Discourse laughter". It merges it into crossed laughter annotation, such as "[laugh]-discourse" by the Algorithm.3. This logic then merges with laughter annotation with another element, such as "S" if it exists a situation where a particular speaks with laughter. Once this annotation proceeds, include the current time interval and associated annotation in the link list. On the other hand, if no laughter macro is present, direct insert the current time interval and annotation into the link list.

4 [*Lines* 22-30] The last code group scrutinises the start time with 0 ms and the end time with the designated end time in the EAF. If it does not fit this condition, insert a silence interval ahead of the link list or tail of the link list. The last step is to transfer the link list to the data frame.

Algorithm 1 Overall procedure of utterance event for specific player

1: **function** OVERALLEVENTCONSTRUCTION
2: $\mathcal{D} \leftarrow$ *GetDataframeFromEAF*()
3: $t_{lastTimenode} \leftarrow$ *GetlastTimeStamp*() ▷ Acquire the last time moment in the current session
4: $\mathcal{P} \leftarrow$ *ParticipantList*() ▷ The set contains unique participant list
5: $\mathcal{L} \leftarrow$ *FilterTier*($\mathcal{D}$, 'laughter_section tier')
6: **for all** $p \in \mathcal{P}$ **do**
7: $\mathcal{D}_p \leftarrow$ *FilterPlayer*($\mathcal{D}, p$) ▷ $\mathcal{D}_p$: The laughter tier vector of particular player
8: $\mathcal{I}_p \leftarrow$ *PlayerMomentLinkedList*() ▷ $\mathcal{I}_p$: LinkedList for p
9: **for all** $(t_S, t_E, annotation) \in \mathcal{D}_p$ **do**
10: $v \leftarrow$ *UtteranceCategorisation*(*annotation*) ▷ Call Algorithm.2 to categorise the utterance in the EAF into our predefined format
11: $laughMarco_str \leftarrow$ '[laugh]'
12: **if** $laughMarco_{str} \in annotation$ **then** ▷ Determine current annotation contains laughter marco
13: $\tau \leftarrow (t_S, t_E)$ ▷ Construct time tuple
14: $\lambda \leftarrow$ *LaughterIntervalAlignmentAlgorithm*($\tau, \mathcal{L}$) ▷ Call Algorithm.3 to align current abstruct laughter to concrete version
15:
16: $\mu \leftarrow$ *UtteranceConncation*(v, λ) ▷ Link laughter with other element in utterance
17: $\mathcal{I}_p$.*AddNodeIntoLinklist*(t_S, t_E, μ)
18: **else**
19: $\mathcal{I}_p$.*AddNodeIntoLinklist*(t_S, t_E, v)
20: **end if**
21: **end for**
22: $t_{\text{currentSpeakerendTime}} \leftarrow \mathcal{I}_p$.*QueryForLastTimeInLinklist*()
23: **if** $t_{\text{currentSpeakerendTime}} < t_{lastTimenode}$ **then**
24: $\mathcal{I}_p$.*AddNodeIntoLinklist*($t_{\text{end}}, t_{lastTimenode}$, 'Silence')
25: **end if**
26: $\mathcal{D}'_p \leftarrow$ *LinklistToDataFrame*($\mathcal{I}_p$)
27: $t_{\text{min\_start}} \leftarrow$ *AcquireMinStartTime*($\mathcal{D}'_p$, 'Start time')
28: **if** $t_{\text{min\_start}} \neq 0$ **then**
29: *InsertSlienceIntoDataFrame*(0, $t_{\text{min\_start}}$, 'Silence', $\mathcal{D}'_p$)
30: **end if**

3.3.2.2 Silence detection

To enhance our program's ability to detect silence gaps, we have implemented dynamic time interval insertion, utilising a linked list structure for swift operations. Each node in this linked list holds crucial information: start time (in milliseconds), end time (in milliseconds), and annotation (a string). If the end time of the previous moment does not align with the start time of the current one, our program intelligently identifies and populates the gap with a node representing the silent interval. Specifically, it sets the start time of the gap as the end time of the previous moment and the end time of the gap as the current moment's start time.

3.3.2.3 Vocal classification

This functionality transforms original utterances in the EAF into a series of self-defied symbols, such as V to represent speaking and [V] to respect non-laughter vocalisation. Additionally, it preserves punctuation, including exclamation marks, question marks, commas, full stops, and semicolons, to rephrase the original utterance without loss of precision at this stage. The algorithm specification and pseudocode are provided in the description below.

Algorithm specification

1. **Input:** *s* be a input string in the annotation.
2. **Output:** The function produces a classified string string.
3. **Sample input and output:** "[eh] the next one [i]." → "[V] V [V]."; "Laugh] Violin" → "[laugh] V" ; "Hello, this is a test + with [brackets] and other symbols." → "V, V [V] V.".

Algorithm 2 Utterance categorisation

```
1:  function UTTERANCECATEGORISATION (inputStr)
2:      inputStr ← TransformIntoLowerCase(inputStr)
3:      strList ← SplitStringBySpace(inputStr)
4:      result_list ← InitEmptyList()
5:      for each currstr in strList do
6:          currstr ← RemoveLeadingTrailingSpce(currStr)
7:          if ContainLaughterMarco(currStr) then
8:              augmentText ← ProcessLaughterMarco(current_str)
9:              new_list.Append(augmentTtext)
10:         else
11:             utteranceStr ← ProcessUtterance(currentStr)
12:             newList.Append(utteranceStr)
13:         end if
14:     end for
15:     returnStr ← TrnasfromStringVectorIntoString (newList)
```

With the Function.2, there are three main steps:

1 [*Lines* 1-4] This logic first transforms the input string into lowercase,and then splits a single string into a string vector.

2 [*Lines* 5-14] Iterate the string vector and remove leading and trailing spaces in each turn. Determine if the current string contains laughter marco ("[laugh]"). If it contains this marco, our program maintain original string string as it is; otherwise, the function replaces each string with "V" to simplify the utterance representation.

3 [*Lines* 15] The last step involves converting string vectors into one string separated by one space.

3.3.2.4 Laughter alignment

This function aims to align the abstract laughter annotation from the “participant tier” into specific laughter in the “laughter_section” tier(cf.figure.A1.2).

Inputs:

- l_{current}: current moment tuple,where $l_{\text{current}} = (s_{\text{p}}, e_{\text{p}})$
- L: Laughter section list, where each tuple in L is the below format:$(s_l, e_l, exact_laugh)$

Output:

- s: Specific laughter or NULL.

Algorithm 3 Laughter Interval Alignment algorithm

```
function LaughterIntervalAlignmentAlgorithm(curr_tuple, laughter_list)
    s_p ← curr_tuple[0]                          ▷ s_p denotes the start time of current participant
    e_p ← curr_tuple[1]                          ▷ e_p denotes the end time of current participant
    for inner_l in laughter_section_list do
        s_l ← inner_l[0]                         ▷ s_l denotes the start time of current laughter moment
        e_l ← inner_l[1]                         ▷ e_l denotes the end time of current laughter moment
        subCond1 ← s_p ≥ s_l
        subCond2 ← e_p ≤ e_l)
        1stCond ← (s_p = s_l ∧ e_p = e_l)        ▷ The former interval equals the later interval
        2ndCond ← (subCond1 ∧ subCond2
                                                 ▷ The former interval is within the later interval
        3rdCond ← (s_p ≤ s_l ∧ e_p ≥ e_l)        ▷ The former interval contains the later interval
        if 1stCond ∨ 2ndCond ∨ 3rdCond then
            specific_laugh ← inner_l[2]
            return specific_laugh
        end if
    end for
    return “null”
```

The above Function.3 aligns with three conditions of the current moment with laughter Marco ([laugh])in the iteration with specific laughter, such as laughter that is laughable in the laughter tier.

3.3.2.5 Simplified representation of self-defined annotation type

The initial dataset algorithm stimulates the symbols in the discourse, but some of them represent the same semantics, such as “V [V] V.” and “V [V].”. Therefore, it is imperative to devise a more coherent representation. This symbolic representation incorporates two types of utterances: single-type utterances and merge-type utterances. A single-type utterance consists of original laughter tags, such as “[laugh]-Mirthful,” silence tag “[V],” and “S” tag. In merge-type utterances, the “M” symbol indicates the combination of “S” and “[V]” when both types of elements appear in the same utterance, with the occurrence of each type being greater than or equal to one.

3.3.3 Varied duration construction algorithm

The overhead of the varied duration dataset follows the format described below (cf. Figure 3.6). To construct each moment within this overhead, the Python Audio Segment API divides each audio into continuous segments, with each segment’s duration matching that of the corresponding moment in the initial dataset (cf. Figure 3.5). Because each moment’s duration may vary, resulting in what we term the varied duration dataset. Subsequently, the Python Opensimile API extracts 6,373 acoustic properties from each audio segment clip per moment, appending these properties to the initial dataset overhead. At the conclusion of each iteration, the generated audio file is deleted to conserve space(cf.Figure.3.7).

Figure 3.6: The overhead of Varied duration dataset

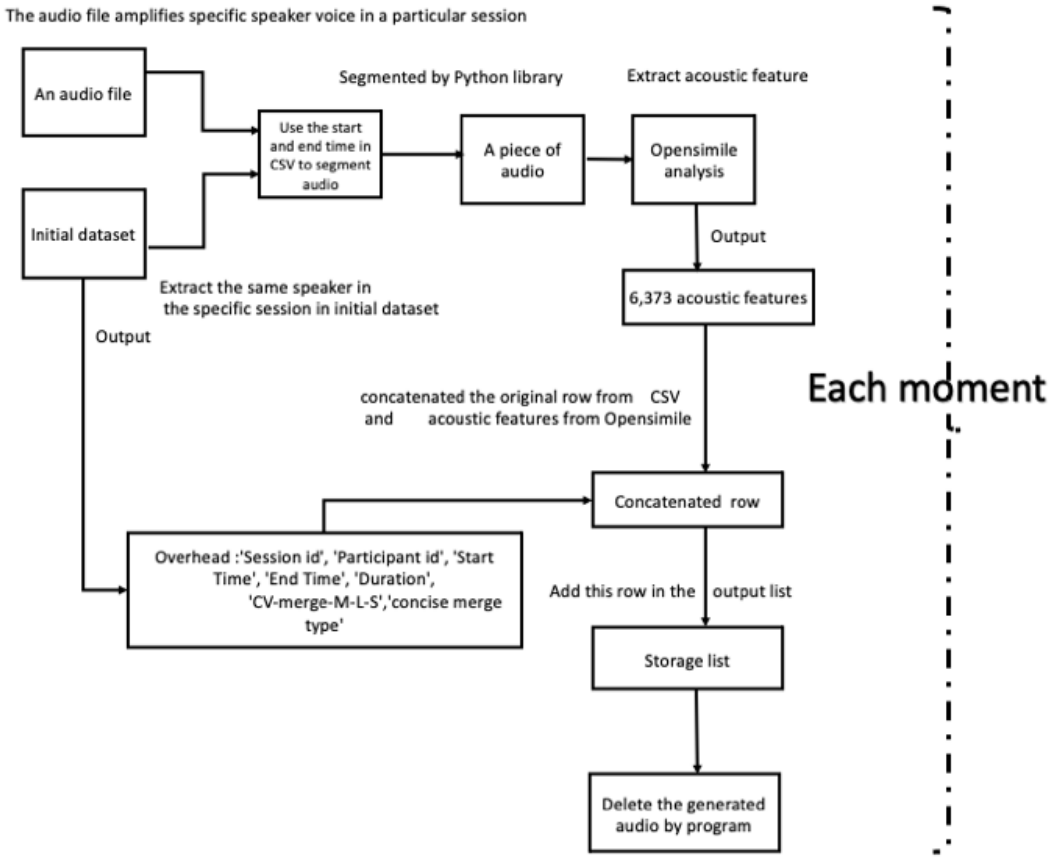


Figure 3.7: The process of construction of varied duration dataset from two CSV system

3.3.4 Fixed duration construction algorithm

This section explains the alignment process for constructing a fixed-duration dataset from two CSV systems. Then, it offers concrete examples of how our solution addresses the alignment issue between different duration systems.

3.3.4.1 Overall process

The overhead of the varied duration dataset is expected in the below format (cf. Figure.3.8). When compared to the overhead of the varied duration dataset (cf. Figure 3.6), the fixed duration dataset overhead includes two additional columns: "First Time Appeared" and "Count for the Same Utterance Type." The first new column contains two values, "true" and "false," while the second extra column records occurrences of the same event.

Session id	Participant id	Start Time	End Time	Duration	CV-merge-M-L-S	concise merge type	First time appear	Count for the same utterance type	6,373 acoustic features

Figure 3.8: The overhead of fixed duration dataset

The term "concise merge type" is defined below as an utterance event. The diagram (cf. Figure 3.9) below

illustrates one participant in one session, introducing the alignment process with two CSV systems: ELAN CSV and OpenSimile CSV. It demonstrates how to construct a fixed-duration dataset.

The general alignment mechanism involves iterating over each moment in the ELAN interval and using one utterance event to match one row or multiple rows in the OpenSimile CSV. For any ELAN CSV session, this program first extracts a unique player list, including two participants and one facilitator, then iterates over different players. For the specific player, the reduced data frame is extracted by filtering the specific players. Once a specific player data frame is acquired, , two cases arise for aligning the two systems.

Firstly, if the current duration is greater than or equal to 200 milliseconds, which surpasses the minimum range of the human hearing system, the program proceeds. The duration at this moment is divided by 200 to determine the quantity of integer iterations that match with the OpenSimile interval. Then, the remainder interval is calculated using modulo 200 to determine whether the current reminder interval belongs to the current utterance event or the next. Secondly, if the duration is less than 200 milliseconds, regardless of how tiny the interval is, this interval belongs to the next utterance event.

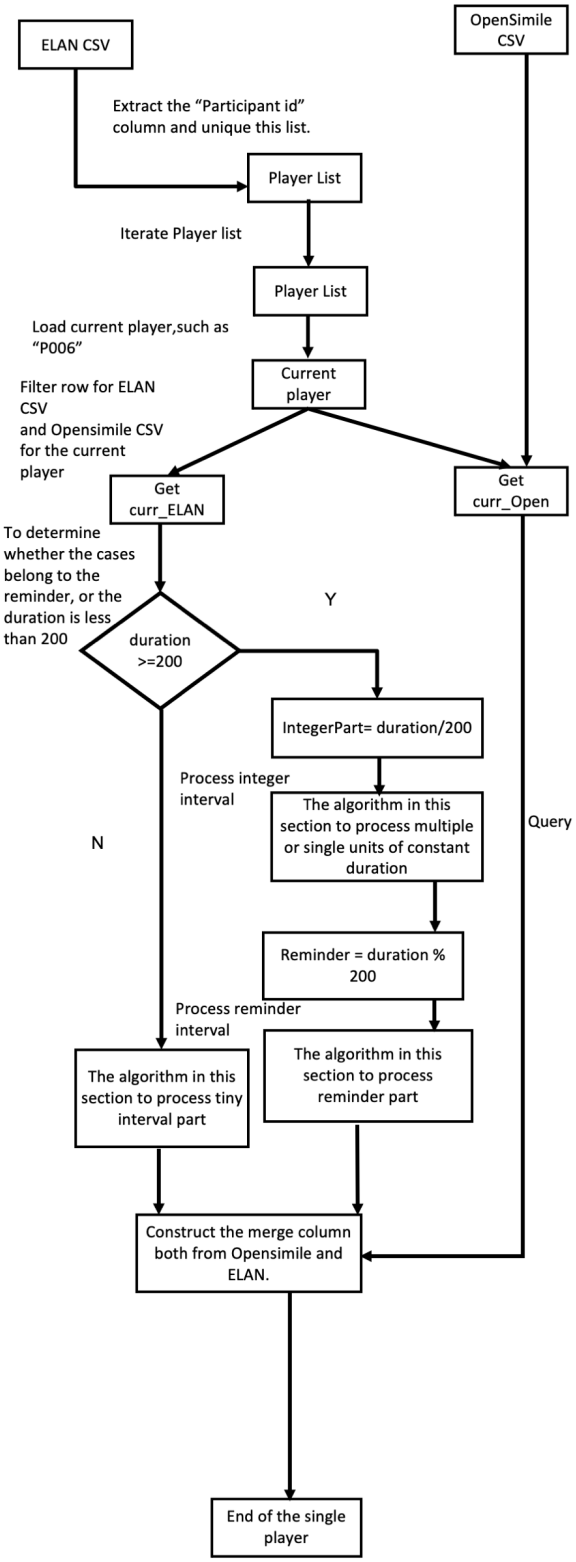


Figure 3.9: The diagram of alignment between fixed duration Opensimile interval and varied ELAN interval

3.3.4.2 Solution to alignment between two duration CSVs

This program utilised ELAN CSV in session 2, as shown in Figure .3.10, and Opensimile CSV in session 2, as shown in Figure .3.11, to generate the merged constant duration CSV. It is important to note that the ELAN CSV (cf. Figure 3.10) contains utterance events, whereas the Opensimile CSV does not.

Session id	Participant id	Start Time-ms	End Time-ms	Duration-ms	CV-merge-M-L-S	concise merge type
S02	M001_S02	0	1375	1375	[laugh]-Discourse	[laugh]-Discourse
S02	M001_S02	1375	6441	5066	V. [V] V.	M
S02	M001_S02	6441	7057	616	V	S
S02	M001_S02	7057	8812	1755	V	S
S02	M001_S02	8812	9380	568	V	S
S02	P006	0	6441	6441	Silience	Silience
S02	P006	6441	7057	616	V.	S
S02	P007	0	8812	8812	Silience	Silience
S02	P007	8812	9380	568	V	S

Figure 3.10: The sample dataset extracted from Session 2 ELAN CSV

session id	participant id	start time-m	end time-ms	duration-ms	audspec_lengthL1norm_sm	audspec_len	audspec_len	audspec_len	audspec_len	audspec_len	audspec_len	audspec_len	audspec_len
S02	M001_S02	0	200	200	0.032610729	0.61538464	0	0.11116292	0.11278435	0.11977001	0.00162143	0.00698566	0.00860709
S02	M001_S02	200	400	200	0.267099977	0.92307693	0	0.26656941	0.28875604	0.35221267	0.02218664	0.06345662	0.08564326
S02	M001_S02	400	600	200	0.670082331	0.84615386	0.53846157	0.48018655	0.7058385	0.85709399	0.22565195	0.15125549	0.37690744
S02	M001_S02	600	800	200	0.47559154	0	0.92307693	0.15833233	0.25835705	0.45445681	0.10002472	0.19609976	0.29612445
S02	M001_S02	800	1000	200	0.507164657	0.69230771	0	0.1657692	0.20949621	0.50023299	0.04372701	0.29073679	0.33446378
S02	M001_S02	1000	1200	200	0.201477885	0.69230771	0	0.86057472	0.89355737	0.93258077	0.03298265	0.0390234	0.07200605
S02	M001_S02	1200	1400	200	0.312408507	0	0.92307693	0.73263162	0.79856056	0.83427811	0.06592894	0.03571755	0.10164648
S02	M001_S02	1400	1600	200	0.71144557	0.61538464	0.15384616	0.22873759	0.60800266	0.72642761	0.37926507	0.11842495	0.49769002
S02	M001_S02	1600	1800	200	1.964694381	0.23076923	0.92307693	2.31669426	3.05255795	3.34354711	0.73586369	0.29098916	1.02685285
S02	M001_S02	1800	2000	200	0.320117176	0	0.53846157	0.21231262	0.26339531	0.32390922	0.05108269	0.06051391	0.1115966
S02	M001_S02	2000	2200	200	1.789293289	0.61538464	0.23076923	0.84340984	1.18737805	1.93334913	0.34396821	0.74597108	1.08993936
S02	M001_S02	2200	2400	200	3.338021755	0.30769232	0.92307693	0.86162907	1.72824514	2.72624779	0.86661607	0.99800265	1.86461878
S02	M001_S02	2400	2600	200	1.997183084	0.92307693	0.38461539	0.7835269	0.86983204	2.34915805	0.08630514	1.47932601	1.56563115
S02	M001_S02	2600	2800	200	2.315701008	0.53846157	0	1.38429749	2.02244687	2.71575928	0.63814938	0.69331241	1.33146179
S02	M001_S02	2800	3000	200	1.144848824	0.15384616	0.92307693	0.73325151	0.94466048	1.15224183	0.21140897	0.20758134	0.41899031
S02	M001_S02	3000	3200	200	1.934340715	0.92307693	0.46153846	0.55802673	0.7753368	1.12337351	0.21731007	0.34803671	0.56534678
S02	M001_S02	3200	3400	200	0.787099957	0.92307693	0.38461539	2.10834837	2.25099564	2.43468142	0.14264727	0.18368578	0.32633305
S02	M001_S02	3400	3600	200	0.998520911	0	0.92307693	0.2616109	0.3521148	0.76153374	0.0905039	0.40941894	0.49992284
S02	M001_S02	3600	3800	200	0.081804909	0.92307693	0.38461539	0.11722324	0.12473774	0.14919634	0.0075145	0.0244586	0.0319731
S02	M001_S02	3800	4000	200	5.090063095	0.92307693	0	0.41520581	1.51386309	4.29460812	1.09865725	2.78074503	3.8794024
S02	M001_S02	4000	4200	200	0.905243874	0	0.92307693	4.65243816	4.72024298	5.08355093	0.06780481	0.36330795	0.43111277
S02	M001_S02	4200	4400	200	1.132924318	0	0.53846157	0.26666579	0.31429562	0.54796481	0.04762983	0.23366919	0.28129902
S02	M001_S02	4400	4600	200	0.285835564	0	0.92307693	1.00280154	1.0407685	1.09907496	0.03796697	0.05830646	0.09627342

Figure 3.11: The sample dataset extracted from Session 2 Opensimile CSV

Definition of threshold The idea that auditory perception operates within milliseconds was initially proposed by Efron (1970). This study adopts a sampling duration of 200 milestones, aiming to keep each sample within the realm of human perceptibility. Specifically, it focuses on 120 milliseconds, identified as the lower threshold for acoustic perception intervals. The ratio between 120 milliseconds and 200 represents the minimum interval employed to recall the duration of the ELAN interval. Referred to as α, this ratio signifies the anchor usage saturation. To categorise any reminder interval, it should be divided by 200 milliseconds, and the resulting ratio compared with the anchor usage saturation to determine the placement of the utterance event.

Case analysis In Figure 3.9, three cases of alignment are mentioned: when reminder interval usage is greater than or equal to the threshold, when reminder interval usage is less than the threshold, and when it is of tiny duration.

•Case 1:

The remainder of the first row in ELAN CSV (cf. Figure.3.10) is calculated as 1375-200$\times$6 = 175, which corresponds to the first six rows in the Opensimile CSV(cf.Figure.3.11) belonging to the current moment event in the ELAN CSV(cf.Figure.3.10). As for the reminder moment from 1300 to 1400 milliseconds in the Opensimile, the reminder usage could be derived from this calculation: 175/200 = 0.875, greater than the anchor usage saturation of 0.6, indicating that this reminder

relatively fully utilises the interval within 200 ms(cf. Figure.3.12) and the event of this reminder should belong to the current utterance.

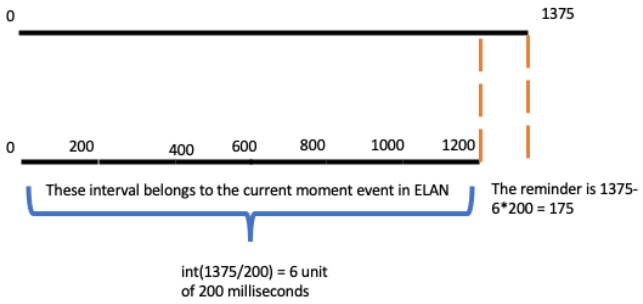


Figure 3.12: Example case 1 for threshold boundary

•Case 2:

For another case, if the duration is 471 milliseconds, the reminder could be derived from this calculation:471 = 200 $\times$2 +71. The remainder is 71. The interval usage is 71/200 = 0.355, which is less than the anchor usage saturation (cf. Figure 3.13), indicating this reminder event should belong to the next utterance event.

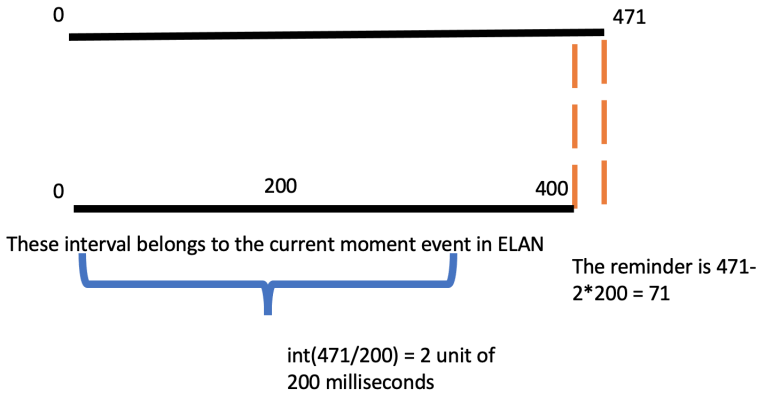


Figure 3.13: Example case 2 for threshold boundary

•Case 3:

To ensure consistency with the ELAN CSV, every event should be covered, even if its duration is less than 200 milliseconds (cf. Figure 3.14). Thus, the utterance of the event of the current moment should belong to the current utterance moment, regardless of how brief the current duration is. This operational logic maintains the coherence of the merge CSV, aligning it with the requirements of ELAN CSV.

Duration – ms	CV – merge – M – L – S	concise merge type	audspec_lengthL1norm_sma_range
All	All	All	All
6	Silience	Silience	NA
9	Silience	Silience	NA
10	Silience	Silience	NA
13	Silience	Silience	NA
14	Silience	Silience	NA
17	Silience	Silience	NA
20	[laugh]–Discourse	[laugh]–Discourse	NA
23	Silience	Silience	NA
27	Silience	Silience	NA
55	Silience	Silience	NA
66	V	S	0.000000000
66	Silience	Silience	0.000000000
73	V	S	0.000000000
74	V?	S	0.000000000

Figure 3.14: The duration in ELAN csv is less than 200

3.3.5 Comparison between different duration datasets

There are four aspects to compare between the two version duration datasets:

- Capacity: the capacity of varied version is 1.67 GB whereas the fixed counterpart is around 15 GB.
- The content in acoustic property extracted from Opensimile: Some moments in varied duration datasets that are less than 200 milliseconds do not have acoustic properties due to the artefact and configuration in OpenSimile(cf.figure.3.15). In contrast, the fixed duration dataset does not have this issue.
- Alignment issue: In the process of construing a varied duration dataset, each moment in Opensimile CSV does not discard as two CSV systems synchronise each other in each moment, while In the process of construing fixed duration dataset, some rows might discord in the Opensimle CSV as two systems moment is not always one-to-one perfect alignment.

Duration - ms	CV - merge - M - L - S	concise merge type	audspec_lengthL1norm_sma_range
[...]	All	All	All
6	[V]	[V]	NA
6	Silience	Silience	NA
6	Silience	Silience	NA
6	Silience	Silience	NA
9	Silience	Silience	NA
10	Silience	Silience	NA
13	Silience	Silience	NA
14	Silience	Silience	NA
17	Silience	Silience	NA
20	[laugh]-Discourse	[laugh]-Discourse	NA
23	Silience	Silience	NA
27	Silience	Silience	NA
55	Silience	Silience	NA
66	Silience	Silience	0.000000000

Figure 3.15: Issue in the varied duration:some laughter moments that are less than 60 ms could not be processed in Opensimile

3.4 Correctness of dataset construction

Dataset quality significantly impacts machine learning performance, thus necessitating thorough verification. Four key aspects must be assessed to ensure the quality of our dataset prior to conducting machine learning experiments.

3.4.1 Verification of code logic for the issue in the construction of the fixed duration dataset

To verify code logic in a fixed-duration dataset, potentially addressing alignment issues, the author of this dissertation employed a button-based mechanism. This method allowed the author to ascertain whether the reminder interval corresponds to the current utterance event or the subsequent one. The relevant code logic is presented in Listing 3.1 and is discussed in the context preceding the iteration before the integer part (highlighted in a yellow rectangle in Figure 3.16).

Code Listing 3.1: Button for verification of control fixed duration dataet alignment

```
if button_for_move_next_event == True:
   count_this_type += 1
   number_intervals -= 1
# When finished it should be shutdown immedicately
button_for_move_next_event = False
```

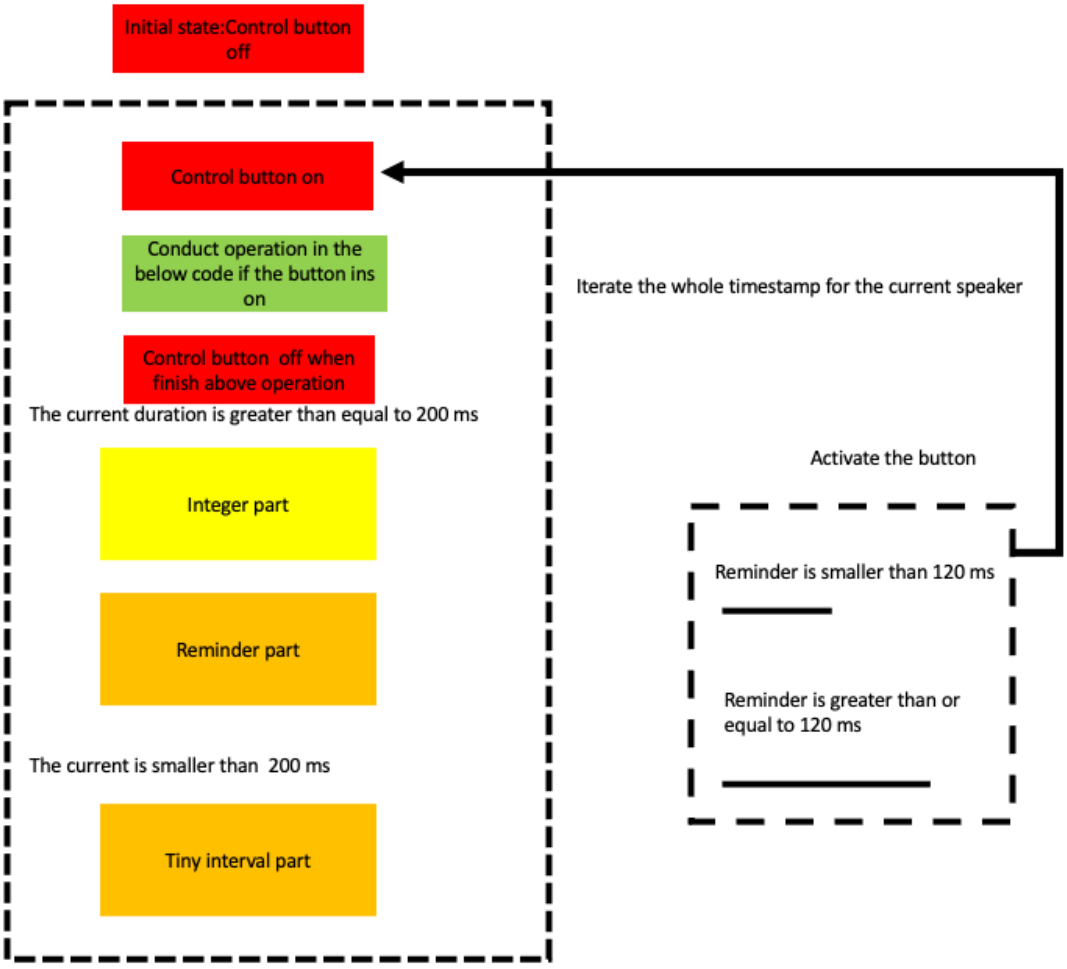


Figure 3.16: Utterance event control in fixed duration dataset

3.4.2 Verification of null category in response variable

Here are the NULL categories in the EAF file, and it needs to transform any laughter category containing "null", such as "[laugh]-null Discourse"(cf. Figure.3.17 and Figure.3.18) to “Ambiguous”.

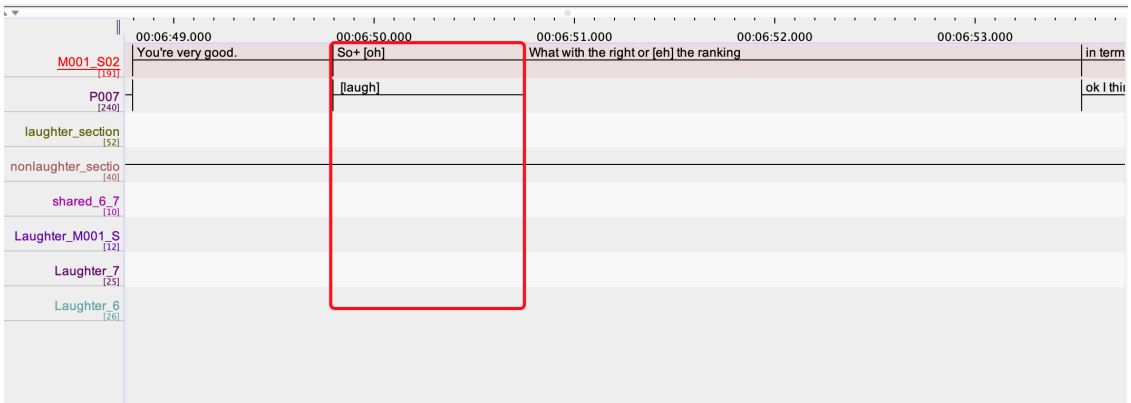


Figure 3.17: Missing annotation in session 2 in visualised by ELAN

Participant id	Start Time-ms	End Time-ms	Duration-ms	CV-merge-M-L-S	concise merge type
P007	408148	408799	651	V.	S
P007	408799	409795	996	Silience	Silience
P007	409795	410753	958	[laugh]-null	[laugh]-null
P007	410753	413536	2783	Silience	Silience
P007	413536	414771	1235	V	S
P007	414771	415464	693	V	S
P007	415464	416245	781	V.	S
P007	416245	417056	811	Silience	Silience
P007	417056	417577	521	V.	S
P007	417577	422722	5145	Silience	Silience
P007	422722	424481	1759	V	S
P007	424481	431211	6730	V	S
P007	431211	431879	668	V	S
P007	431879	433021	1142	V	S
P007	433021	433497	476	Silience	Silience
P007	433497	435427	1930	V	S

Figure 3.18: Inspection of missing annotation in session 2

Besides, this project also changes “[laugh]-Discourse S” and “[laugh]-Discourse [V]” into “[laugh]-Discourse” as it is hard to distinguish the speaking or non-laughter vocalisation is interleaved with laughter.

3.4.3 Verification of quantity of utterance event in the fixed duration dataset

To address the alignment issue, the fixed duration dataset requires merging two CSVs: the ELAN CSV and the OpenSimile version CSV.

	Session id	Participant id	Start Time - ms	End Time - ms	Duration - ms	CV - merge - M - L - S	concise merge type	Count for the same utterance type	First time appear	audspec_lengthL1norm_sma_range	...	mfcc_sma_d
0	S02	P007	0	200	200	Silience	Silience	1	True	0.016241	...	
1	S02	P007	200	400	200	Silience	Silience	2	False	0.035103	...	
2	S02	P007	400	600	200	Silience	Silience	3	False	0.036810	...	
3	S02	P007	600	800	200	Silience	Silience	4	False	0.049637	...	
4	S02	P007	800	1000	200	Silience	Silience	5	False	0.112556	...	
...	...	...	...	...	...	...	...	...	...	...	...	
147342	S23	P049	361200	361400	200	Silience	Silience	6	False	0.019752	...	
147343	S23	P049	361400	361600	200	Silience	Silience	7	False	0.039127	...	
147344	S23	P049	361600	361800	200	Silience	Silience	8	False	1.102097	...	
147345	S23	P049	361800	362000	200	Silience	Silience	9	False	0.456577	...	
147346	S23	P049	362000	362200	200	Silience	Silience	10	False	0.151887	...	

147347 rows × 6382 columns

Figure 3.19: The screenshot of constant dataset for the total utterance event inspection

This study uses two columns to verify the count number (cf. Figure 3.19). If the "count for the same utterance type" is one, it can be inferred that the "First time appears" column is True. Otherwise, if any value is not 1, it indicates that the "First time appear" column should be false.

3.4.4 Verification of total event in fixed duration dataset alignment with varied duration dataset

When processing the ELAN CSV and Opensimile CSV datasets with varied versions, each moment is aligned to the same duration. Utilising the varied duration dataset as a reference point, we can assess the completeness and accuracy of the fixed duration dataset. This involves selecting the 'count for the same utterance type' column and filtering the final events by choosing all instances where this count equals 1.

Here is screenshot of varied duration dataset:

	Session id	Participant id	Start Time - ms	End Time - ms	Duration - ms	CV - merge - M - L - S	concise merge type	audspec_lengthL1norm_sma_range	audspec_lengthL1norm_sma_maxPos
0	S02	P007	0	8812	8812	Silience	Silience	0.601711	0.359268
1	S02	P007	8812	9380	568	V	S	0.458016	0.673469
2	S02	P007	9380	18149	8769	Silience	Silience	0.348826	0.000000
3	S02	P007	18149	18620	471	V	S	1.032122	0.375000
4	S02	P007	18620	44627	26007	Silience	Silience	0.425555	0.023911
...	...	...	...	...	...	...	...	...	...
14436	S23	P049	357649	361628	3979	Silience	Silience	1.614275	0.353846
14437	S23	P049	361628	361883	255	V	S	1.109623	0.555556
14438	S23	P049	361883	363365	1482	Silience	Silience	0.394520	0.007092
14439	S23	P049	363365	363828	463	V	S	2.253242	0.794872
14440	S23	P049	363828	365917	2089	Silience	Silience	0.633792	0.000000

14441 rows × 6380 columns

Figure 3.20: The screenshot varied duration dataset

Here are all the rows where the value in the "count for the same utterance type" column equals 1:

	Session id	Participant id	Start Time - ms	End Time - ms	Duration - ms	CV - merge - M - L - S	concise merge type	Count for the same utterance type	First time appear	audspec_lengthL1norm_sma_range	...	mfcc_sma_d
0	S02	P007	0	200	200	Silience	Silience	1	True	0.016241	...	
44	S02	P007	8800	9000	200	V	S	1	True	0.045487	...	
47	S02	P007	9400	9600	200	Silience	Silience	1	True	0.221935	...	
91	S02	P007	18200	18400	200	V	S	1	True	0.962688	...	
93	S02	P007	18600	18800	200	Silience	Silience	1	True	0.175560	...	
...	...	...	...	...	...	...	...	...	...	...	...	
147307	S23	P049	354200	354400	200	Silience	Silience	1	True	0.042384	...	
147327	S23	P049	358200	358400	200	V	S	1	True	0.162915	...	
147328	S23	P049	358400	358600	200	Silience	Silience	1	True	0.044638	...	
147335	S23	P049	359800	360000	200	V	S	1	True	0.092651	...	
147337	S23	P049	360200	360400	200	Silience	Silience	1	True	0.069276	...	

14441 rows × 6382 columns

Figure 3.21: All unique utterance events in the fixed duration dataset

Based on figures 3.20 and 3.21 above, it is apparent that both the total row count of the varied dataset and the total row count of the fixed duration dataset, under the condition where the value in the "count for the same utterance type" column equals 1, amount to 14,441 rows. This observation suggests that the fixed duration dataset exhibits 100% accuracy.

3.5 Data analysis methods

The section on Data analysis methods involves dataset exploration, identification of discriminating features and interaction among acoustic features. It is important to note that all experiments are currently in the

design phase, and the results will be presented in the subsequent chapter dedicated to results.

3.5.1 Dataset exploration

Data exploration presents seven types of response variables in “consider merge type”, with these labels serving as the independent variables, while 6,373 acoustic properties act as the dependent variable Table 3.1 below provides explanations for each label.

Table 3.1: Classified utterance and explanation

Self-defined response variable	Explanation
S	Spoken words such as "ahaha"
Silence	Silence moment
M	Merge type consists of any number of speaking and non-laughter vocalisation,such as "[ah] I ....[em]"
[laugh]-Discourse	Discourse laughter
[V]	Non-laughter vocalisation
[laugh]-Mirthful	Mirthful laughter
Ambiguous	Due to missing labels in the EAF file, some annotations exist, such as "[laugh]-null.". We transform all these types into "Ambiguous".

This section also presents the quantity of discourse and mirthful laughter in the varied and fixed-duration datasets. Upon examining the figures (cf. Figure 3.22 and Figure 3.23), it becomes evident that laughter accounts for only a small fraction of the total label quantity.

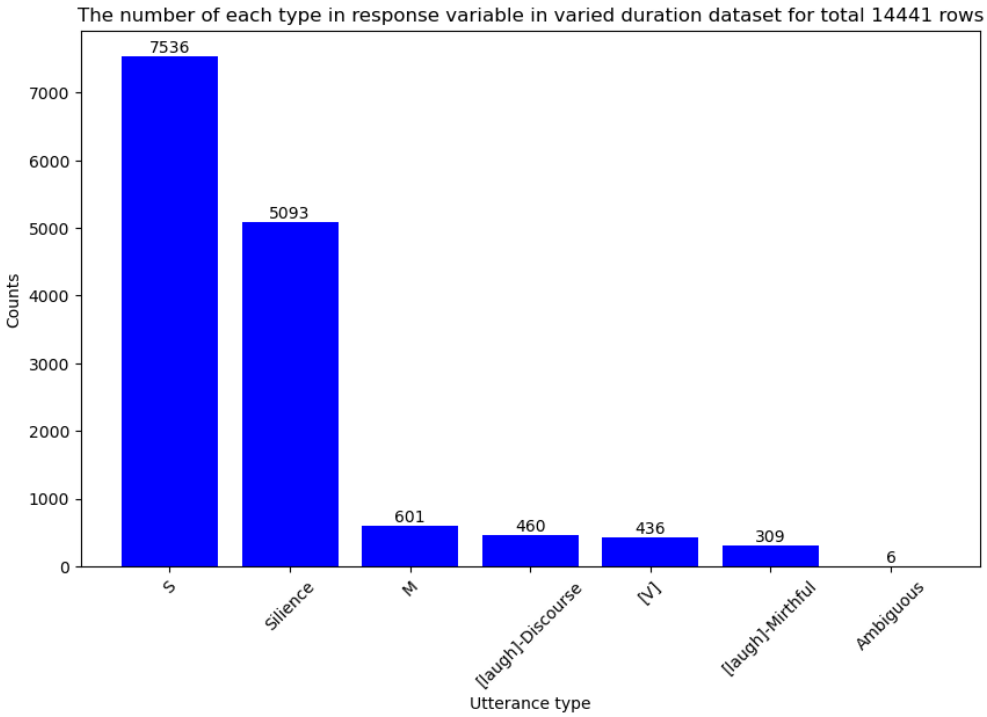


Figure 3.22: The label distribution of varied duration dataset

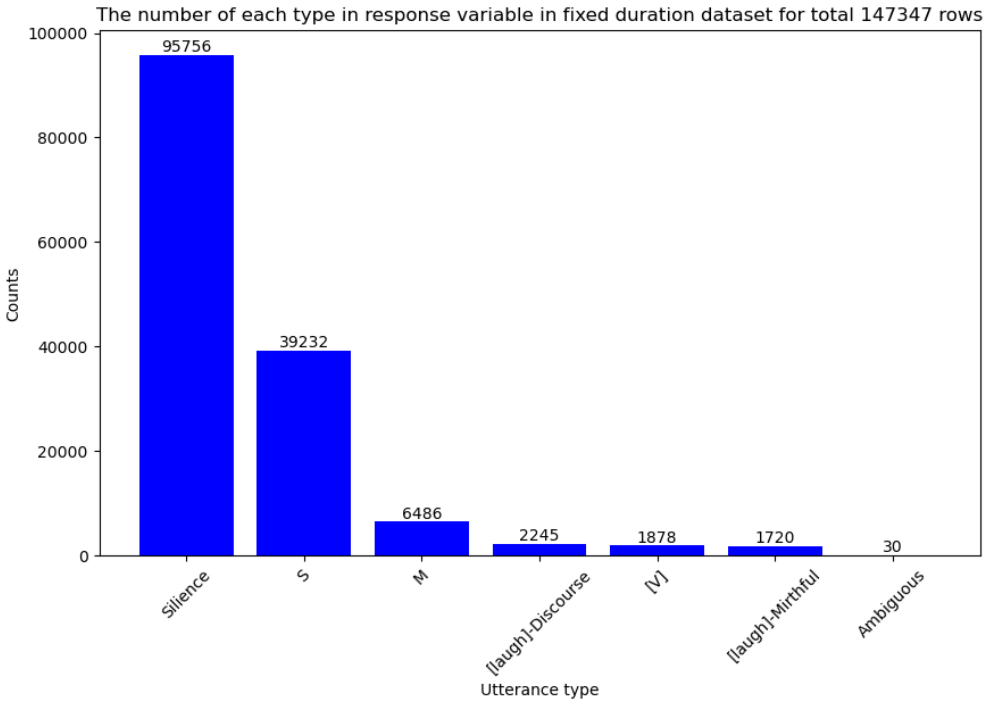


Figure 3.23: Fixed duration dataset label distribution

3.5.2 Identification of discriminating acoustic features

In this section and results chapter, both acoustic properties and acoustic features are interchangeably referenced. This project employs two models, Decision Tree and Multinomial Regression, to identify

distinctive features in various types of laughter. Before inputting the dataset into the Decision Tree model, the target variable must be converted into binary classification; otherwise, the model won't discern the discriminating features effectively. This decision model design process is depicted in Figure 3.24. Regarding multi-label classification (refer to Figure 3.25), this project leverages its characteristics to extract the coefficients of features corresponding to different labels, thus identifying the distinguishing features. In both models, a Decision Tree is employed to visualise the discriminant features.

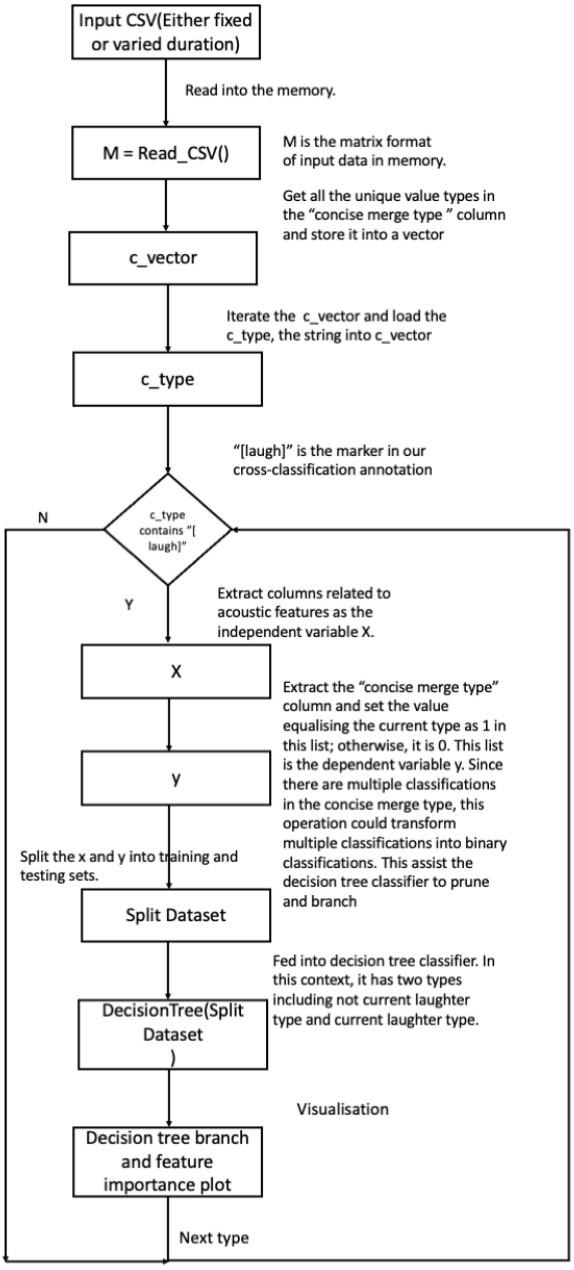


Figure 3.24: The diagram for the discriminating acoustic feature process generated by decision tree model

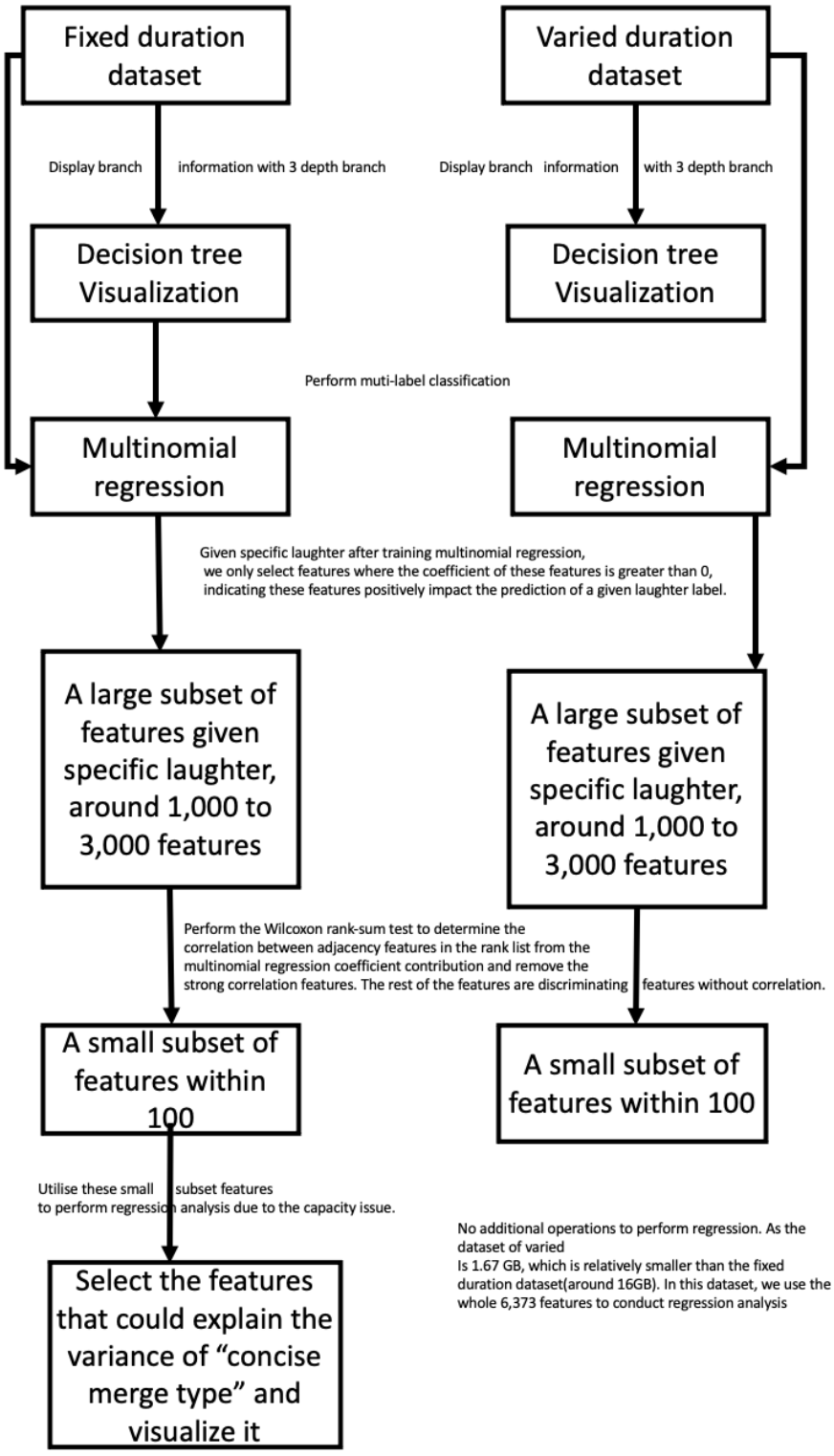


Figure 3.25: The diagram for the discriminating acoustic feature process generated by multinomial logistic regression model

3.5.3 Interaction among discriminating acoustic features

To examine the interaction among acoustic features, we employed the Wilcoxon signed rank test to compare the adjacency feature generated by the machine learning approach utilised in this project after conducting a normality test,such as Shapiro-Wilk test (cf. Figure 3.26).

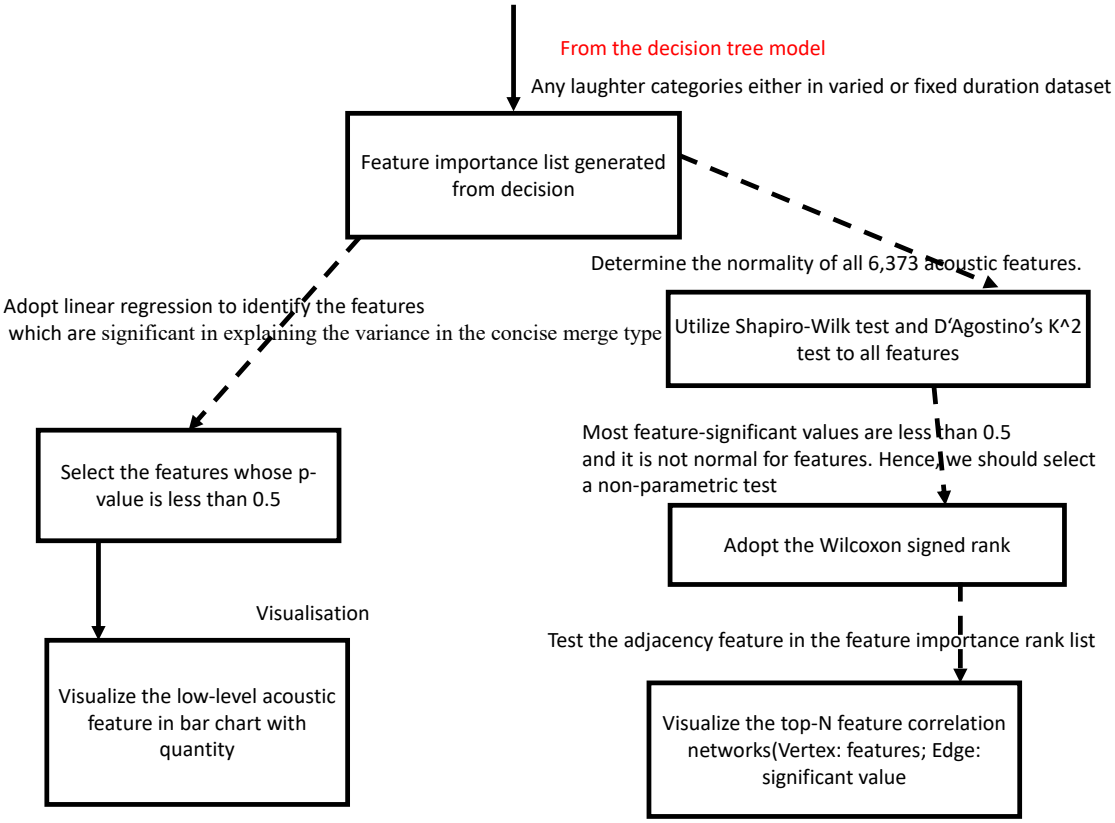


Figure 3.26: The diagram of interaction between discriminating acoustic feature

3.6 Summary of methodology

The methodology chapter introduces the research design in a high-level description and proposes research methods, encompassing data construction and dataset analysis. Prior to conducting machine learning experiments, this work rigorously validates dataset annotations across various dimensions to ensure accuracy. In essence, this chapter furnishes a fundamental framework for the subsequent chapter.

4 Results

The results chapter presents the findings of the machine learning experiment in the methodology chapter. This chapter will begin by detailing the experiment's parameter settings. Subsequently, it will present the binary classification results followed by the multi-classification results. Next, a synopsis of discriminative features in our study will be provided. Finally, the chapter will compare our results with those of Tanaka & Campbell (2014), as well as discuss the findings regarding acoustic feature interactions.

To ensure clarity, we consistently employ the term "feature" to denote acoustic properties within the experimental context. Here, acoustic properties serve as features, while the response variable (concise merge type) functions as the target or dependent variable. When referring to acoustic features, we intend the same semantic meaning as acoustic properties in this study. Furthermore, each low-level acoustic property extracted from the Opensimile includes acoustic property names and functional information, such as statistical details. We interpret acoustic property names exclusively by querying them on this website, excluding statistical information to facilitate summarisation and comparison with others' work(cf. link https://github.com/rupafn/CulturalClassifier).

4.1 Experiment setup

Before presenting the experimental results, it is helpful to provide an overview of the experimental setups to assist in replicating this work. The following four subsections detail each component configuration used in our experiment.

4.1.1 Machine learning model parameter configuration

This dissertation employed two machine learning models: a decision tree for binary classification and multinomial logistic regression for multi-class classification. Given that the computation time required for hyperparameter tuning can be substantial, we predefined parameter collections for each model. Below are two configurations.

Code Listing 4.1: Decision tree model parameter

```
from sklearn.tree import DecisionTreeClassifier
model = tree.DecisionTreeClassifier(criterion='gini',
max_depth=3,min_samples_split=10,
random_state=42,class_weight='balanced')
```

Here is the explanation for decision tree model parameter setting (cf.listing.4.1):

•criterion: The ‘criterion‘ parameter denotes the partition algorithm in the decision tree model. There are two mainstream approaches based on entropy and the Gini index. We select the Gini index based on computationally efficient and high-dimensional data orientation compared to associated properties in entropy(cf.link:https://www.javatpoint.com/gini-index-in-machine-learning#:~:text=Advantages%3A,example%2C%20entropy%20or%20misclassification%20rate.).

•max_depth:To fit the page size properly, we set the max depth of decision tree as 2 or 3 depend on varied duration dataset or fixed duration dataset.

•min_samples_split :To prevent over-split, we set the minimum sample partition in each node as ten units.

•random_state:To achieve the same results in different executions, we set random states to a specific number forcibly.

•class_weight':As our dataset has a relative imbalanced distribution towards laughter and other elements in the discourse, such as non-speaking vocalisation, we need to allow the classifier to adjust imbalanced target labels.

Here is the multinomial regression parameter setting, and this configuration comes from the modification of Karishma (2022)'s work as they adopted basic configuration. It is convenient to adjust in our work(cf.listing.4.2).

Code Listing 4.2: Multinomial logesitic regression model parameter

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression (multi_class='multinomial',
penalty='l2',
 solver ='lbfgs',
max_iter=1000,class_weight='balanced')
```

•multi_class:We set this parameter as the multinomial label forcibly to signal the classifier to deal with the multi-labelling tasks.

•penalty:We opt for the L2 penalty in the penalty parameters to prevent overfitting in multinomial regression based on squaring hyperparameters.

•solver:As for the optimisation solver, we selected “lbfgs”,as this quasi-Newtown method could handle large computations for a relatively small capacity dataset(around 1GB to 20 GB)and multinomial issues.

•max_iter:As our dataset might need large amount of time to execute, we set 1000 max iteration to control converge turn.

•class_weight: The same setting and explanation as in the decision tree model(cf.listing.4.1).

4.1.2 Considerations of discriminating feature range

In the decision tree model, we selected acoustic properties with feature importance scores above 0, determined through ranking in the built-in classification package, according to different types of laughter such as discourse or mirthful laughter. In examining feature importance for the multi-label labeling task, the

coefficient matrix produced by multinomial logistic regression highlights each significant feature for various utterance types. Specifically, we consider coefficient values greater than 0 for a given laughter type as distinguishing features for that particular laughter.

4.1.3 Hypothesis testing for feature correlation and feature reduction

Before conducting hypothesis testing, this study utilised the Shapiro-Wilk test and D'Agostino's K2 test to verify the normal distribution of discriminating properties before selecting types of statistical hypothesis testing (non-parametric or parametric) to verify the normality of each feature. After normality testing, it was revealed that our data followed a non-normal distribution in most cases. Consequently, we opted for the Wilcoxon signed-rank test. The research employed the Wilcoxon signed rank test for the decision tree model to inspect the top-N acoustic properties, as the number of top-N acoustic properties is within 10.

Another application of the Wilcoxon signed-rank test for multinomial regression involved filtering adjacency acoustic properties. This was based on the criterion that the coefficient feature exceeds 0 in the rank list generated by multinomial regression. The justification for this approach hinges on the number of these acoustic properties, which falls within the range of approximately.

4.1.4 Regression analysis for identification of significant feature explaining variance of target variable

Regression analysis identifies acoustic properties that could explain the variance of the response variable, indicating that these significant properties correlate with the response variable when the coefficient of this independent variable is less than the alpha value(0.05).

This research selects different sets of acoustic properties for different sampling datasets for regression analysis. For the varied duration dataset, which has a relatively minor capacity (1.67 GB), we utilise all acoustic properties.

However, for the fixed duration dataset with a larger capacity issue (around 15 GB), before conducting regression analysis, we apply the Wilcoxon rank test to reduce features by removing strongly correlated ones. Additionally, each unique utterance event in the varied duration dataset might occupy multiple rows. After this operation, the acoustic properties fed into regression analysis are limited to 100, a feasible number for our machine to execute.

With this configuration, we can acquire the acoustic properties that explain the variance in different duration datasets given different types of laughter. Then, we elucidate our process of analysing acoustic properties by employing linear regression to examine each feature, explaining the significant variance in the associated laughter in this context, specifically "[laugh]-discourse". We select the features whose p-value is less than 0.05.

Sequentially, for each select feature, such as "jitterLocal_sma_quartile3", we retain the text before the first underscore("_") as this text represents the lower-level acoustic feature, while the text after the first underscore indicates functional information, such as statistical details. We can derive the following visualisation by tallying each processed text.

4.1.5 Decision tree branch visualisation explanation

This study employed a decision tree for visualising both binary and multi-labels within the response variable. In the experimental setup, the "Sklearn" library facilitated the acoustic selection process through the implementation of a decision tree model (cf. link.https://scikit-learn.org/stable/modules/generated/sklearn.tree.export_graphviz.html).

Each node or leaf has several parameters related to branching and here is the statement of the parameters in each row per node:

•The row related to the comparison between specific acoustic properties with some threshold indicates the discriminating feature at the current level.

•The row related to the Gini index measures the impurity of this node. The less impurity, the better the separation is.

•The text related to samples denotes the quantity of each response label in the current separation.

•The row related to class shows the predominant class in this node.

4.2 Binary classification results

This section initially presents decision tree visualisation for binary labelling tasks,and discriminating features of discourse and mirthful laughter in varied and fixed duration datasets, ranked by the decision tree model. Subsequently, this subsection focuses on decision tree visualisation for binary labelling task.

4.2.1 Feature selection visualisation

The decision tree depicted above is constructed for binary classification, distinguishing between non-discourse laughter (“Not [laugh]-Discourse”) and discourse laughter (“[laugh]-discourse”) within the varied duration dataset (cf.Figure.4.1). In this tree, the root node’s determiner is spectral Features. If a sample’s spectral features are less than or equal to 0.036 Hz, the predominant utterance type is deemed to be discourse laughter. Subsequently, if this condition holds, the model predicts the class as non-discourse laughter; otherwise, it predicts it as discourse laughter. The tree continues to branch based on different discriminators until the third depth level is reached.

It is important to note that the presented tree only exhibits thresholds for the third-level depth without comparisons involving different acoustic properties. Analysis of this tree highlights that several low-level acoustic properties, such as spectral features, mel-frequency cepstral coefficients, auditory spectrum, and jitter, play pivotal roles in determining discourse laughter within the varied duration dataset.

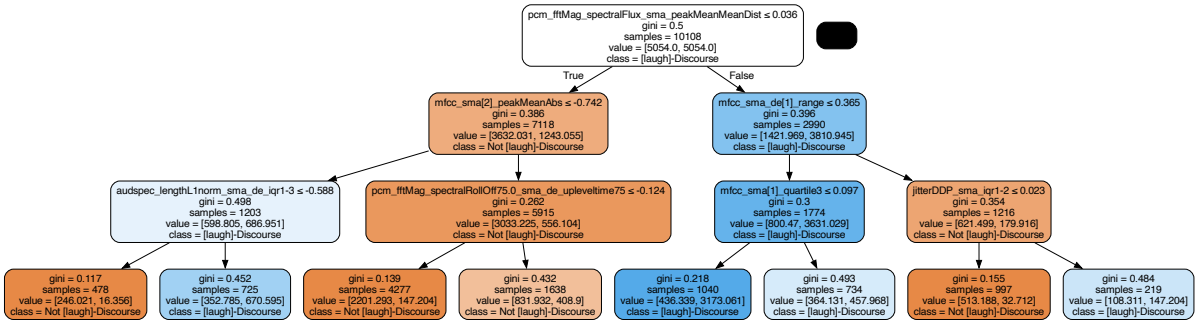


Figure 4.1: Feature selection process of discourse laughter in varied duration dataset

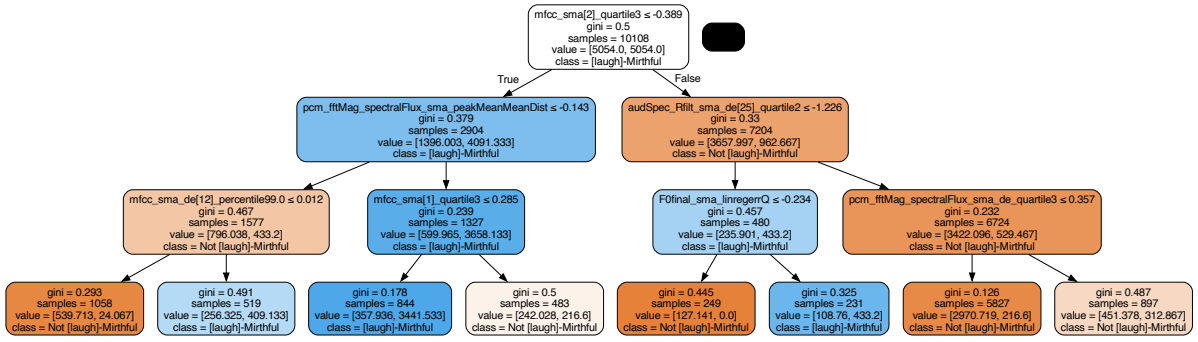


Figure 4.2: Feature selection process of mirthful laughter in varied duration dataset

The decision tree depicted in Figure.4.2 is designed for binary classification, distinguishing between non-mirthful laughter (“Not [laugh]-Mirthful”) and mirthful laughter (“[laugh]-Mirthful”) within the varied duration dataset (cf.Figure.1.2). At its core, the root node hinges on mel-frequency cepstral coefficients. When a sample’s Mel-frequency cepstral coefficients fall below or equal to -0.389 W/Hz, the prevailing utterance type is identified as discourse laughter. Notably, since Mel-frequency cepstral coefficient represents a power spectrum measurement, the unit of measurement remains consistent. The negative sign denotes an inverted cosine wave.

If the first inequality is satisfied, the model predicts that the class is mirthful laughter; otherwise, it predicts it as non-mirthful laughter. This tree exhibits various branching determiners until the third depth level is reached. It is important to note that at this depth, there are no threshold comparisons involving different acoustic properties (features); we only present the third-level depth. From this tree above, it is evident that the following low-level acoustic properties, including mel-frequency cepstral coefficient, spectral features, auditory spectrum, and fundamental frequency, are key determiners for identifying mirthful laughter within the diverse duration dataset.

The decision tree depicted below is constructed for binary classification, distinguishing between non-discourse laughter (“Not [laugh]-Discourse”) and discourse laughter (“[laugh]-discourse”) within the fixed duration dataset (cf.Figure.4.3). The fundamental frequency determines the root node of this tree. When a sample’s fundamental frequency is less than or equal to 291.659 Hz, the predominant utterance type is classified as non-discourse laughter. If this condition is met, the model predicts the class as non-discourse laughter; otherwise, it predicts it as discourse laughter. This decision tree proceeds with different branching

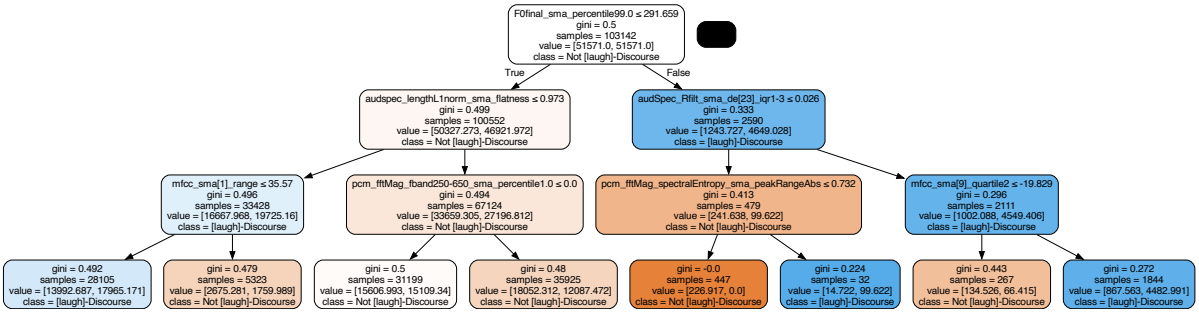


Figure 4.3: Feature selection process of discourse laughter in fixed duration dataset

determiners until the third depth level is reached. Notably, no threshold comparisons with various acoustic properties (features) are presented, as only the third-level depth is showcased. From the above tree, it is evident that several low-level acoustic properties play crucial roles in determining discourse laughter within the fixed-duration dataset. These properties include fundamental frequency, auditory spectrum, mel-frequency cepstral coefficient, and spectral features.

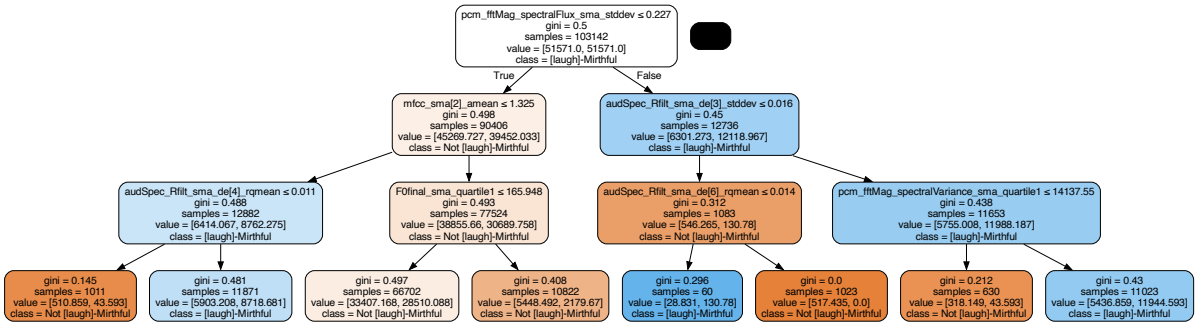


Figure 4.4: Feature selection process of mirthful laughter in fixed duration dataset

The decision tree depicted above is constructed for binary classification between non-mirthful laughter (“Not [laugh]-Mirthful”) and mirthful laughter (“[laugh]-Mirthful”) within the fixed duration dataset (cf. Figure.4.4). At the root node, spectral features serve as the primary determinant. Mirthful laughter is identified as the predominant utterance type if a sample’s spectral features measure less than or equal to 0.227 Hz. If this condition is met, the model classifies the sample as non-mirthful laughter; otherwise, it categorizes it as mirthful laughter. This decision tree proceeds through various branching determinants until the third depth level is revealed. It is important to note that no threshold comparisons are made with different acoustic properties (features) beyond the third-level depth presented here. From the analysis of this decision tree, it becomes evident that low-level acoustic properties such as spectral features, mel-frequency cepstral coefficients, auditory spectrum, and fundamental frequency play pivotal roles in determining mirthful laughter within the fixed duration dataset.

4.2.2 Quantitative analysis of discourse laughter in a varied duration dataset using decision trees

By utilising a decision tree model, we can ascertain the contribution of each feature to the response variable through its feature importance value. Features with a feature importance greater than zero are selected, signifying their positive contribution to the response variable.

Figure.4.5 displays seven features meeting the aforementioned criteria. Subsequently, by selecting the acoustic property name, a ranked list within the diverse duration dataset provided by discourse laughter is presented as follows: spectral features, mel-frequency cepstral coefficients, auditory spectrum, and jitter.

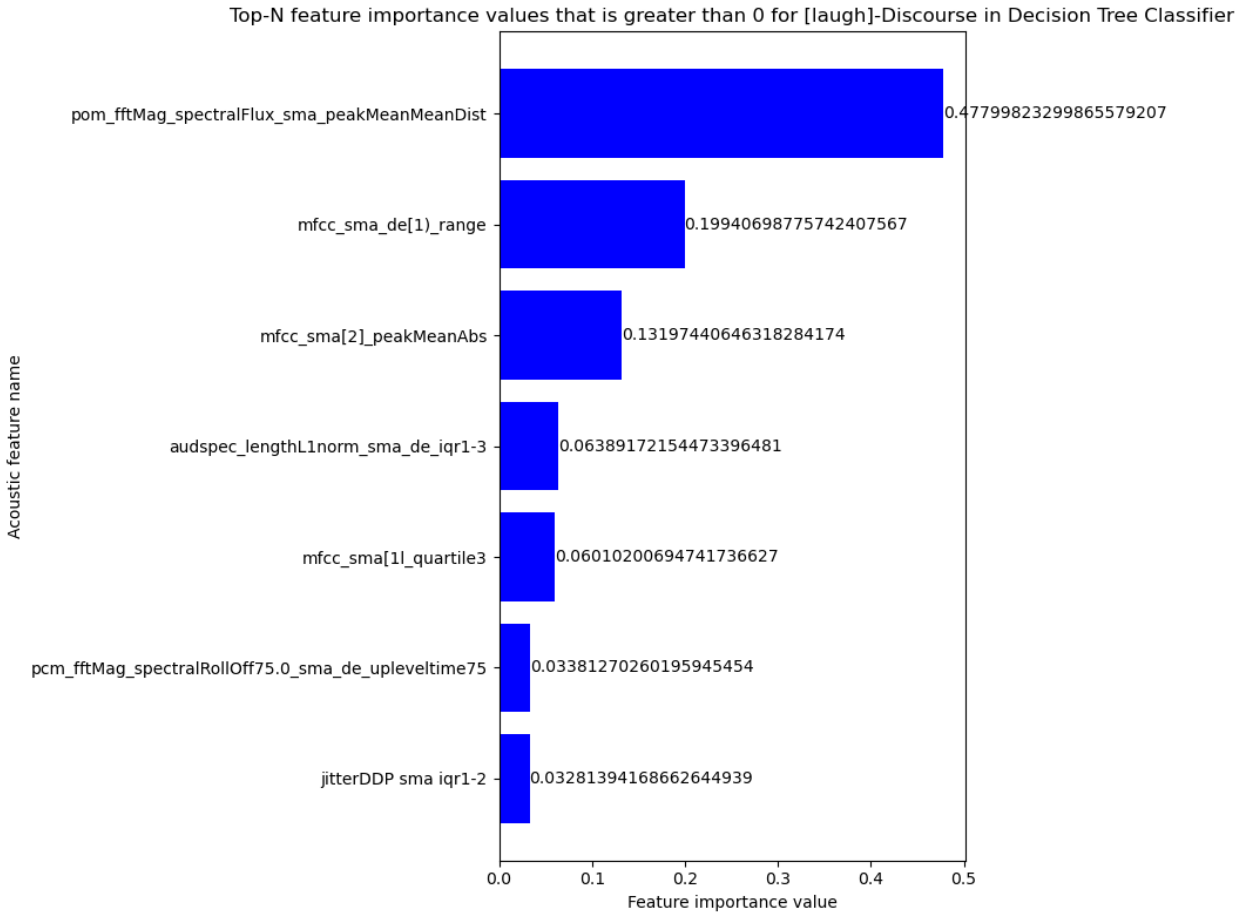


Figure 4.5: Feature importance ranked by decision tree model given by discourse laughter in varied duration dataset

4.2.3 Quantitative analysis of mirthful laughter in a varied duration dataset using decision trees

Figure.4.6 shows seven features that satisfy the above requirement. Then, by selecting the acoustic property, a ranking list within the diverse duration dataset, based on mirthful laughter, is presented as follows: Mel-frequency cepstral coefficients, spectral features, the auditory spectrum and fundamental frequency.

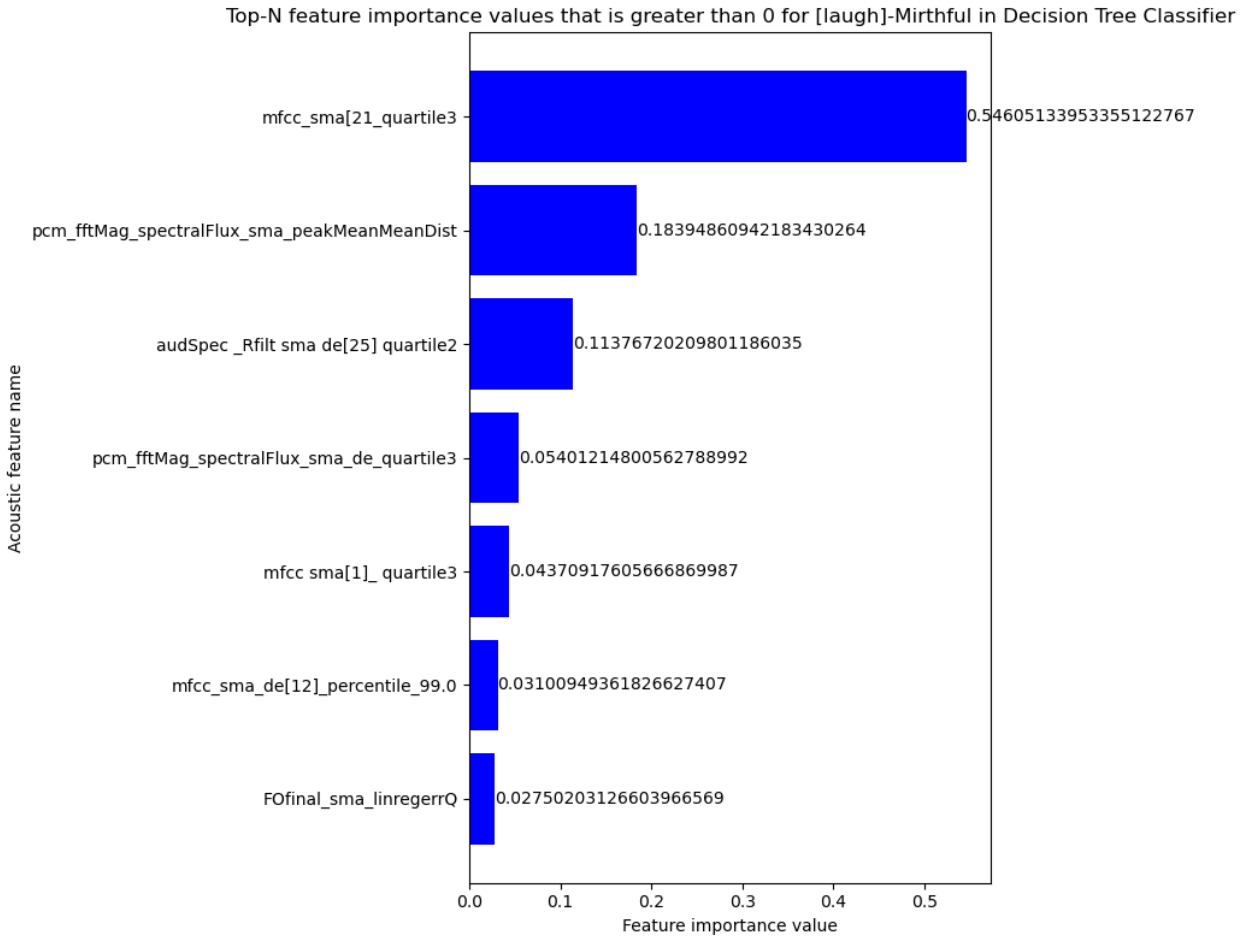


Figure 4.6: Feature importance ranked by decision tree model given by mirthful laughter in varied duration dataset

4.2.4 Quantitative analysis of discourse laughter in a fixed duration dataset using decision trees

Figure.4.7 shows seven features that meet the aforementioned criteria. Subsequently, in the fixed duration dataset provided by discourse laughter, the ranked list based on acoustic properties is as follows: fundamental frequency,the auditory spectrum,spectral features and mel-frequency cepstral coefficients.

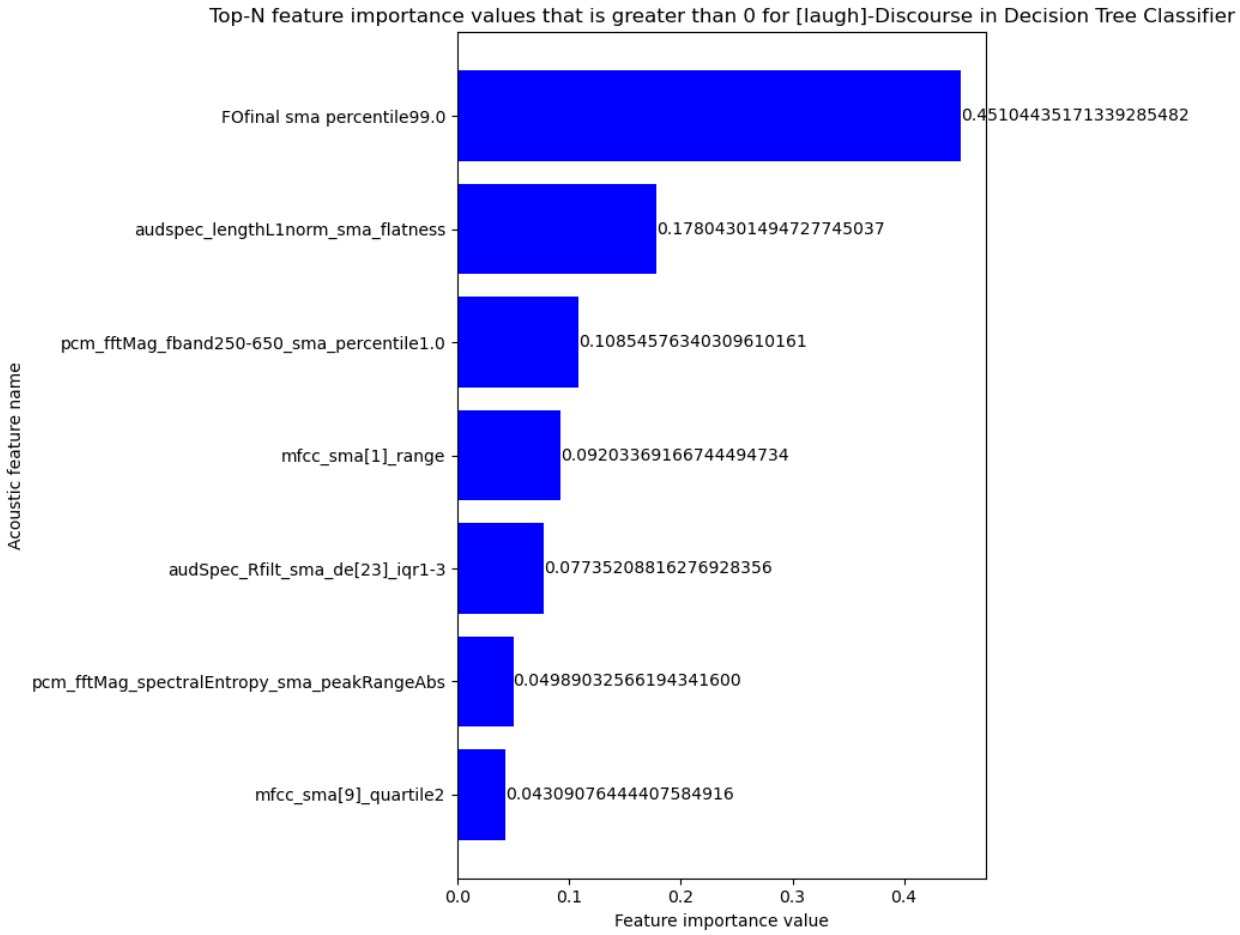


Figure 4.7: Feature importance ranked by decision tree model given by discourse laughter in fixed duration dataset

4.2.5 Quantitative analysis of mirthful laughter in a fixed duration dataset using decision trees

Figure.4.8 illustrates seven features that meet the aforementioned requirement. Subsequently, focusing on the acoustic property, a ranked list from the fixed duration dataset, based on mirthful laughter, is presented as follows: spectral features, mel-frequency cepstral coefficients, and fundamental frequency.

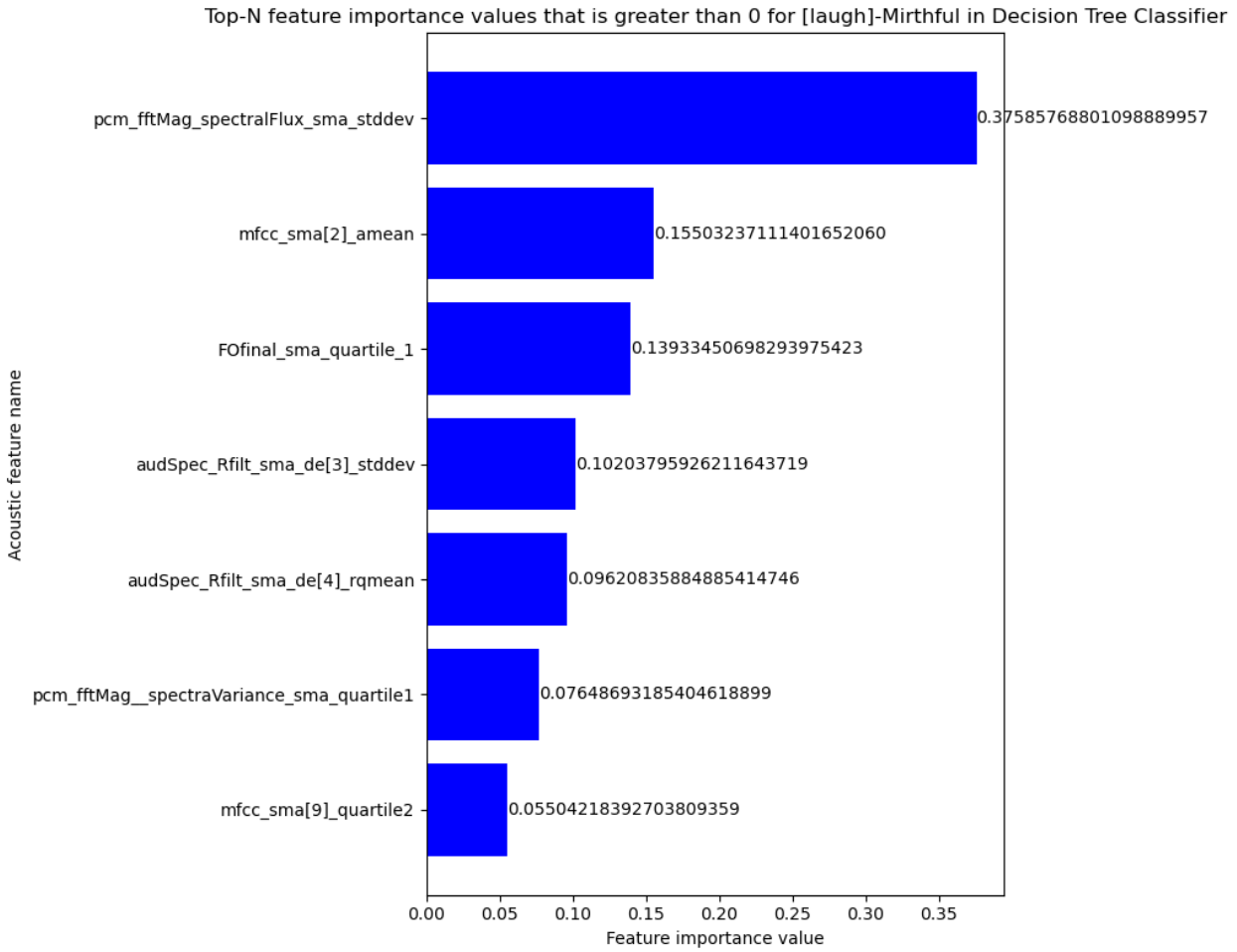


Figure 4.8: Feature importance ranked by decision tree model given by mirthful laughter in fixed duration dataset

4.3 Multi-classification results

This section presents decision tree visualisation of decision trees for multi-labelling tasks,and discerning features of discourse and mirthful laughter across varied and fixed duration dataset, ranked by the multinomial regression model.

4.3.1 Feature selection visualisation

In the context of multi-label classification, which encompasses seven distinct utterance types (’Ambiguous’, ’M’, ’S’, ’Silence’, ’[V]’, ’[laugh]-Discourse’, ’[laugh]-Mirthful’), decision trees demonstrate a capacity for feature selection tailored to multi-classification scenarios.

The decision tree depicted in Figure 4.9 pertains to a multi-classification analysis of the varied duration dataset(cf.figure.4.9) is based on a multi-classification in the varied duration dataset. The root node of this tree is the mel-frequency cepstral coefficient. Mirthful laughter predominates as the utterance type if a sample’s mel-frequency cepstral coefficient is less than or equal to 0.118 W/Hz. If this condition is met, the model predicts that the class is engaged in speaking within the discourse; otherwise, it is categorised as a merged type involving both speaking and non-laughter vocalisation. This tree utilises various branching

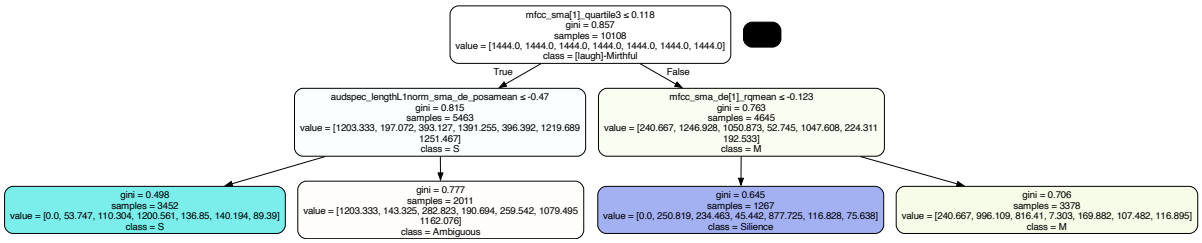


Figure 4.9: Feature selection process for all utterances in varied duration dataset

determiners until the second depth level is reached. It is important to note that there is no threshold comparison with different acoustic properties (features) at this stage, as we only present the second-level depth. From the tree above, it is evident that low-level acoustic properties, such as mel-frequency cepstral coefficient and auditory spectrum, play crucial roles in determining different utterance types in the varied duration dataset. However, it cannot identify a discriminating feature rank order for specific laughter.

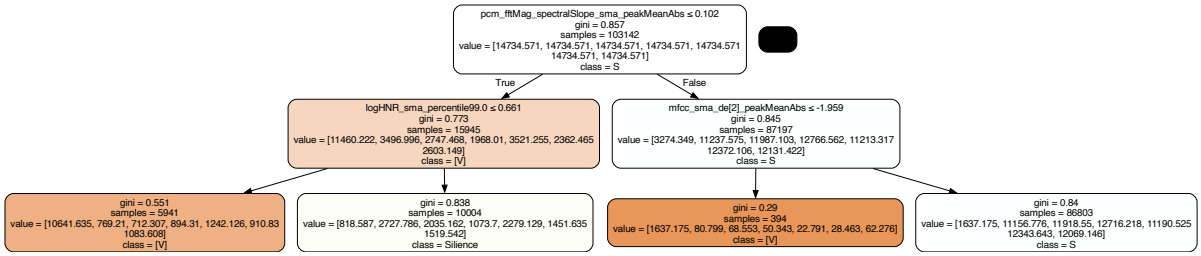


Figure 4.10: Feature selection process for all utterances in fixed duration dataset

The decision tree depicted in Figure 4.10 is constructed for multi-classification within the fixed duration dataset (cf. Figure.4.10). At the root node of this tree, spectral features serve as the primary determinant. If a sample's spectral features measure less than or equal to 0.102 Hz, the predominant utterance type is categorised as spoken words. In case this condition is met, the model classifies the class as non-laughter vocalisation within the discourse; otherwise, it predicts it as the spoken words in the discourse. The tree proceeds through various branching determiners until the second depth level is reached.

Notably, the analysis does not involve threshold comparisons with different acoustic properties beyond the second depth level. The examination of this tree reveals that low-level acoustic properties such as spectral features, harmonic-to-noise ratio, and mel-frequency cepstral coefficients play pivotal roles in determining different utterance types within the fixed duration dataset. However, it's important to highlight that the tree does not provide a discerning feature rank order specifically for laughter.

4.3.2 Quantitative analysis of discourse laughter in a varied duration dataset using multinomial regression

Using a multinomial logistic regression model, this approach constructs a matrix detailing each feature's coefficient with respect to every response variable within a multi-label target. Through analysis of various laughter instances, features are selected based on coefficients greater than zero, signifying a positive contribution to the corresponding response variable.

There are 36 features whose coefficients are greater than 0. We selected the first five features to present in the main text.

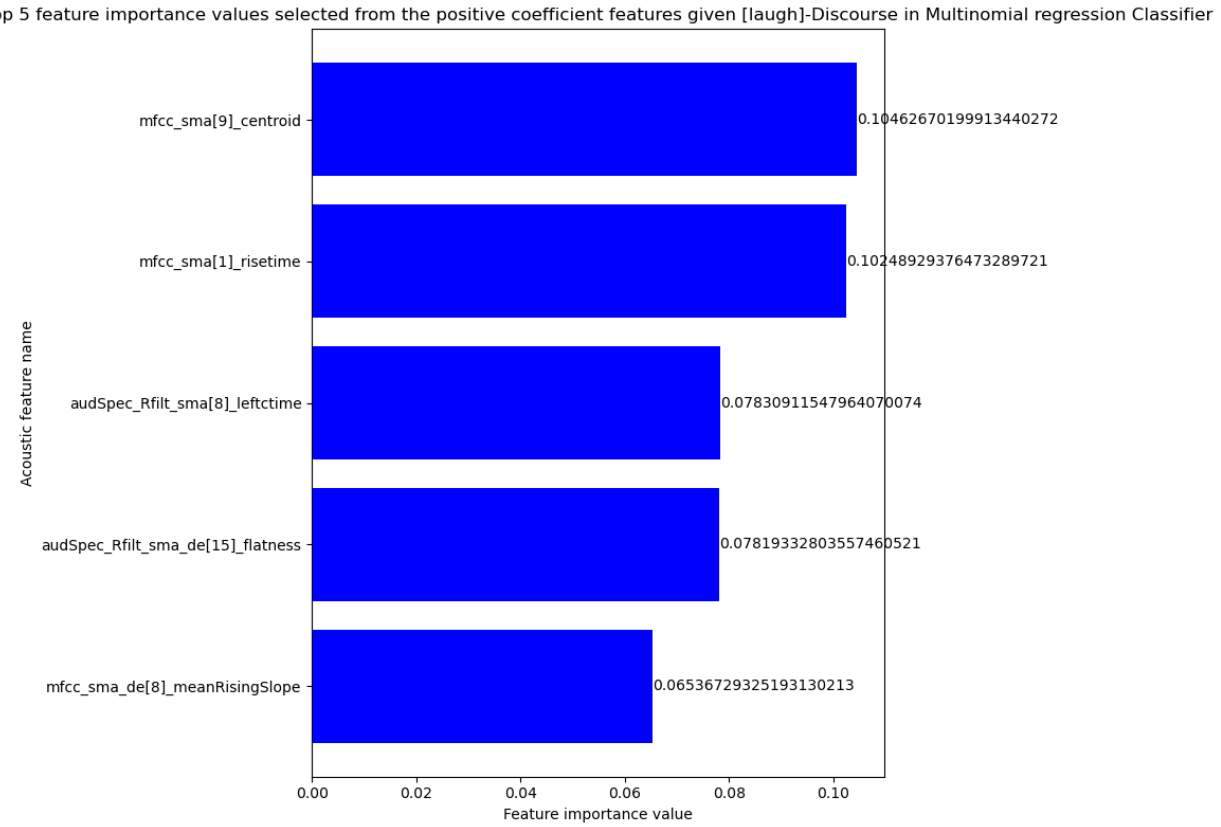


Figure 4.11: Top-5 feature importance ranked by multinomial regression model given by discourse laughter in varied duration dataset

Figure.4.11 shows the first five features that meet the aforementioned criteria. Subsequently, prioritising the acoustic property name, a ranked list within the varied duration dataset provided by discourse laughter is presented as follows: mel-frequency cepstral coefficients and auditory spectrum.

4.3.3 Quantitative analysis of mirthful laughter in a varied duration dataset using multinomial regression

There are 49 features with coefficients greater than 0. We have chosen to present the first five features to present in the main text.

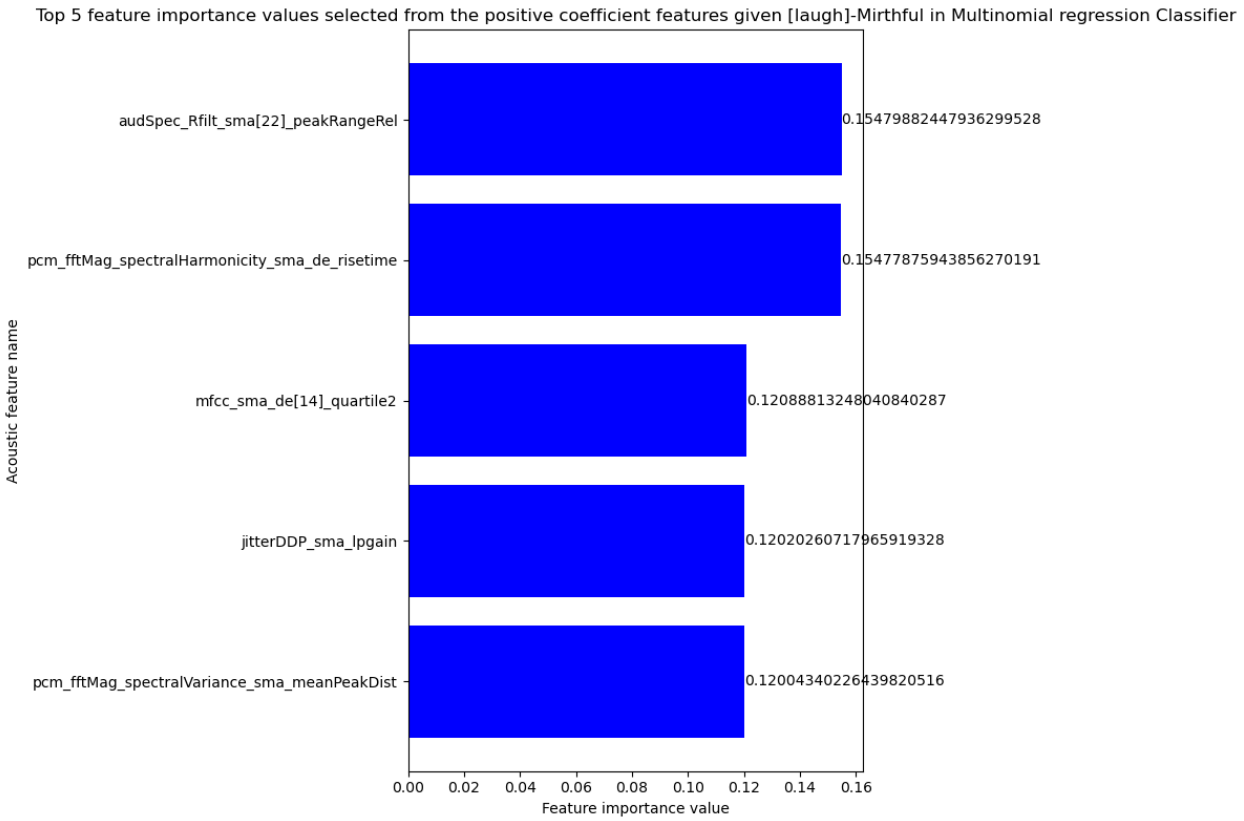


Figure 4.12: Top-5 feature importance ranked by multinomial regression model given by mirthful laughter in varied duration dataset

Figure.4.12 shows the initial five features that meet the aforementioned criteria. Subsequently, focusing on the acoustic property name, a ranked list within the varied duration dataset provided by mirthful laughter is presented below: auditory spectrum, spectral harmonicity, mel-frequency cepstral coefficients, jitter and Spectral Feature.

4.3.4 Quantitative analysis of discourse laughter in a fixed duration dataset using multinomial regression

There are 52 features with coefficients greater than 0. In the main text, we chose to present the first five features.

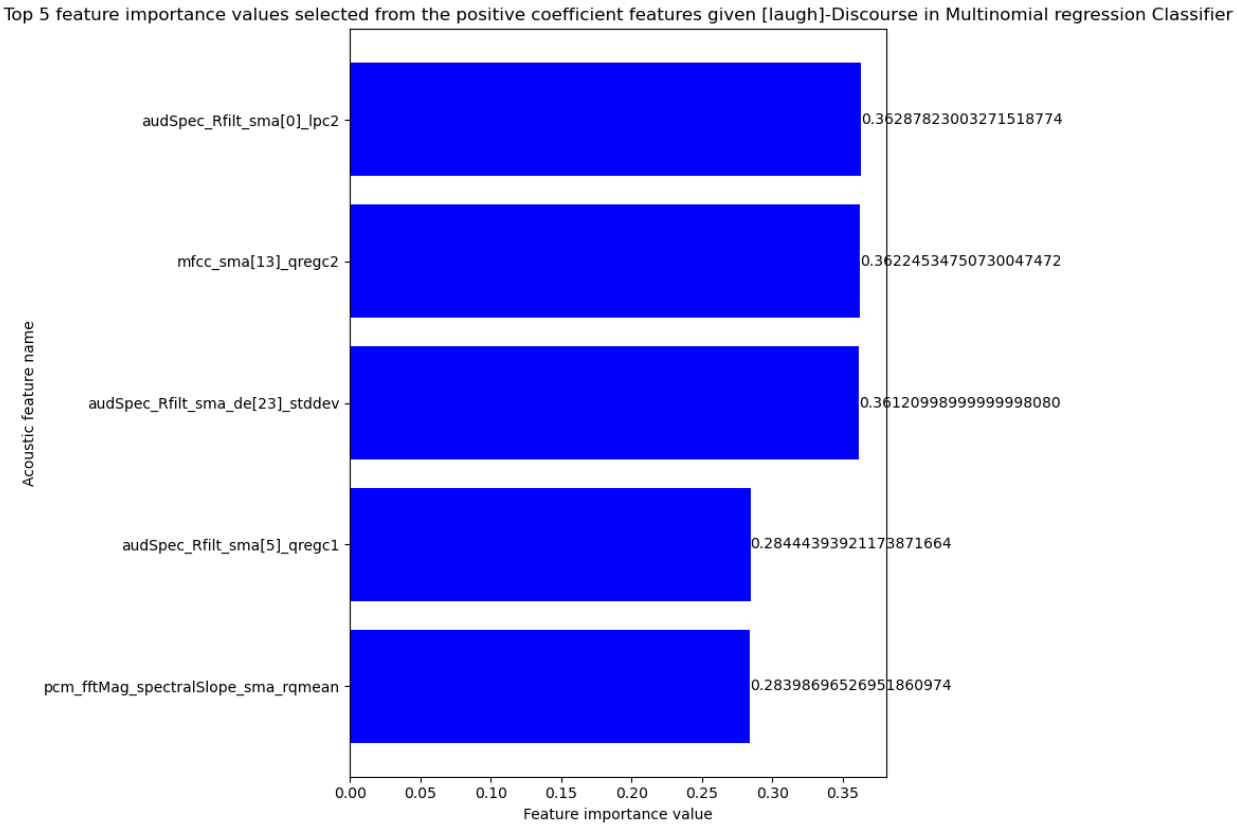


Figure 4.13: Top-5 feature importance ranked by multinomial regression model given by discourse laughter in fixed duration dataset

Figure.4.13 depicts the top five features that meet the aforementioned criteria. Subsequently, in the fixed duration dataset provided by discourse laughter, a ranking based on the acoustic property names is presented as follows: auditory spectrum,mel-frequency cepstral coefficients and spectral features.

4.3.5 Quantitative analysis of mirthful laughter in a fixed duration dataset using multinomial regression

There are 60 features with coefficients greater than 0. We have chosen to highlight the first five features in the main text.

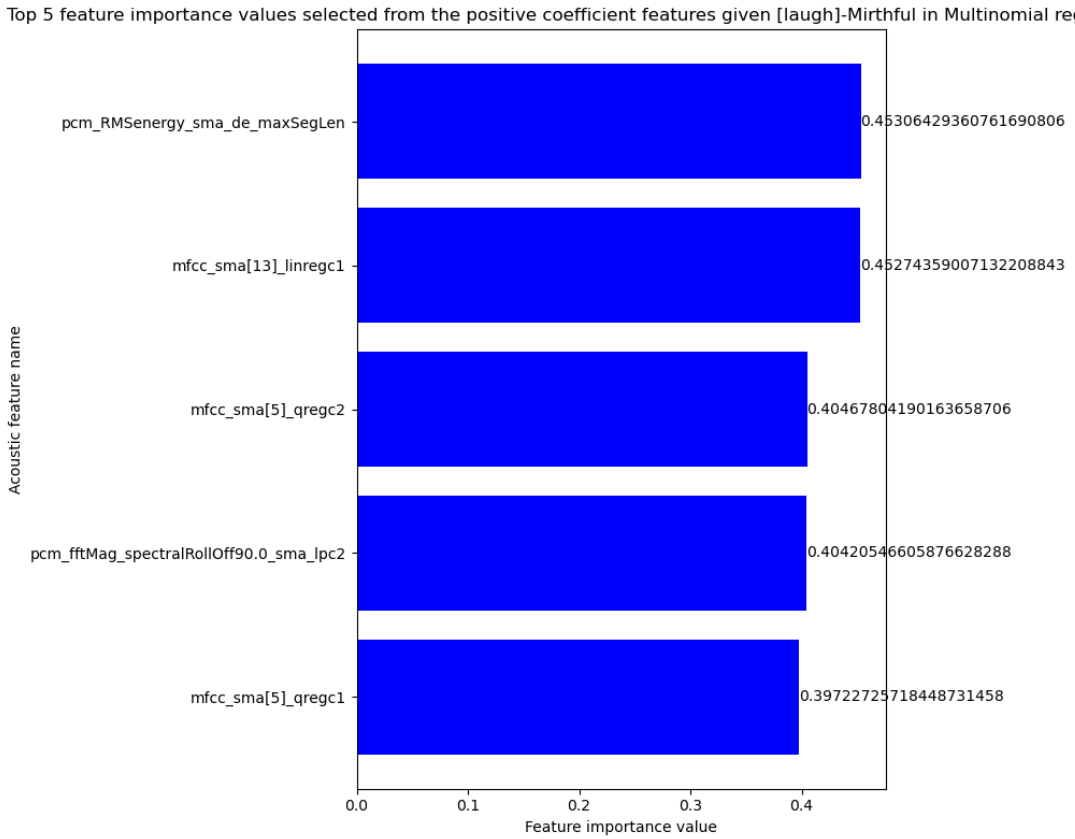


Figure 4.14: Top-5 feature importance ranked by multinomial regression model given by mirthful laughter in fixed duration dataset

Figure.4.14 shows first 5 features that satisfy the above requirement. Then, focusing on the acoustic property name, a ranked list in the fixed duration dataset provided by mirthful laughter is presented as follows:root-mean-square signal frame energy,mel-frequency cepstral coefficients and spectral feature.

4.4 Summary of discriminating features

To discern the distinctive features within different laughter categories across varied and fixed duration datasets, we utilised website[1] to retrieve the names of low-level acoustic features for each category. We then isolated the acoustic features, excluding functional components such as “sma”, to enhance clarity. The discriminating features are summarised in rank order in the table.4.1 below.

Table 4.1: The summary of discriminating acoustic properties towards different model given different laughter per dataset type

Classifier	Dataset type	Discriminating properties of Discourse laughter	Discriminating properties of Mirthful laughter
Decision tree	Varied duration dataset	**Spectral Features**	**Spectral Features**
		Mel-frequency cepstral coefficients	**Mel-frequency cepstral coefficients**
		Auditory Spectrum	**Fundamental frequency**
		Jitter	
	Fixed duration dataset	**Fundamental frequency**	**Spectral Features**
		Auditory Spectrum	**Auditory Spectrum**
		FFT	**Fundamental frequency**
		Mel-frequency cepstral coefficients	
		Spectral Features	
Multinomial logistic regression	Varied duration dataset	**Mel-frequency cepstral coefficients**	**Auditory Spectrum**
		Pulse code modulation	**Fundamental frequency**
		Auditory Spectrum	**Jitter**
			Pulse code modulation
	Fixed duration dataset	**Auditory Spectrum**	**Auditory Spectrum**
		Mel-frequency cepstral coefficients	**Mel-frequency cepstral coefficients**
		Pulse code modulation	Pulse code modulation
		Jitter	**Jitter**

The table.4.1 above highlights common discriminating features in two dataset, showing that the most dominant acoustic properties include the auditory spectrum,spectral features, fundamental frequency,mel-frequency cepstral coefficients, and jitter.

4.5 Comparison between our result with Tanaka & Campbell (2014)’s work

This section compares our results with those of Tanaka & Campbell (2014) in two key aspects, including identifying agreements and disagreements regarding discriminating features in both studies, and comparing

[1] https://github.com/rupafn/CulturalClassifier

the total number of instances of utterance events in each.

4.5.1 Agreement and disagreement for discriminating features in both two work

In comparison to Tanaka & Campbell (2014)'s study, which employed the Expressive Speech Processing (ESP) corpus, the domain acoustic properties include the following: the mean value of fundamental frequency (fmean), the maximum value of power (pmax), the position of the power maximum in relative percentage values (ppct), the difference between the first harmonic and the third formant (h1a3), duration of the laugh (dn), the number of calls in a bout (No.call), and pitch change between the first and the second call (F0moveAB).

Comparing our results in the table.4.1 with their results, we could notice that fundamental frequency,the auditory spectrum and spectral features are shared discriminating features in both works. Differences in the acoustic features utilised between Tanaka & Campbell (2014) study and ours could contribute to distinct discriminatory properties observed. Tanaka & Campbell (2014) employed the "Snack Speech Processing tool"" for acoustic feature extraction, developed in 1997, which offered a relatively limited selection of acoustic features. In contrast, we adopted the ComParE 2016 feature set, comprising over 6,000 low-level acoustic features. Furthermore, we evaluated each acoustic feature across a spectrum of statistical parameters, including mean, skewness, kurtosis, quartile, among others. The extensive range of acoustic features and comprehensive statistical descriptions in our approach rendered our results more nuanced, despite not achieving perfect alignment with the discriminating acoustic properties observed in their work.

4.5.2 Total number of instances of utterance event

Tanaka & Campbell (2014) conducted a study that recorded three 30-minute audio sessions featuring various types of laughter, including mirthful laughter, discourse laughter, derisive laughter, non-laughter vocalisation, and other types of utterances. They presented a 30-minute conversation between two males, and total utterance events were 7,875 instances, including non-laugh vocalisation(6,999 instances), discourse laughter(579 instances), mirthful laughter(244 instances), derisive laughter(49 instances), and other utterance type that they have not clearly noted (4 instances).

In our work, the diverse duration dataset contains continuous unique utterance events at each moment, segmented using the same start and end times for each dialogue session with the EAF file. In this dataset, our dataset consisted of a total of 14,441 utterance event instances, including discourse laughter (460), mirthful laughter (309), spoken words (7536), silence (5093), a mixture of spoken words and non-laughter vocalisation (601), non-laughter vocalisation (436), and ambiguous type (6). A comparison of the total utterance events in our varied duration dataset with that of previous work reveals a moderately noticeable difference in the quantity of laughter events, with only approximately 100 instances of variance (discourse laughter: 579 in their work compared to 460 in ours; mirthful laughter: 244 in their work compared to 309 in ours). Regarding the remaining utterance events, we have 13,726 instances, excluding the two types of laughter, whereas their work reported 7,052 instances. Consequently, we have a greater number of negative samples compared to their work.

The fixed-duration dataset has 147,347 instances of utterance. It emerges from each row utterance event from ELAN CSV to align single or multiple rows from 200 milliseconds constant duration in Opensimile CSV, resulting in a large capacity (around 15 GB). This relatively narrow segmentation(200 ms) in our setting

results in finer extracted features compared to the varied duration dataset (ranging from around 10 to 10000 ms). Short-time segmentation is more conducive to capturing variations in laughter acoustic properties, as it allows for more frequent changes in the original laughter instances within a shorter time frame. Additionally, it's worth noting that the total number of utterance events in our dataset far exceeds that of Tanaka & Campbell (2014), even though the fixed duration dataset has the same total number of utterance events as its varied duration counterpart.

4.6 Results from interaction among acoustic properties

To demonstrate the interaction between various acoustic features, we initially employ regression analysis to ascertain which acoustic features account for the variability in laughter across varied and fixed duration datasets. Subsequently, we illustrate the correlations among the discriminative features generated by the decision tree model in the network graph, highlighting the interplay of these discriminative features.

However, this visualisation has not been tested with the decision tree model for feature correlation analysis in multinomial regression as, due to the capacity issues, we have removed all discriminating features from the positively correlated features that could potentially correlate.

It is important to note that the edges between nodes signify the correlation between these two nodes as determined by the Wilcoxon signed rank test. In the second subsection of this section, we exclusively present the correlation between discriminating features correlation generated by decision tree model in both two types of duration datasets.

4.6.1 The summary of regression analysis

Given the relatively large capacity of the fixed duration dataset, we utilised a subset of features with weak correlations. Conversely, in the varied duration dataset, we incorporated all 6,373 acoustic features. Under this configuration, we assessed features whose p-value fell below the predefined alpha threshold of 0.05. This signifies that these features could elucidate the variance of the response variable, laughter. In this context, the specific numerical value of a feature's significance is immaterial; our primary aim is to identify significant features.

Table 4.2: The summary of significant acoustic properties yield by ordinary least squares

Dataset type	Significant properties of Discourse laughter	Significant properties of Mirthful laughter
Varied duration dataset	Auditory Spectrum	Auditory Spectrum
	Mel Frequency Cepstrum Coefficient	Pulse code modulation
	Pulse code modulation	Mel Frequency Cepstrum Coefficient
	Shimmer	Fundamental frequency
	The ratio of the energy of harmonic signal components to the energy of noise	The ratio of the energy of harmonic signal components to the energy of noise
	Fundamental frequency	Shimmer
	Jitter	Jitter
Fixed duration dataset	Auditory Spectrum	Auditory Spectrum
	Mel Frequency Cepstrum Coefficient	Mel Frequency Cepstrum Coefficient
	Pulse code modulation	Pulse code modulation

The table above.4.2 illustrates that across different type duration dataset, several acoustic features notably account for the variance in acoustic laughter. For the varied duration dataset,these features include auditory spectrum, mel-frequency cepstrum coefficients,jitter, and harmonic-to-signal ratio. Conversely, for the dataset with fixed durations, significant explanatory features for acoustic laughter variance comprise auditory spectrum, mel-frequency cepstral coefficients, and pulse code modulation.

4.6.2 Feature correlation visualisation in decision tree model

To visualise feature correlation, we used a network graph from the Python package to present the top-N features in the adjacency feature rank list. We have showcased the top 10 distinguishing properties, a slightly higher number compared to the typical count of distinguishing properties generated by different models across various datasets (around 7 features). This choice allows for a more comprehensive exploration of correlations.

In this subsection, we focus on the top 10 adjacency feature correlations within discourse laughter in the varied duration dataset. The remaining correlations are detailed in the “Results Chapter Supplement Material” in the appendix.A1.3.

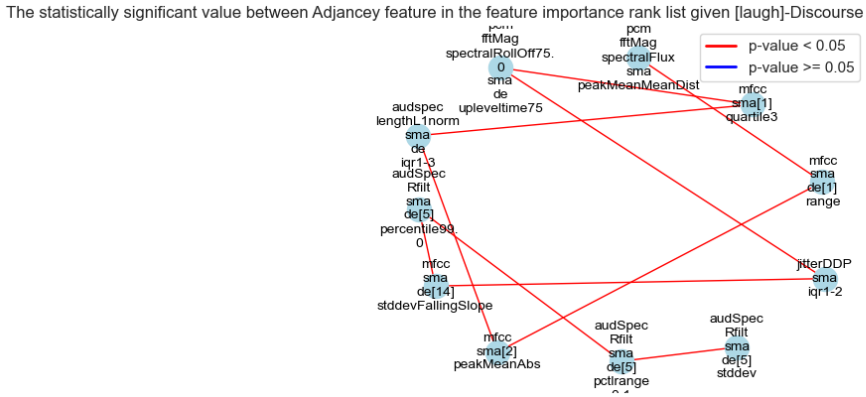


Figure 4.15: Top-N adjacency features correlation given discourse laughter in varied duration datasets.

In the above figure.4.15, each node represents an adjacency feature ranked by the decision tree model, while edges depict the correlation values between these adjacency features determined by the Wilcoxon signed-rank test. To illustrate varying levels of significant correlation between adjacency features, we use red to indicate a significant correlation ($p < 0.05$) and blue to denote a non-significant correlation ($p >= 0.05$). Additionally, the most discriminating feature (the auditory spectrum) is positioned at the top centre of the figure.

Observing this figure, it becomes evident that each selected discriminating feature correlates with the subsequent adjacency feature in the rank list. This phenomenon may be attributed to synchronised variations among these discriminating features.

Notably, the measurement of adjacency feature correlation follows a consistent pattern across the remaining three analyses: mirthful laughter in the varied duration dataset(cf.figure.A1.3), discourse laughter in the fixed duration dataset(cf.figure.A1.4), and mirthful laughter in the fixed duration dataset(cf.figure.A1.5).

5 Evaluation

The methodology chapter has provided a rationale for our chosen methods and confirmed the accuracy of the dataset. The evaluation chapter assesses the machine learning model using classification accuracy and Cohen's kappa coefficient. While there are numerous metrics available for assessing a machine learning model's performance, we opted for these two specific metrics for several reasons. Firstly, by employing classification accuracy, our results can be directly compared with Tanaka and Campbell's (2014) work, as both studies involved the classification of discourse and mirthful laughter, despite differences in sample size. Additionally, the instances of laughter in our dataset, occurring in varied and fixed durations, are significantly fewer than non-laughter instances. Consequently, Cohen's Kappa coefficient addresses this imbalance issue, ensuring that the assessment score is more justified.

Additionally, this section examines the dynamics of specific discriminating properties for three types of utterance events in the time domain, including discourse laughter, mirthful laughter, and spoken words. The objective is to assess the level of randomness exhibited by these utterance events. Furthermore, we investigate acoustic laughter duration and topic termination patterns to align with certain assertions made in related literature.

The last section aims to determine whether the particular session/participant demonstrates distinctive patterns in discriminating properties compared to others, specifically concerning laughter. This evaluation employs both quantitative and qualitative assessments. Initially, quantitative assessment is utilised to gauge the model's performance. Conversely, qualitative assessments are employed in the following sections, utilising diverse duration datasets incorporating discriminating acoustic properties generated by the decision tree model. These experiments are designed to either validate or refute claims posited in the literature.

5.1 Machine learning model performance

The metrics employed to evaluate the machine learning model in this study include classification accuracy and Cohen's Kappa coefficient to assess its performance. Notably, our dataset comprises more non-laughter utterance types than laughter types. Relying solely on classification accuracy may not provide a comprehensive evaluation, as this metric encompasses all label situations. Therefore, Cohen's Kappa coefficient is utilised alongside the confusion matrix to scrutinise and address the imbalance issue.

Classification accuracy We used classification accuracy to validate the testing set on both models at each duration dataset. In binary classification, the response label contains two types: "Discourse laughter" and "Non-discourse laughter." In the table below, we used "laugh]-discourse" to represent the positive label in binary classification. However, for multi-classification, we provide both overall accuracy and accuracy for

each utterance.

For classification accuracy tested on the decision tree model, Table 5.1 shows classification accuracy on the varied duration dataset is relatively higher than the fixed duration counterpart in the corresponding laughter type, showing around 75% accuracy on the varied duration dataset and around 70% accuracy in the fixed duration dataset given discourse laughter. Accuracy for binary classification in mirthful laughter in varied duration datasets (around 84%) still beats discourse laughter counterpart (around 77%).

Even though classification accuracy is relatively higher than Tanaka and Campbell's (2014) results, classification accuracy alone does not explain anything. The score evaluated by classification accuracy only presents the ratio of correct predicated samples of the total samples. Reliability might decrease if the classification result is solely trusted, as the predicated samples include positive and negative samples. In our dataset, instances of laughter are much fewer than instances of non-laughter.

Another reason why classification alone is unreliable is evident in Tanaka and Campbell's (2014) work. They conducted several classification tasks on diverse racial groups, including native English speakers and non-English speaker groups. In contrast, in our classification task, we did not consider this factor and added these factors as categorical variables to feed into the machine learning model. Considering these factors, we need other dimensional metrics to evaluate the model's performance, such as the confusion matrix and Cohen's kappa coefficient.

Table 5.1: Classification accuracy on decision tree

Duration	Main binary variable	Acc on test set
Varied duration	[laugh]-Discourse	0.756981306
	[laugh]-Mirthful	0.843295638
Fixed duration	[laugh]-Discourse	0.709241036
	[laugh]-Mirthful	0.773396675

To further the accuracy of each label's prediction, we combined confusion matrix visualisation, and this could clearly illustrate positive and negative samples. From this visualisation (cf. Figure. A1.8, Figure. A1.9, Figure. A1.10, and Figure. A1.11), laughter's true positive sample is relatively scarce due to the small number of laughter samples.

This situation also happens in the multi-label classification, and the classification accuracy test of multinomial regression is presented in Table. 5.2. The overall accuracy in varied duration datasets is doubled that of the fixed duration counterpart (around 80% versus 40%). Similarly, a similar trend emerges when it comes to laughter classification, encompassing both discourse and mirthful laughter. Each laughter classification sees significantly higher accuracy in the varied duration dataset than in the fixed duration dataset. It is worth noting that the accuracy for both laughter classifications is close to 0.1 in the fixed duration dataset. This low percentage may be attributed to a pattern where non-laughter instances dominate, leading to more negative samples and reducing the accuracy of laughter classification. This trend is further reflected in the confusion matrix in the appendix, which indicates fewer true positive samples for laughter (cf. Figure.A1.6 and Figure. A1.7).

Table 5.2: Classification accuracy on multinomial regression

Duration	Feature name	Acuracy on test set
Varied	Overall	0.816985922
	Ambiguous	**0**
	M	0.38
	S	0.86
	Silience	0.91
	[V]	0.54
	[laugh]-Discourse	0.36
	[laugh]-Mirthful	0.4
Constant	Overall	0.387625834
	Ambiguous	**0**
	M	0.25
	S	0.29
	Silience	0.46
	[V]	0.09
	[laugh]-Discourse	0.11
	[laugh]-Mirthful	0.07

From the above table, it is evident that regardless of whether it's a varied or a fixed duration dataset, the precision of the ambiguous type is consistently zero. The ambiguous type constitutes only a minuscule portion of the overall dataset, accounting for 6 out of 14,441 instances in the varied duration dataset and 30 out of 147,347 instances in the fixed duration dataset. Subsequently, we employed random spoliation on the dataset to generate both training and testing sets. Following these operations, it's highly probable that no instances of the ambiguous type appear in the testing set, resulting in zero precision for both datasets.

Cohen's kappa coefficient Based on the relatively small sample of laughter instances, the Cohen Kappa coefficient was adopted to measure imbalanced data classification performance more accurately than classification accuracy. Table 5.3 and Table 5.4 show the Cohen kappa coefficients of the decision tree.

Table 5.3 shows that the decision tree model generates most values less than 0.02 in both varied duration and fixed duration datasets, indicating significant agreement. However, the Cohen kappa coefficient on the multinomial regression-generated dataset with varied duration is around 0.7(cf. Table 5.4), indicating substantial agreement. However, that value in the fixed duration is around 0.02.

As McHugh (2012) suggested, the Cohen Kappa coefficient ranges from 0.41 to 0.61, indicating moderate alignment between the prediction label and the true label. Additionally, for measuring the value of another numeral range in our results, the value of the Cohen Kappa coefficient ranges from 0.1 to 0.2, indicating a slight alignment between the prediction label and the true label.

Based on these criteria, our results of both classifiers could be a better match. But this does not mean our work is worse, and an imbalanced sample is the central issue causing this phenomenon.

Table 5.3: The Cohen kappa coefficient of decision tree

Duration	Main binary variable	Cohen Kappa value on test set
Varied	[laugh]-Discourse	0.132194302
	[laugh]-Mirthful	0.123472128
Fixed	[laugh]-Discourse	0.011690505
	[laugh]-Mirthful	0.016283401

Table 5.4: Cohen kappa coefficient for multinomial regression

Duration	Cohen Kappa coefficient for multinomial regression model on test set
Varied	0.699875026
Fixed	0.098180117

5.2 Inspection of some discriminating properties' internal structure generated by the decision tree model by given utterance event tested on varied duration dataset

This section intends to conduct a small experiment to assess the randomness of certain discriminating properties generated by the decision tree model in a binary labelling task within selected utterance events, such as mirthful and discourse laughter. This experiment will be tested on datasets of varied duration, as this dataset preserves unique utterance events temporally. Three stand-alone experiments presented here align with claims from the literature.

5.2.1 Discriminating features comparison between laughter and spoken words

Previous research has indicated the internal structure of laughter and spoken words. Koutsombogera & Vogel (2022) state that the distribution of discourse is more systematic than its mirthful counterpart, and discourse laughter is usually associated with topic termination. Dunbar (2014) also found that laughter and speech have some shared acoustic properties, such as prosody. By combining the statement from two works(Dunbar, 2014; Koutsombogera & Vogel, 2022), a claim could be speculated, stating that the internal structure of mirthful laughter has a certain level of randomness compared to related factors in spoken words and discourse laughter.

To validate this assertion, we conducted experiments to confirm or refute this hypothesis. In our prior research, we identified the distinguishing characteristics of mirthful and discourse laughter, while the distinguishing properties of spoken words still need to be identified. In this small-scale study, we utilised a decision tree model to perform binary labelling tasks on datasets of varying duration, aiming to compare the outcomes with laughter under similar conditions.

Figure 5.1 presents the top N discriminating features that positively contribute to spoken words, indicating that these features primarily involve the MFCC and auditory spectrum. This presentation aims to enhance clarity.

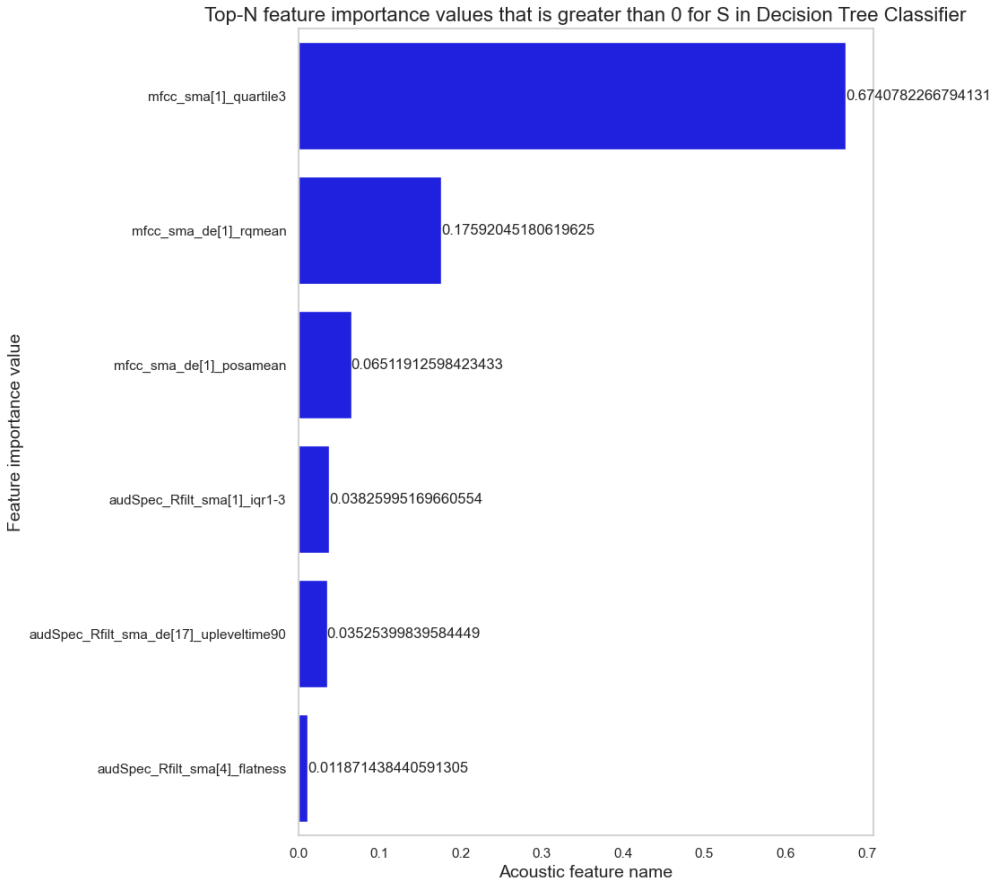


Figure 5.1: Feature importance ranked by decision tree model given by spoken words in varied duration dataset

Three distinct experiments were conducted to investigate whether the pattern of discourse laughter and spoken words leans more towards rhythmic than mirthful laughter. These experiments encompassed comparing the time gaps between utterances and examining the statistical significance of three types of utterances.

The objective of the four experiments is as follows.

- The first experiment aims to observe the level of randomness across three types of utterances by examining the time gap between the end time of the previous moment and the start time of the current moment within specific sessions, focusing on particular acoustical properties.
- The second experiment presents the duration of each moment in three utterance types that progressed in the whole session flow.
- The third experiment presents statistical information for three utterance types, including the median, mean, and standard deviation, and identifies pairwise statistical significance among them.

Here are the reasons why we designed these three experiments. We aim to assess the level of randomness exhibited by mirthful laughter, discourse laughter, and spoken words. We utilise columns such as 'Start Time -ms', 'End Time -ms', and 'Duration -ms' to analyse the temporal dynamics that discriminate among these three types of utterances. Different allocations at timestamps may reflect the randomness of specific utterance types. These three experiments encompass various aspects of verification.

The first experiment examines the time span from the end time of a previous utterance event to the start time of the next moment, which can provide insights into the frequency of specific utterances.

The second experiment independently plots the time flow duration for each continuous moment. This approach allows us to observe distinct patterns for each utterance type at different timestamps and to infer relationships with other types.

It is important to note that the previous two experiments were conducted only on specific sessions (session 3), potentially limiting their generalisability to some extent. Therefore, in experiment 3, we considered all sessions to explore the acoustic properties identified in the previous two experiments. This broader approach assists us in identifying general patterns within the dataset.

Comparison of utterance appearance time gap

Sequentially, we used the distinguishing properties of discourse laughter, mirthful laughter, and spoken words in time aspects. Due to variations in session duration across the 18 sessions, we selected session 3, which had a relatively short duration of around 5 minutes. In this subsection, we meticulously categorised all distinguishing properties among the three types of utterance events: discourse laughter, mirthful laughter, and spoken words. We provided an analysis of the temporal dynamics for one specific low-level acoustic property, while the remaining properties were included in the supplemental material for evaluation in the appendix chapter. Utilising our dataset, which established a continuous timeline for each utterance type along with their corresponding acoustic properties, we assessed the level of randomness in the distribution of the three utterance types by tracking the energy magnitude of specific acoustic properties in the time domain.

As we mentioned the term "energy" several times, there are reasons why we empathise with it. Our focus lies within audio processing, where an acoustic wave embodies energy over time. Various units exist to

quantify the power of distinct acoustic properties, including Hertz. Therefore, in this context, "energy" signifies the magnitude of specific acoustic attributes.

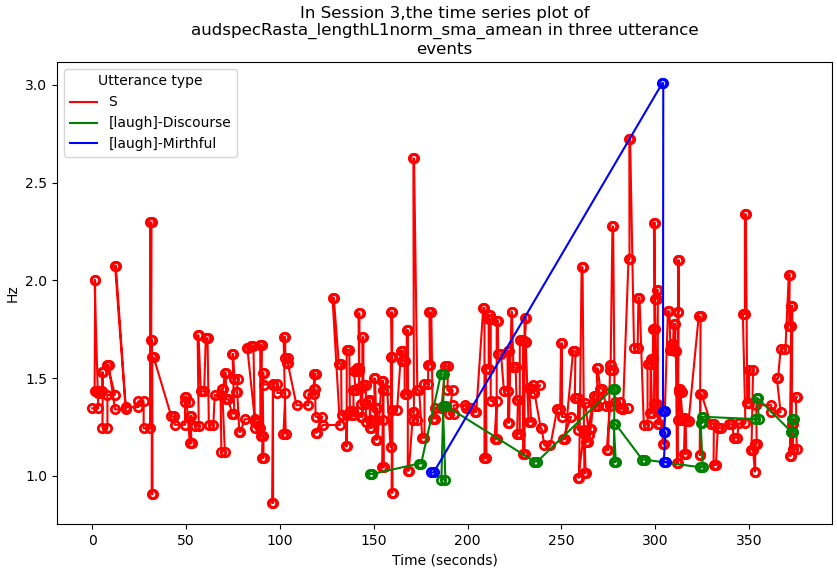


Figure 5.2: The dynamics of “Magnitude of L1 norm of Auditory Spectrum” in the session 3 for three utterance types

In this study, auditory spectrum properties were employed in session 3 to analyse the figure above, facilitating an examination of the underlying pattern behind the observed phenomenon.

In session 3, we selected “Magnitude of L1 norm of Auditory Spectrum” as selected properties, as these properties are common discriminating properties from three utterance events. From the above Figure.5.2, “S”, “[laugh]-Discourse”, and “[laugh]-Mirthful” represent spoken words, discourse laughter, and mirthful laughter, respectively. We utilised distinct colours to denote the commencement and conclusion times of each type of utterance: spoken words were represented in red, discourse laughter in green, and mirthful laughter in blue. Nonetheless, this visualisation could potentially exhibit a narrow overlap within brief intervals. The connecting line between dots illustrates the temporal gap between the end time of the previous moment and the current moment.

It was observed that between approximately 170 to 310 seconds, the pattern of three utterances becomes more pronounced compared to other time intervals. Specifically, the frequency of mirthful laughter undergoes significant variation, increasing from around 1.0 Hz at approximately 170 seconds to around 3.0 Hz at around 300 seconds. Following this increase, during the subsequent short-time continuous event, the energy of mirthful laughter rapidly decreases to 1 Hz. In contrast, the discourse and spoken word patterns within the same time windows exhibit relatively stable oscillations, with frequencies around 1.2 Hz and 1.4 Hz, respectively.

The rest of the discriminating properties pattern in the 200 s to 300 s time windows, including fundamental frequency (cf. Figure. A1.16), auditory spectrum (cf. Figure.A1.12 and Figure.A1.13), MFCC(cf. Figure.A1.14) and jitter(cf. Figure. A1.15), shows that the magnitude of mirthful laughter has a significant jump either increasing or decreasingly. In most discriminating properties, the pattern of discourse laughter constantly varied, even though the magnitude of some properties for discourse laughter changed significantly, such as f0 (cf. Figure A1.16). As for the spoken words’ pattern, this utterance oscillates at the certain value.

The substantial time span between the previous moment’s end time and the current moment’s start time for mirthful laughter signifies volatile energy levels (measured in hertz, the magnitude of the auditory spectrum) compared to other types of utterances, including discourse laughter and spoken words. Higher energy levels (in hertz for the auditory spectrum) correspond to a richer sound of laughter, suggesting that specific acoustic characteristics of laughter distinctly impact mirthful laughter, resulting in a more randomly distributed energy pattern in the time domain, characterised by significant fluctuations.

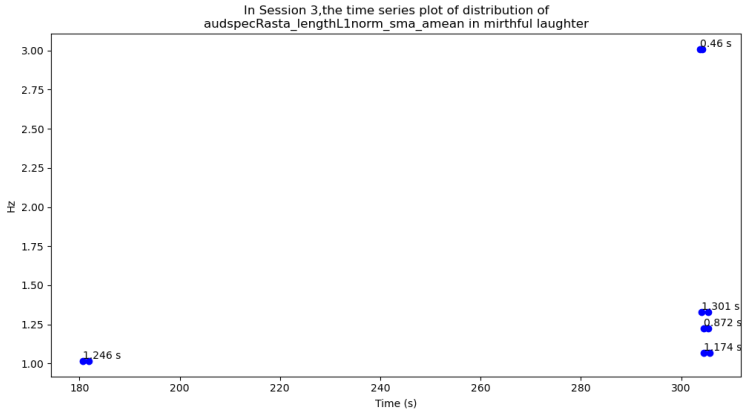
In contrast to the energy variations observed in discourse laughter and spoken words, the energy levels of these two utterance types show relatively minor changes from the end time of the previous moment to the start time of the next, as depicted in this chapter and the appendix. This observation highlights the stability of energy levels within a specific range for these two utterance types, contrasting sharply with the pronounced fluctuations observed in mirthful laughter. The relatively high level of randomness in energy variation indicates mirthful laughter, whereas rhythmic energy variations are characteristic of discourse laughter and spoken words.

Comparison of duration and distribution

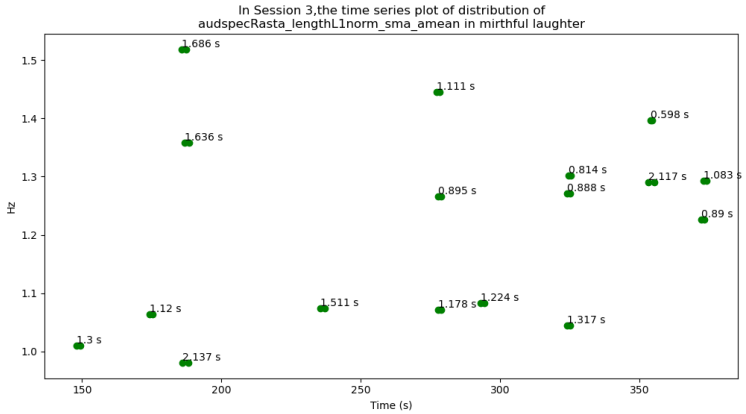
However, it is noted that Figure 5.2 does not show the duration and allocation of each utterance in terms of the auditory spectrum in the temporal flow. Hence to identify the duration distribution progress in the temporal flow, we visualise the below illustration (cf. Figure.5.3) to present the duration progression in the time flow.

The figure below consists of three subplots for mirthful laughter, discourse laughter, and spoken words, respectively; the x-axis of each figure stands for the time stamps, and the overall x-axis represents the total duration of the specific session. The y-axis stands for the energy of the particular properties. In this context, we used Hertz to describe the energy power of these acoustic properties. Additionally, each dot in the sub-figures represents the duration of particular utterance types at a given moment. By amalgamating this information, we can interpret each plot as illustrating the distribution pattern of specific acoustic properties across different utterance types over time.

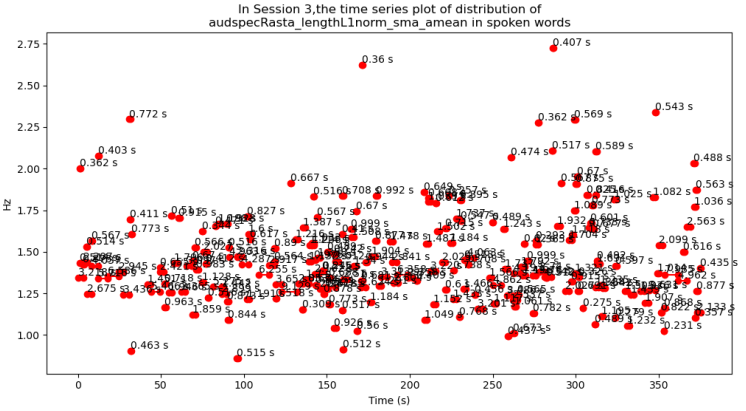
Figure 5.3 shows the energy of auditory spectrum variation regarding time flow for mirthful laughter, discourse laughter, and spoken words from the top(cf.Figure.5.3(a)),middle(cf.Figure.5.3(b)),and bottom(cf.Figure.5.3(c)). These sub-pictures represent the energy of auditory spectrum variation regarding time flow for mirthful laughter, discourse laughter, and spoken words.



((a)) Duration and distribution of mirthful laughter for auditory spectrum property



((b)) Duration and distribution of discourse laughter for auditory spectrum property



((c)) Duration and distribution of spoken words for auditory spectrum property

Figure 5.3: Duration and distribution in terms of auditory spectrum of three utterance types

To show the duration of each moment, we also add text related to each moment's duration. From these three figures, it is noted that the pattern of discourse and spoken words follow a seemly fixed trajectory, even though the quantity of spoken words is larger than its discourse counterpart, as this might be since the "MULTSIMO" dataset is a dialogue dataset and spoken words predominate the most category. Compared to an incremental variation in the pattern of discourse laughter and spoken words, the variation of mirthful

laughter follows a random oscillation from around 1 Hz at 180 s to around 3 Hz at 300 s.

Investigation of statistical significance fro three utterance types

As shown in Figure 5.3, the duration and allocation of three types of utterances differ in session three, and we need to know the situation in other sessions. Besides, this figure cannot present information related to the internal structure of each utterance in the same measurement and loses some generality to some extent. Hence, to guarantee our results, we tested the properties of the auditory spectrum for all 18 sessions using these three types of utterances to convert the. Additionally, to address the same measurement issue, we used each moment's energy of acoustic properties (Hz) divided by the associated moment's duration.

We called this ratio the normalised energy magnitude. After this conversion, the statistical information and statistical hypothesis testing are presented in the below table.

Table 5.5 shows the mean, median, and standard deviation of the normalised energy magnitude of the auditory spectrum in three utterance types. Notably, the numerical values for spoken laughter and spoken words exhibit proximity across the mean, median, and standard deviation, approximately at 0.002096 for the mean, 0.00164 Hz/ms for the median, and 0.00168 Hz/ms for the standard deviation (cf. Table.5.5). This proximity indicates that the numerical value in different statistical measurements(mean,median,std) for both utterance types is the same.

However, the numerical value in different statistical measurements (mean, median, std) of mirthful laughter has different numbers after three decimal places compared with the associated statistical measurement of spoken words and discourse laughter, even though these numerical differences between spoken words/discourse laughter and mirthful laughter are within 0.001. Such findings shed light on the comparable energy usage of these two types of utterances (discourse laughter and spoken words).

Table 5.5: The statistical information related to normalised energy magnitude of auditory spectrum in three utterance types

	Mean(Hz/ms)	Median(Hz/ms)	Standard deviation(Hz/ms)
Discourse laughter	0.002096224	0.001645573	0.001686378
Mirthful laughter	0.00167502	0.001301723	0.001289957
Spoken words	0.002096224	0.001645573	0.001686378

To further inspect more significance differences among the three types, we adopted the Wilcoxon rank sum test to assess the pairwise correlation between every two types without self- and duplicate mutual comparisons. Given that the normalised energy magnitude does not adhere to a normal distribution in most cases, we opted for non-parametric hypothesis testing, specifically the Wilcoxon rank sum test, to assess the correlation. The null hypothesis of this testing is that two pairwise variables have the same continuous distribution between each other if the statistical significance (p-value) is less than the predefined alpha value (0.05).

Table 5.6: Pairwise comparison in terms of statistical significance towards normalised energy magnitude of auditory spectrum among three utterance types

	Discourse laughter	Mirthful laughter
Mirthful laughter	4.70×10^{-6}	-
Spoken words	1	4.70×10^{-6}

Table 5.6 shows that the value between mirthful laughter and discourse laughter is less than 0.05. This indicates a statistical difference in the normalised energy magnitude of the auditory spectrum between mirthful laughter and discourse laughter, leading to the rejection of the null hypothesis.

Similarly, the statistical significance value between spoken words and mirthful laughter shows that the null hypothesis can be rejected(p-value < alpha:$4.70 \times 10^{-6} < 0.05$).

As for the comparison between spoken words and discourse laughter, the statistical significance value shows that the null hypothesis failed to be rejected (p-value >= alpha:1>=0.05), indicating that the normal energy magnitude of the auditory spectrum property for discourse laughter is not statistically different from the related property in spoken words.

Based on the above interpretation, the normalised energy magnitude of the auditory spectrum property for mirthful laughter has a different continuous distribution with the same property in discourse laughter and spoken words. In contrast, the normalised energy magnitude of discourse laughter, and spoken words has the same continuous distribution in relation to auditory spectrum. This experiment validates the similar distribution of spoken words and discourse laughter under the same conditions in one facet.

Short summary

While the findings from experiments 1 and 2 were derived solely from the selected session (session 3) and specific discriminating acoustic properties, the findings from experiment 3 encompassed all sessions and focused on particular acoustic properties. Despite these variations, these pilot studies confirm the relatively strong randomness of mirthful laughter compared to the associated patterns in discourse laughter and spoken words, aligning with the claim we seek to validate.

5.2.2 Previous event observation and topic termination verification

Koutsombogera & Vogel (2022) claimed that discourse laughter normally accompanies topic termination. To verify this statement, we select the 'CV-merge-M-L-S' column (cf. Figure.1.4) containing punctuation, which could assist us in finding the topic termination. The column after the third column is the response variable in the experiment and the simple version of the value in column three.

This assessment selects full stops, question marks, and exclamation marks to count topic termination. During the original data frame iteration, check and store the previous utterance before the specific laughter type, such as "[laugh]-Discourse" detected in the "concise merge type" column. Based on this operation, we could acquire related data and the result is presented in the below Table.5.7.

.

Note that this experiment is carried out on the varied duration dataset for all 18 sessions.

Start Time-ms	End Time-ms	Duration-ms	CV-merge-M-L-S	concise merge type
0	1375	1375	[laugh]-Discourse	[laugh]-Discourse
1375	6441	5066	V. [V] V.	M
6441	7057	616	V	S
7057	8812	1755	V	S
8812	9380	568	V	S
9380	13459	4079	V	S
13459	18149	4690	V.	S
18149	18620	471	[V]	[V]
18620	21375	2755	V	S
21375	23275	1900	V.	S
23275	29931	6656	V	S
29931	33993	4062	V, V	S
33993	35566	1573	V? V	S

Figure 5.4: 'CV-merge-M-L-S' type and 'concise merge type' in varied duration dataset

Table 5.7 presents termination occupancy in terms of terminated punctuation before the laughter event in the first column. The values in this column indicate the number of topic punctuation symbols in the "CV-merge-M-L-S" type before two types of laughter. The second value indicates the total number of laughter events across varied durations(refer to figure.3.22). The third column represents the ratio between the first and second columns, indicating the percentage of previous events before laughter, which may signal potential topic termination. In some aspects, this percentage of topic termination reflects the occurrence of topic-related punctuation before responding with laughter. However, it's important to note that topic termination may not always be explicitly indicated within our selected scope in conversation.

Table 5.7: The occurrence related to topic termination signal in our definition in both laughter(Round to four decimal places)

	Topic termination before laughter	Total number	Percentage of topic termination usage
Discourse laughter	63	460	0.1370
Mirthful laughter	30	309	0.0971

The Table.5.7 notes that the percentage for discourse laughter (0.1370) is relatively more significant than that for mirthful laughter(0.0971), which shows some degree, that discourse laughter is associated with topic termination. This moderate difference in percentage might stem from the fact that there is no multiple relationship and overlapping between the quantity of mirthful laughter (309) and discourse laughter (460).

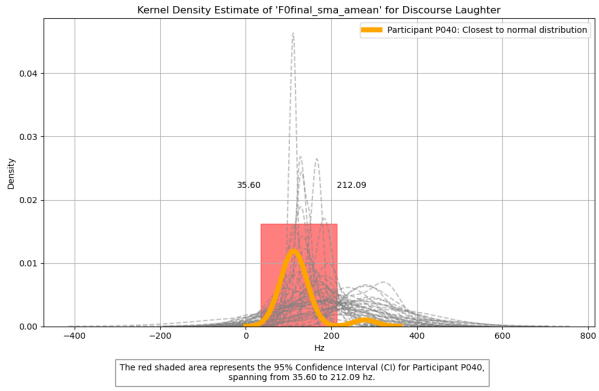
5.3 Identification the key participant/session in terms of discriminating properties generated by decision tree model in varied duration dataset

This section investigates the key participants and sessions regarding discriminating properties generated by the decision tree model in the varied duration dataset. During the experiment, participants were engaged in discerning properties across 18 sessions, with 49 participants included in the exploration. Kernel density estimation plots were utilised to illustrate the distribution and likelihood of each point, emphasising noteworthy individuals. The notable consistency observed across all participants and sessions in this experiment is reflected in the shape of the kernel density estimation plot, which closely resembles a normal distribution.

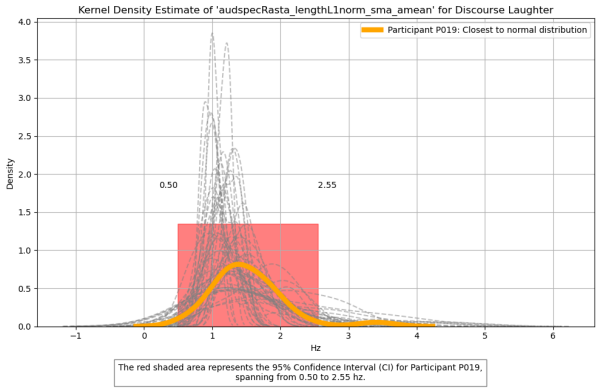
5.3.1 Identification of the key participant for given discriminating properties

To examine the primary contributors among all individuals, we focus on two acoustic properties: fundamental frequency and auditory spectrum. These properties are pivotal in distinguishing disclosure and mirthful laughter within our dataset of varied durations. Specifically, we first extract all participants from the "Participant ID" column. Then, we iterate through the dataset, isolating subsets of data corresponding to each unique participant identifier, such as "P001". During each iteration, we extract specific laughter instances from the target response column and their associated properties, including fundamental frequency. Once the operation was finished, we used this final conversion to feed into the kernel density estimation plot to draw normal distribution across 49 participants highlighting the most significant participant based on the shape close to normal distribution.

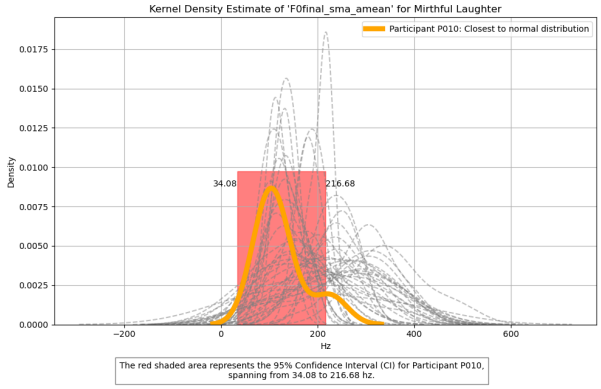
Figure.5.5 shows the key participants in each laughter for different discriminating properties. We used orange to make the key participant whose shape is close to the normal distribution while employing red to fill the 95% confidence interval and label the lower and upper bound of this range on the figure. Horizontal observation shows that the key participant has different acoustic properties regarding the same type of laughter. In contrast, in the vertical comparison, the key participant laughing different type of laughter impacts the same acoustic properties.



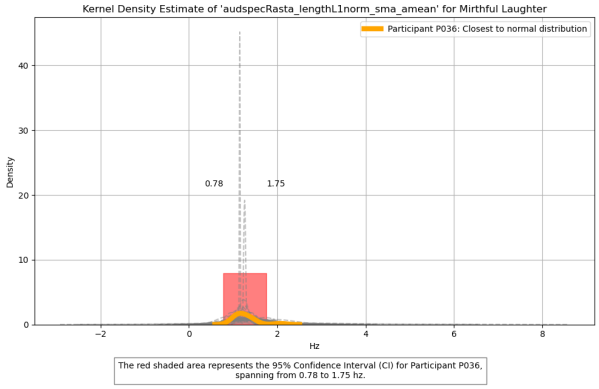
((a)) Kernel density of all participants in discourse laughter related fundamental frequency property



((b)) Kernel density of all participants in discourse laughter related to auditory spectrum property



((c)) Kernel density of all participants in mirthful laughter related to fundamental frequency property



((d)) Kernel density of all participant in mirthful laughter related to auditory spectrum property

Figure 5.5: Kernel density of all participant in both types of laughter related to fundamental frequency and auditory spectrum property

From the result, it is noted that Participant 40 stands out from all participants owing to its distinct impact on the frequency of discourse laughter. The value is clustered at 35.6 to 212.09 Hz for the fundamental frequency of discourse laughter (upper left figure) and has more data for the 95% confidence interval. The probability of this range is roughly 0.01 to 0.02. This range of information reflects the variation of fundamental frequency in the discourse laughter and explains why Participant 40 could stand out from all participants in the fundamental frequency in the discourse laughter with a specific energy range distinct from the rest of the participants

The remaining three figures have the same interpretation, even though they have different cluster ranges in different laughter for different acoustic properties. From these four subfigures, we could observe the following phenomena: Participant 19 has a distinct impact on the auditory spectrum in mirthful laughter; Participant 10 has a distinct impact on fundamental frequency in mirthful laughter; Participant 36 has a distinct impact on the auditory spectrum in mirthful laughter.

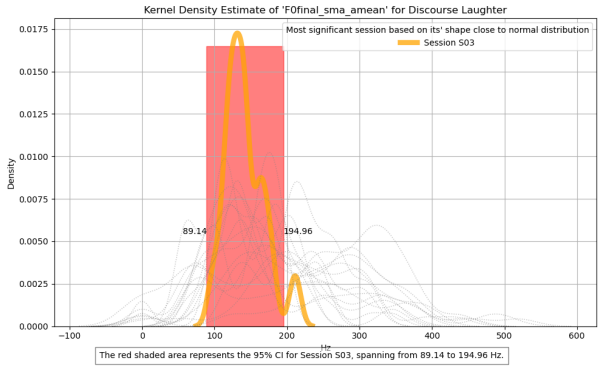
This analysis aligns with our initial anticipation that the specific participant has a distinct impact on specific laughter and stands out on particular discriminating acoustic properties by testing different laughter within all participants. Given the inherent differences in voice characteristics, such as pitch, between male and female participants, our initial hypothesis posits that each participant contributes uniquely to the acoustic profile of specific laughter types.

5.3.2 Identification of the key session for given discriminating properties

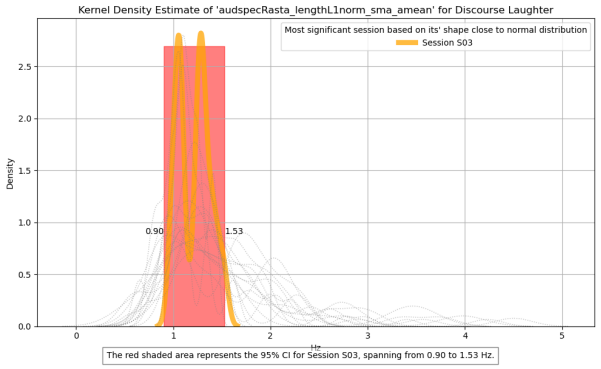
The process of identifying key themes across all 18 sessions mirrors that of identifying key participants, ensuring consistency in discerning acoustic properties and selection criteria, such as fundamental frequency and auditory spectrum. The observation approach remains consistent with previous comparisons regarding the impact of participants on laughter.

Similar to the previous key participant experiment, we used orange to make the key participant whose shape is close to the normal distribution. Additionally, we employed red to shade the 95% confidence interval and label the lower and upper bounds of this range on the figure.

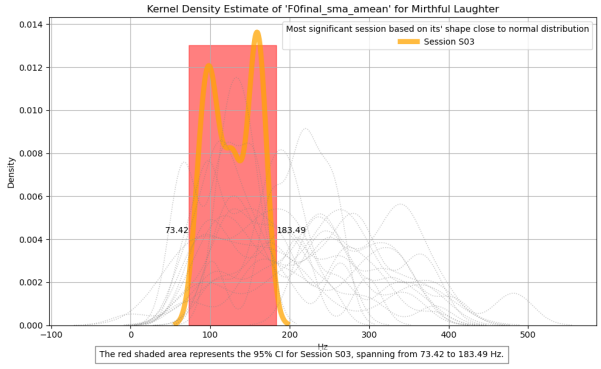
The result (cf.Figure.5.6) shows that session 3 stands out from all participants owing to having a distinct impact on the fundamental frequency of discourse laughter. The value is clustered at 89.14 to 194.96 Hz for the fundamental frequency of discourse laughter (upper left figure) and has more data for the 95% confidence interval.



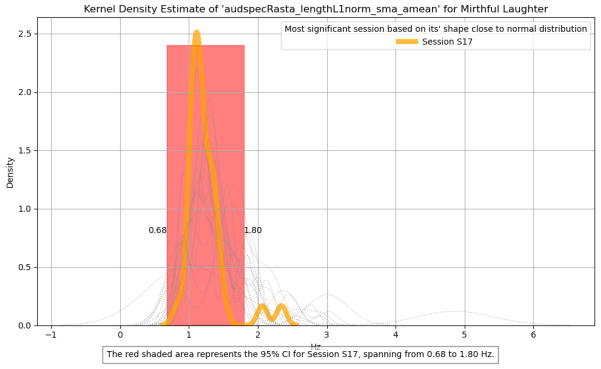
((a)) Kernel density of all sessions in discourse laughter related to fundamental frequency property



((b)) Kernel density of all sessions in discourse laughter related to auditory spectrum property



((c)) Kernel density of all sessions in mirthful laughter related to fundamental frequency property



((d)) Kernel density of all sessions in mirthful laughter related to auditory spectrum property

Figure 5.6: Kernel density of all sessions in both types of laughter related to fundamental frequency and auditory spectrum property

The remaining three figures share the same interpretation, despite featuring different cluster ranges for various acoustic properties across different types of laughter. Among these four sub-figures, it's notable that session 3 distinctly affects the fundamental frequency in mirthful laughter, as well as the auditory spectrum in discourse laughter. Meanwhile, session 17 primarily impacts the auditory spectrum in mirthful laughter.

This analysis aligns with our initial anticipation that the specific session has a distinct impact on specific laughter and stands out on particular discriminating acoustic properties by testing different laughter within all sessions. Our initial hypothesis suggests that sessions vary in duration, ranging from 5 to 10 minutes, and feature recordings of different participants' voices. Given this context, we aim to ascertain whether specific sessions distinctly influence particular acoustic properties in the laughter observed across all sessions.

6 Conclusion

The conclusion chapter encapsulates an overview of the entire process discussed in this dissertation. It reflects on the project undertaken and offers a comprehensive discussion of the results within their broader context. Furthermore, it outlines potential directions for future work.

6.1 Overview

This dissertation aimed to identify systematic differences in acoustic properties among laughter categories within the "MULTISIMO" dataset. To elucidate the complexity of this task, the motivation section within the introduction and background chapter reviewed previous research, highlighting the scarcity of datasets suitable for our research objectives. Our motivation also stemmed from the desire to replicate the work of Tanaka & Campbell (2014), given their comprehensive analysis of acoustic properties. Building upon this background investigation, the methodology chapter aimed to extract pertinent information from the raw MULTSIMO dataset to generate a dataset tailored to our research question. Subsequently, machine learning models were employed to discern features that distinguished the decision tree from multinomial regression models. The methodology encompassed dataset construction, validation, and analysis techniques. Dataset production generated both fixed-duration and varied-duration datasets, followed by rigorous verification procedures to ensure dataset accuracy. Furthermore, the chapter delineated the process by which our constructed dataset was utilised in machine learning models. Additionally, regression analysis was incorporated to pinpoint statistically significant properties explaining variance in the target variable.

The methodology chapter drew upon the foundation laid by the background chapter to devise appropriate research methods aligned with our research question, thereby guiding the experiment chapter in generating data directly relevant to answering the research question. In the Results chapter, multinomial regression and decision tree models were employed to identify discriminatory properties across two types of datasets per machine learning approach. Regression analysis was also utilised to identify significant properties capable of elucidating variance in the target variable. It is noted that fundamental frequency, mel-frequency cepstral coefficient, auditory spectrum, spectral features, and jitter emerged as common properties between machine learning experiments and regression analysis. Although the research question was only partially addressed in the results chapter, this marks a crucial stage rather than the conclusion of the investigation.

In order to elucidate the impact of these distinctive attributes within the two datasets, the evaluation chapter undertakes several preliminary studies to examine factors influencing these attributes in the varied duration dataset, thereby corroborating certain assertions found in the relevant literature. For instance, it is observed that mirthful laughter exhibits a more arbitrary distribution compared to discourse laughter and spoken words counterparts. Furthermore, we assess the efficacy of the machine learning model through classification accuracy and Cohen's Kappa coefficient. Despite achieving an overall classification accuracy of 70%, the

precision of laughter samples is relatively lower. This discrepancy is largely attributable to the imbalance between samples annotated as laughter and those labelled as non-laughter.

In the main chapter, we established the dataset following the methodology employed by Hegarty (2022) and implemented a comprehensive acoustic property identification pipeline inspired by the work of Tanaka & Campbell (2014). Our approach enhanced their methodology by integrating additional tests that were not previously considered. By employing this refined pipeline, we effectively addressed our research question.

6.2 Reflection

Overall, this research has successfully addressed all the objectives outlined in the introduction chapter. The significant aim of this work, dataset construction, has been accomplished and verified correctly. However, there are still some limitations to this work. These limitations are elaborated upon in the subsequent sections, encompassing programming language selection, dataset construction, model selection, acoustic properties selection, and influential factors for discriminating properties.

- Programming language selection: In this project, we used Python programming language to implement the whole project. Even though this language could achieve our objectives quickly, the execution time of dataset construction is almost 1 hour, especially for the fixed-duration datasets. As Opensimile also contains C++ programming language, a lower-level architecture programming language to manipulate the operation system, the execution time will be rocketed if we apply this programming language to construct datasets and use this programming language in an industrial setting. Besides, in the evaluation chapter, we employed an indirect approach to assess the impact of session/participant variables on discriminating properties. Unlike Python, R or Weka(Hall et al., 2009) support categorical features, which provides a more robust analysis framework. Furthermore, the evaluation chapter exclusively utilised varied duration data to validate claims from existing literature, potentially limiting the generalisability of results in certain scenarios.
- Dataset construction: This research only uses a 200-ms duration to split the audio due to capacity constraints. The rationality of choosing this value is based on it being slightly greater than the minimum human's auditory perception (170 ms). This number selection is also arbitrary, and a more constant duration application will be more trustworthy, even though it requires more capacity, such as 100 GB.
- Model selection: The decision tree model theoretically can handle multi-label tasks. However, both Python and R lack the functionality to clearly differentiate discriminating properties in feature ranks for multi-classification. Consequently, our approach with the decision tree is adapted to binary labelling tasks. This adaptation potentially mitigates feature race conditions and amplifies the impact of specific acoustic properties, introducing bias towards the generated feature ranks. Additionally, in regression analysis on the fixed duration dataset, we have yet to use the full 6.373 properties to feed into the regression model due to capacity. We implemented this project using Python, which only accepts numeral values for the independent variables and does not accept categorical variables.
- Acoustic properties selection: For each model in the varied duration and fixed duration datasets, we only consider discriminating properties whose feature rank coefficient is greater than 0. In practice, this operation sounds to some extent. However, some meaningful discriminating properties will be discarded if their coefficient is close to 0. The range of candidates for discriminating properties could be larger to explore more meaningful acoustic properties.

- Influential factors discriminating properties: in the evaluation chapter, we focused solely on verifying the selected session, Session 3, within the temporal dynamics experiment. This session specifically examined the disorienting properties of mirthful laughter, discourse laughter, and spoken words. The findings from this experiment might be niched, and it is better to test all 18 sessions, even though it requires more time. We only used one discriminating property in all sessions to inspect the normalised energy in the evaluation chapter. The mean, median, and standard deviation of normalised energy are not distinct, indicating mirthful laughter' normalised energy in the auditory spectrum is moderately larger than related numeral values in discourse laughter and spoken words. It is better to test on other discriminating properties to draw a more objective conclusion.

6.3 Discussion

The nature of acoustic laughter is intricate(Bachorowski et al., 2001), and acoustic laughter is a component of laughter per se, even though we have identified some reasonable discriminating properties. In the discussion section, we will discuss our work's contribution to laughter classification and broader context.

6.3.1 Contribution in Laughter classification

This study reveals several intriguing discoveries, such as discriminating acoustic properties and the factors influencing these properties' discrimination.

Discussion towards discriminating acoustic properties

This work identifies the fundamental frequency, MFCC, auditory spectrum,spectral features and jitter as common discriminating acoustic properties intersected from machine learning experiments and regression analysis. In this discussion, we discuss the discriminating acoustic properties identified in previous literature directly or indirectly, and these acoustic properties depict different aspects of voice information.

Fundamental frequency is the first peak in the formants, and formants are the peak of the audio wave. The first peak preserves the lower frequency of the formant and preserves the most important voice fingerprint as the f1, f2, and f3 are the amplified frequencies of the fundamental frequency. Based on this information, fundamental frequency is the unique identification of voice. This thought that fundamental frequency is associated with vowels is also shared by Akagi et al. (1998), verifying that they found some fluctuation between fundamental frequency and vowels displayed in the electroglottography experiment. In Tanaka & Campbell (2014) work,they concluded that acoustic laughter is mixed with vowels and constants at lower levels. Based on these two researchers' statements(Akagi et al., 1998; Tanaka & Campbell, 2011), this property could uncover some unique information in acoustic laughter, which is also confirmed in Tanaka & Campbell (2014) work, our replicated work.

The effect of MFCC (mel-frequency cepstral coefficient) is to select representative audio information in each frame in select audio clips, and an explanation of these concepts needs some phonetic knowledge. To explain MFCC (mel-frequency cepstral coefficient) clearly, we provide some phenetic information herein. Given any audio length, we could acquire its spectrum envelope, a smooth curve connecting all formats. However, hearing perception focuses only on a specific religion(Deng et al., 2004) rather than the whole spectrum developed in human auditory perception. The Mel frequency is based on a human being's auditory perception, acting as a filter to emphasise certain frequencies and allowing relevant signals to be highlighted.

As the auditory system operates as a unique non-linear system, it responds differently to various frequencies, enabling the extraction of semantic and personal acoustic information from audio signals. MFCC leverages this by converting the linear spectrum into a non-linear Mel frequency spectrum, effectively capturing human auditory features. To reconstruct the voice in the time domain, the next step involves employing a discrete cosine transform to convert the audio wave from the frequency domain back into the time domain. This process results in a feature vector for each frame, encapsulating detailed voice information across specific frequencies within each frame.

Previous research confirmed that MFCC correlates with a higher emotional recognition accuracy (Wang & Shen, 2023). Additionally, laughter could convey emotion(Gilmartin et al., 2013; Koutsombogera & Vogel, 2022). In our project, we aim to discern the distinctive properties of mirthful and discourse laughter. Utilising MFCC properties, we seek to quantify and present the emotional states associated with these laughter types, potentially through means such as electroencephalography(Ismail et al., 2016).

Even though two temporal-associated discriminating acoustic properties have not appeared in most literature, including the auditory spectrum and jitter, these properties still inspire us to explore some exciting findings. The auditory spectrum measures the range of specific wavelengths of laughter; this information is also associated with the duration of laughter(Tanaka & Campbell, 2014) as it reflects the wave fluctuation range of acoustic properties in the time domain and could be translated into the frequency domain by Fourier transformation to identify the energy of specific acoustic properties via time.

Jitter is the cycle of frequency variation, measuring and extracting frequency oscillation in the time domain. To the best of our knowledge, no studies have identified this acoustic property, and the article mentioned in the background chapter(cf. Chapter 2) has not identified it. However, this property could reflect the cycle of laughter from some perspectives and might assist us in tracking the trajectory of specific acoustic projects to identify some rhythmic patterns.

Following this conjecture, we could further explore this temporal property, as the number of cycles for specific acoustic discriminating properties might also be an undiscovered and interesting topic, as this thought is indirectly confirmed by Brockmann et al. (2011), showing that jitter has a significant impact on vowels, especially in males' voice. Additionally, previous research in the literature review chapter(cf. Chapter 2) presented that some acoustic properties correlate with vowels(Tanaka & Campbell, 2011, 2014; Trouvain & Schröder, 2004). By amalgamating these two statements, we could speculate that our hypothesis might be true, even though further exploration needs to be verified by an EEG test(electroencephalogram)(Ringer et al., 2023). As these two acoustic properties have not appeared in previous literature, they are probably related to the configuration in Opensimile, resulting in this phenomenon.

Lastly, spectral features are a general acoustic property in the "Opensimple" framework that describes a feature transforming the temporal signal into a frequency domain by utilising the Fourier transform. Even though this properties could not be aligned perfectly with others' work due to different namings, Tanaka & Campbell (2014) depict some features in the frequency domain that might correspond to these two feature semantically,such as "the maximum value of power (pmax)",as our work and theirs utilised different sets of acoustic property extraction features.

Discussion towards factors impact on discriminating acoustic properties

In the evaluation chapter, we found evidence suggesting that mirthful laughter exhibits greater randomness in both the auditory spectrum and fundamental frequency compared to discourse laughter and spoken words. This observation aligns with the findings of Koutsombogera & Vogel (2022), who assert that

discourse laughter demonstrates a more structured pattern than mirthful laughter. Furthermore, they note that discourse laughter shares similarities in its topical termination function with spoken words. Compared to their work, this research has provided nuance information regarding the disorienting properties variation in temporal flow and normalised feature energy of discriminating property, and this additional experiment has yet to be done by others, even though we only conduct some sessions and some discriminating properties

Another pilot study, which previous work has overlooked, involves verifying the influence of key participants or sessions on specific acoustic properties, whether related to discourse or mirthful laughter. Given that participants may vary in pitch, certain acoustic properties may be more pronounced in particular participants than in others. Additionally, each session may vary in duration, ranging from 5 to 10 minutes, and this variation in length could potentially affect the perceptual properties of laughter. Thus, this pilot study aims to investigate these factors, providing motivation for its implementation.

6.3.2 Wider context discussion

This work thoroughly classifies the acoustic properties of two types of laughter: discourse and mirthful laughter. Most of the work in this research project is algorithm design and machine learning selection, which is partly theoretical. In a wider context, our findings could be embedded into a large system to track laughter related to patients' emotional states in a clinical setting.

The therapeutic potential of laughter in alleviating depression has been suggested(Navarro et al., 2014), and our research delves into identifying discriminating properties inherent in laughter. This exploration extends beyond mere discourse laughter and mirthful laughter. We aim to establish correlations between discriminating acoustic properties in various types of laughter and depression, using Navarro et al. (2014)'s work as inspiration. This potential product could assist psychiatrists and psychologists in accurately quantifying the emotional state of patients with mental disorders via the dynamics of specific acoustic properties in different acoustic laughter over time. The time series plot we envision is similar to what we investigated in the evaluation chapter in terms of the dynamics of discriminating acoustic properties in three utterance types(cf.Figure.5.2).

Not only could our work benefit in a clinical setting, but our work may also bring benefits to more areas, such as improving participant engagement. Based on findings of this research, the product born from our work might be embedded into a laughter response agent, such as the work of Türker et al. (2017) in the human-rot interaction to identify the discriminating properties in the specific laughter, find the correlation, and enhance participant engagement in a fine-grained manner.

6.4 Future work

No research work is perfect, and we proposed two directions for future work, even though other directions might exist. One potential direction is to explore the MUTLSIMO dataset to identify some associated laughter-related elements.

6.4.1 Influence of different genders on acoustic properties in laughter

Males and females exhibit distinct vocal pitches(Latinus & Taylor, 2012), with laughter potentially serving as a cue for gender recognition(Folorunso et al., 2020). Given the differing acoustic properties between genders and the availability of annotations in the MULTISIMO dataset, incorporating a new column is necessary to

retain this information. Consequently, for future investigations, we could pose the following research question: “Are there discernible variations in the acoustic properties of laughter based on the genders of the individuals involved?”. This inquiry delves into the fundamental differences in acoustic laughter across genders, highlighting their significance in interaction with discourse and mirthful laughter.

6.4.2 Consideration of ratified/ratifying laughter to explore more undiscovered phenomena in laughter research

In the background review chapter, we delved into the research conducted by Hegarty (2022), which involved the construction of a dataset comprising both ratified and ratifying laughter. To examine the distinguishing characteristics associated with laughter leadership (defined by Hegarty (2022) as the individual who initiates laughter), such as ratified laughter, we could integrate her ratified/ratifying laughter detection program into our project, given that we have obtained approval to access their code. Then, we could investigate degree of correlation between social laughter (ratified and ratifying) and natural laughter (discourse and mirthful laughter) regarding discriminating acoustic properties. Another further research direction in this category could be identifying discriminating acoustic properties strongly correlated with a laughter leader.

Bibliography

Akagi, M., Iwaki, M., & Minakawa, T. (1998). Fundamental frequency fluctuation in continuous vowel utterance and its perception. In *Fifth International Conference on Spoken Language Processing*.

Alluri, K. R., & Vuppala, A. K. (2020). A study on the emotional state of a speaker in voice bio-metrics. In *Advances in Ubiquitous Computing*, (pp. 223–236). Elsevier.

Bachorowski, J.-A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *The journal of the Acoustical Society of America*, *110*(3), 1581–1597.

Boyd, A., Puri, R., Shoeybi, M., Patwary, M., & Catanzaro, B. (2020). Large scale multi-actor generative dialog modeling. *arXiv preprint arXiv:2005.06114*.

Brockmann, M., Drinnan, M. J., Storck, C., & Carding, P. N. (2011). Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of voice*, *25*(1), 44–53.

Deng, Y., Chakrabartty, S., & Cauwenberghs, G. (2004). Analog auditory perception model for robust speech recognition. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, vol. 3, (pp. 1705–1709). IEEE.

Dunbar, G. (2014). Phonetic and functional aspects of speech laughter: towards an expressive cognitive phonology. *CogniTextes. Revue de l'Association française de linguistique cognitive*, (Volume 11).

Efron, R. (1970). The minimum duration of a perception. *Neuropsychologia*, *8*(1), 57–63.

Eric, M., Goel, R., Paul, S., Kumar, A., Sethi, A., Ku, P., Goyal, A. K., Agarwal, S., Gao, S., & Hakkani-Tur, D. (2019). Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190–202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, (pp. 1459–1462).

Folorunso, C. O., Popoola, O. P., & Asaolu, O. S. (2020). Laughter signature, a new approach to gender recognition. *Engineering Reports*, *2*(11), e12267.

Gilmartin, E., Bonin, F., Campbell, N., & Vogel, C. (2013). Exploring the role of laughter in multiparty conversation. *Proceedings of the SemDial*, (pp. 191–193).

Ginzburg, J., Mazzocconi, C., & Tian, Y. (2020). Laughter as language. *Glossa: a journal of general linguistics (2021-...)*, *5*(1).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Hegarty, K. (2022). Social and individual interactions with ratification of laughter in dialogue. *The final year project report of Trinity college dublin*.

Ismail, W. W., Hanif, M., Mohamed, S., Hamzah, N., & Rizman, Z. I. (2016). Human emotion detection via brain waves study by using electroencephalogram (eeg). *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 1005–1011.

Karishma, R. (2022). Political position estimation of politicians using social media communication. *The master dissertation of Trinity college dublin*.
URL https://publications.scss.tcd.ie/theses/diss/2022/TCD-SCSS-DISSERTATION-2022-095.pdf

Kipper, S., & Todt, D. (2003). The role of rhythm and pitch in the evaluation of human laughter. *Journal of Nonverbal Behavior*, *27*, 255–272.

Knox, M. T., & Mirghafori, N. (2007). Automatic laughter detection using neural networks. In *Interspeech*, (pp. 2973–2976).

Koutsombogera, M., & Vogel, C. (2018a). Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Koutsombogera, M., & Vogel, C. (2018b). Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).

Koutsombogera, M., & Vogel, C. (2022). Understanding laughter in dialog. *Cognitive Computation*, *14*(4), 1405–1420.

Latinus, M., & Taylor, M. J. (2012). Discriminating male and female voices: differentiating pitch and gender. *Brain topography*, *25*, 194–204.

Ludusan, B., & Schuppler, B. (2022). To laugh or not to laugh? the use of laughter to mark discourse structure. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (pp. 76–82).

Ludusan, B., & Wagner, P. (2019). No laughing matter. an investigation into the acoustic cues marking the use of laughter. In *Proceedings of ICPhS*, (pp. 2179–2182).

Ludusan, B., & Wagner, P. (2022a). ha-ha-hha? intensity and voice quality characteristics of laughter. *Proc. Speech Prosody 2022*, (pp. 560–564).

Ludusan, B., & Wagner, P. (2022b). Laughter entrainment in dyadic interactions: temporal distribution and form. *Speech Communication*, *136*, 42–52.

Ma, Z., Dou, Z., Zhu, Y., Zhong, H., & Wen, J.-R. (2021). One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 555–564).

Maggie, B. (2021). Classifying laughter: An exploration of the identification and acoustic features of laughter types. *An Undergraduate Thesis of THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA*.
URL https://repository.upenn.edu/server/api/core/bitstreams/8dc8c28a-c58b-48fb-9dc3-bf2e94b1b60b/content

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Mittal, V. K., & Yegnanarayana, B. (2015). Analysis of production characteristics of laughter. *Computer Speech & Language*, *30*(1), 99–115.

Mohan, V. (2019). Analysis of laughter in task based interactions. *The master dissertation of Trinity college dublin*.
URL https://publications.scss.tcd.ie/theses/diss/2019/TCD-SCSS-DISSERTATION-2019-051.pdf

Morishima, T., Miyashiro, I., Inoue, N., Kitasaka, M., Akazawa, T., Higeno, A., Idota, A., Sato, A., Ohira, T., Sakon, M., et al. (2019). Effects of laughter therapy on quality of life in patients with cancer: An open-label, randomized controlled trial. *PloS one*, *14*(6), e0219065.

Navarro, J., Del Moral, R., Alonso, M., Loste, P., Garcia-Campayo, J., Lahoz-Beltra, R., & Marijuán, P. (2014). Validation of laughter for diagnosis and evaluation of depression. *Journal of affective disorders*, *160*, 43–49.

Park, Y. S., Konge, L., & Artino Jr, A. R. (2020). The positivism paradigm of research. *Academic medicine*, *95*(5), 690–694.

Pietrowicz, M., Agurto, C., Casebeer, J., Hasegawa-Johnson, M., Karahalios, K., & Cecchi, G. (2019). Dimensional analysis of laughter in female conversational speech. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6600–6604).

Ringer, H., Schröger, E., & Grimm, S. (2023). Neural signatures of automatic repetition detection in temporally regular and jittered acoustic sequences. *Plos one*, *18*(11), e0284836.

Sherman, A., Sweeny, T. D., Grabowecky, M., & Suzuki, S. (2012). Laughter exaggerates happy and sad faces depending on visual context. *Psychonomic bulletin & review*, *19*, 163–169.

Szameitat, D. P., Darwin, C. J., Szameitat, A. J., Wildgruber, D., & Alter, K. (2011). Formant characteristics of human laughter. *Journal of voice*, *25*(1), 32–37.

Tanaka, H., & Campbell, N. (2011). Acoustic features of four types of laughter in natural conversational speech. In *International Congress of Phonetic Sciences*, (pp. 1958–1961).

Tanaka, H., & Campbell, N. (2014). Classification of social laughter in natural conversational speech. *Computer Speech Language*, *28*(1), 314–325.

Trouvain, J., & Schröder, M. (2004). How (not) to add laughter to synthetic speech. In *Tutorial and Research Workshop on Affective Dialogue Systems*, (pp. 229–232). Springer.

Türker, B. B., Buçinca, Z., Erzin, E., Yemez, Y., & Sezgin, T. M. (2017). Analysis of engagement and user experience with a laughter responsive social robot. In *Interspeech*, (pp. 844–848).

Vettin, J., & Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, *28*, 93–115.

Wang, F., & Shen, X. (2023). Research on speech emotion recognition based on teager energy operator coefficients and inverted mfcc feature fusion. *Electronics*, *12*(17), 3599.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

A1 Appendix

The appendix chapter encompasses source code and the supplementary material from the methodology,result,and evaluation chapter.

A1.1 Source code

All source codes in this project serve as supplemental material and are stored in Trinity Computer Science's GitLab repository.

https://gitlab.scss.tcd.ie/mshi/acuosticpropertiesidentification-trinitydissertation.

A1.2 Supplementary material in the methodology chapter

The figures and listings below are referred to in the methodology chapter for the motivation of dataset construction.

	Tier	Start Time	End Time	Annotation
436	P006	6441	7057	Ok.
437	P006	39850	40514	Ok
438	P006	50697	51572	Ok.
439	P006	53186	53716	Right.
440	P006	70735	71403	Which pu
...	...	...	...	...
664	P006	613379	613848	great
665	P006	614729	615259	That was fast
666	P006	616330	616844	[r]
667	P006	617727	618864	That would be great
668	P006	618864	619439	We'll have a chat

Figure A1.1: Sample of P006 tier in S02_Final.eaf

Code Listing A1.1: XML structure in each EAF file containing time information and annotation information

```
<TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="UtteranceType" TIER_ID="shared_6_7">
  <ANNOTATION>
      ....

   <ANNOTATION>
     ....
</TIER>
```

Table A1.1: The Tier ID count for the S2 session extracted from EAF

Tier ID
M001_S02
Sections

Continued on next page

Table A1.1: The Tier ID count for the S2 session extracted from EAF (Continued)

P007
P006
comment
laughter_section
nonlaughter_section
shared_6_7
shared_6_M001
merge_6
merge_7
merge_S02
solo_6
solo_7
solo_M001
Laughter_M001_S02
Turns
Laughter_7
Laughter_6
subsections
secsubsec
subsub

Code Listing A1.2: The annotation of continous moments in the TIME_ORDER tag

```
<TIME_ORDER>
     <TIME_SLOT TIME_SLOT_ID="ts1" TIME_VALUE="0"/>
     <TIME_SLOT TIME_SLOT_ID="ts2" TIME_VALUE="0"/>
     <TIME_SLOT TIME_SLOT_ID="ts3" TIME_VALUE="0"/>
     <TIME_SLOT TIME_SLOT_ID="ts4" TIME_VALUE="0"/>
     <TIME_SLOT TIME_SLOT_ID="ts5" TIME_VALUE="0"/>
     <TIME_SLOT TIME_SLOT_ID="ts6" TIME_VALUE="1375"/>
     ....
     <TIME_SLOT TIME_SLOT_ID="ts2815" TIME_VALUE="622200"/>
     <TIME_SLOT TIME_SLOT_ID="ts2816" TIME_VALUE="622817"/>
     <TIME_SLOT TIME_SLOT_ID="ts2817" TIME_VALUE="622817"/>
   </TIME_ORDER>
```

Code Listing A1.3: The annotation of continous moments with utterance content in the ANNOTATION tag

```
   <ANNOTATION>
```

```
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a106" TIME_SLOT_REF1="ts1385"
        TIME_SLOT_REF2="ts1394">
       <ANNOTATION_VALUE>Are you ready for the second question
          now?</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
</ANNOTATION>

  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a466" TIME_SLOT_REF1="ts373"
        TIME_SLOT_REF2="ts375">
       <ANNOTATION_VALUE> [i] [eh] </ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>

 <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a672" TIME_SLOT_REF1="ts155"
        TIME_SLOT_REF2="ts163">
       <ANNOTATION_VALUE>Mirthful</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
</ANNOTATION>


  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a680" TIME_SLOT_REF1="ts746"
        TIME_SLOT_REF2="ts759">
       <ANNOTATION_VALUE>Discourse</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
</ANNOTATION>



  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a726" TIME_SLOT_REF1="ts430"
        TIME_SLOT_REF2="ts679">
       <ANNOTATION_VALUE>[non_laugh]</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
</ANNOTATION>
```

	Tier	Start Time	End Time	Annotation
670	laughter_section	0	1375	Discourse
671	laughter_section	69343	69869	Mirthful
672	laughter_section	76272	77049	Discourse
673	laughter_section	104875	105545	Discourse
674	laughter_section	105545	106409	Mirthful
675	laughter_section	106409	107005	Mirthful
676	laughter_section	145715	146689	Mirthful
677	laughter_section	199365	199995	Discourse
678	laughter_section	201134	202290	Discourse
679	laughter_section	216708	217296	Discourse
680	laughter_section	217296	218064	Discourse
681	laughter_section	219626	221176	Discourse
682	laughter_section	227668	228749	Mirthful
683	laughter_section	228749	229845	Mirthful
684	laughter_section	245913	247326	Discourse
685	laughter_section	263529	265139	Mirthful
686	laughter_section	296663	297305	Discourse
687	laughter_section	298937	299411	Discourse
688	laughter_section	309432	310028	Discourse
689	laughter_section	316783	317192	Discourse

Figure A1.2: Sample screenshot of laughter tier row instances

A1.3 Supplementary material in the results chapter

The figures and listings below are referred to in the results chapter for the regression analysis task.

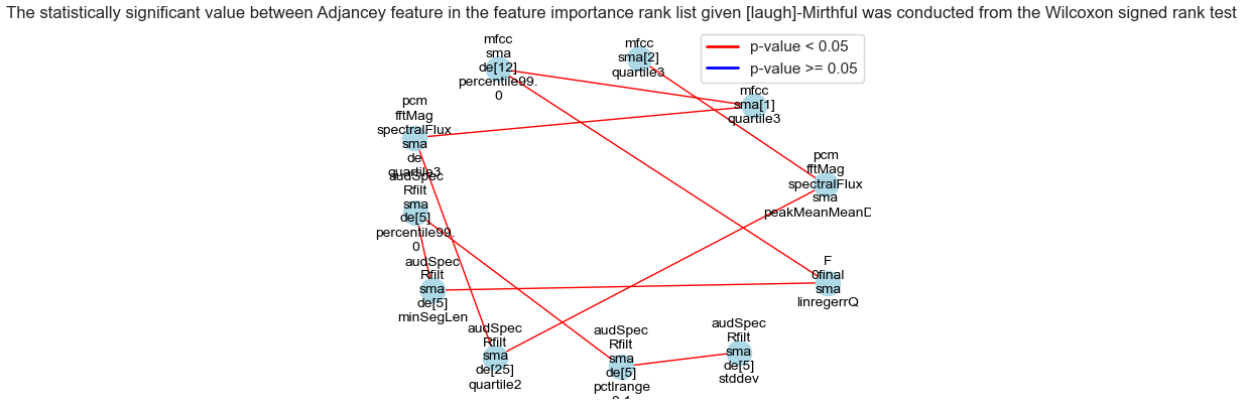


Figure A1.3: Top-N adjacency features correlation given mirthful laughter in the varied duration datasets.

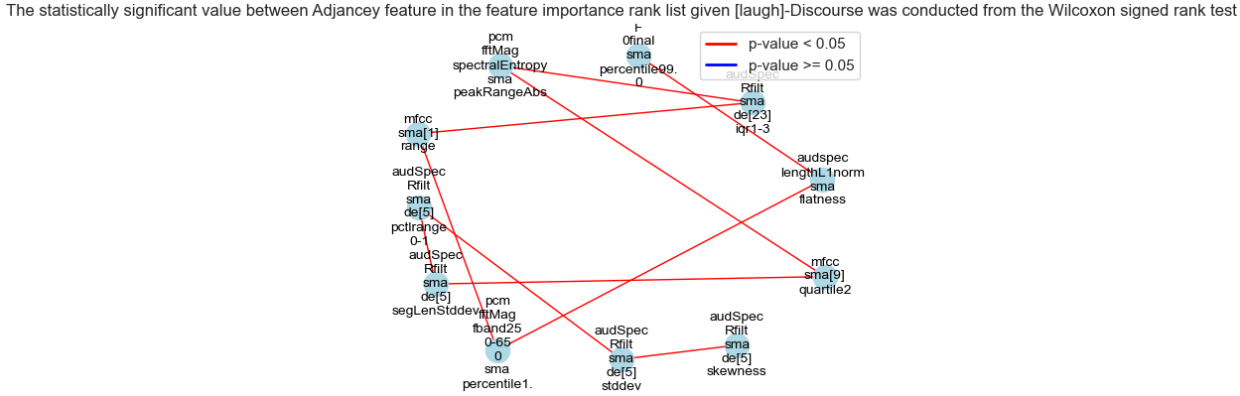


Figure A1.4: Top-N adjacency features correlation given discourse laughter in the fixed duration datasets.

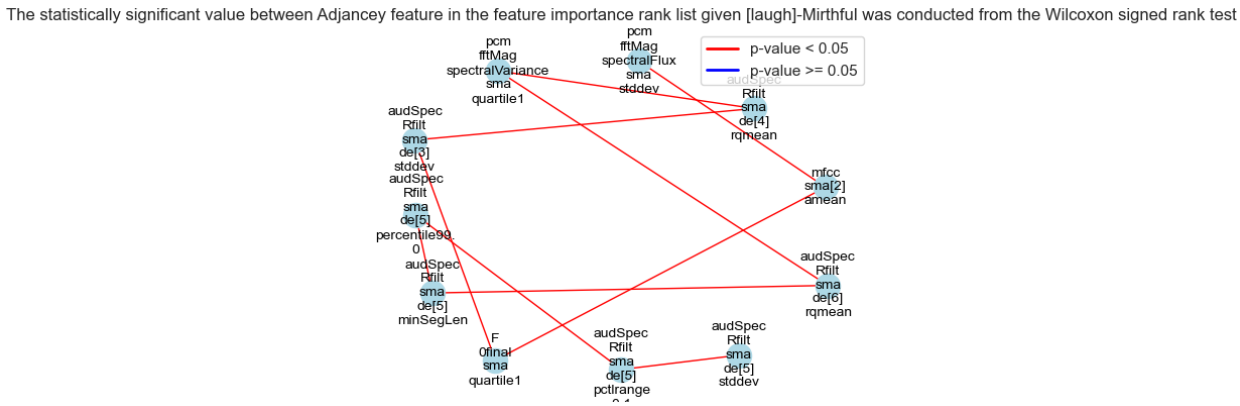


Figure A1.5: Top-N adjacency features correlation given mirthful laughter in the fixed duration datasets.

A1.4 Supplementary material in the evaluation chapter

Below figures are referred to in the machine learning model performance in the evaluation chapter.

A1.4.1 Confusion matrix

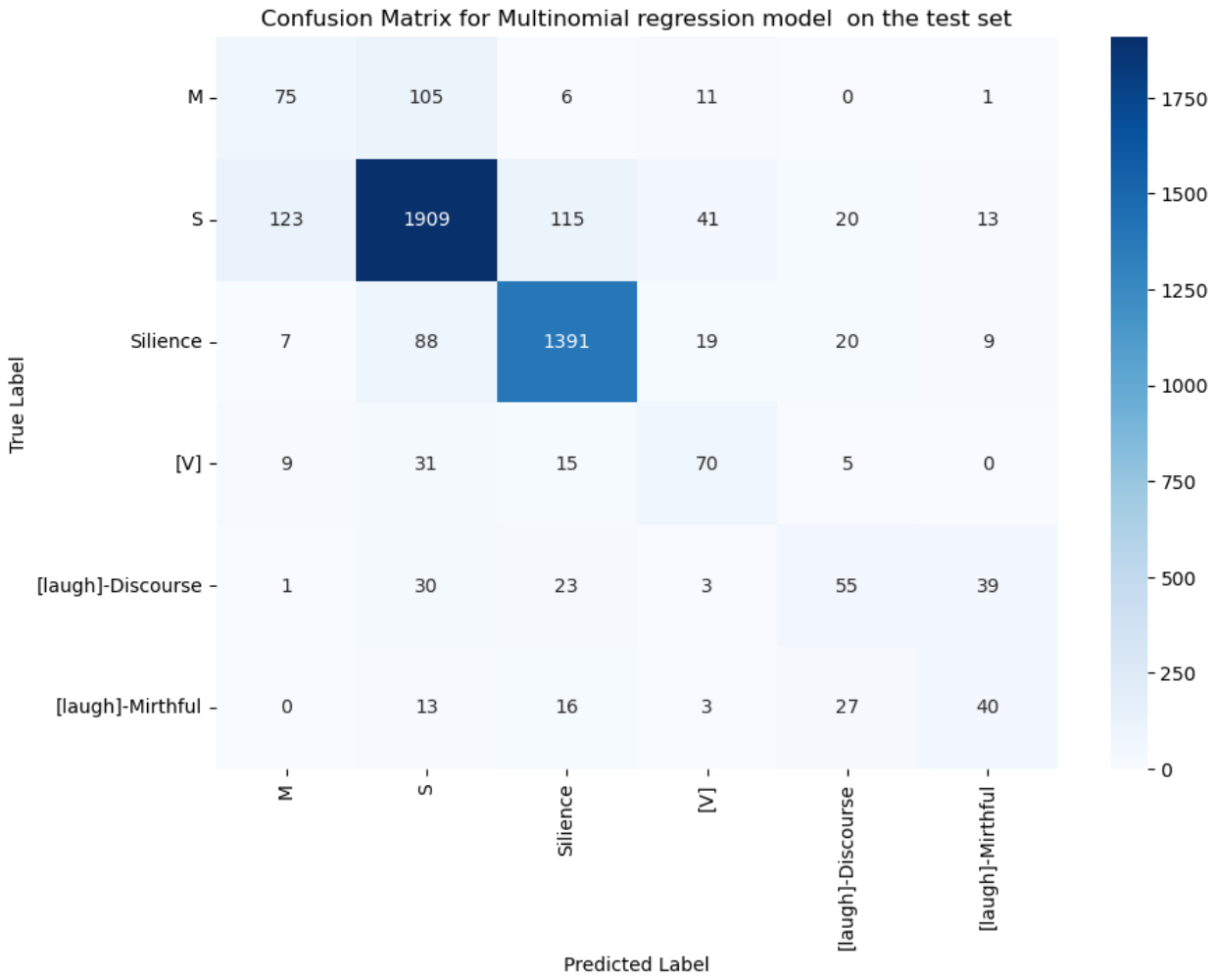


Figure A1.6: The confusion matrix for multinomial logistic regression on the varied duration dataset

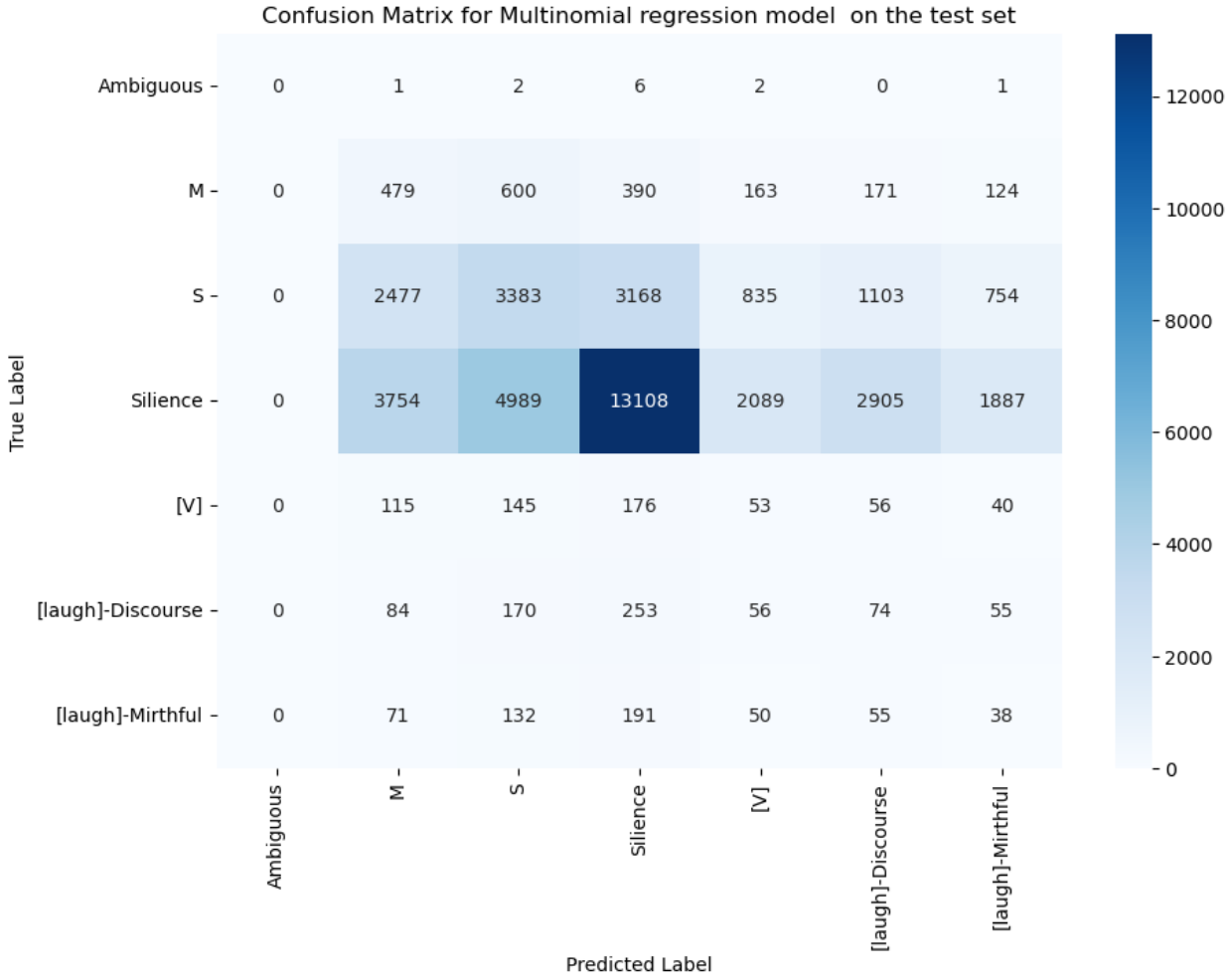


Figure A1.7: The confusion matrix for multinomial regression on the fixed duration dataset

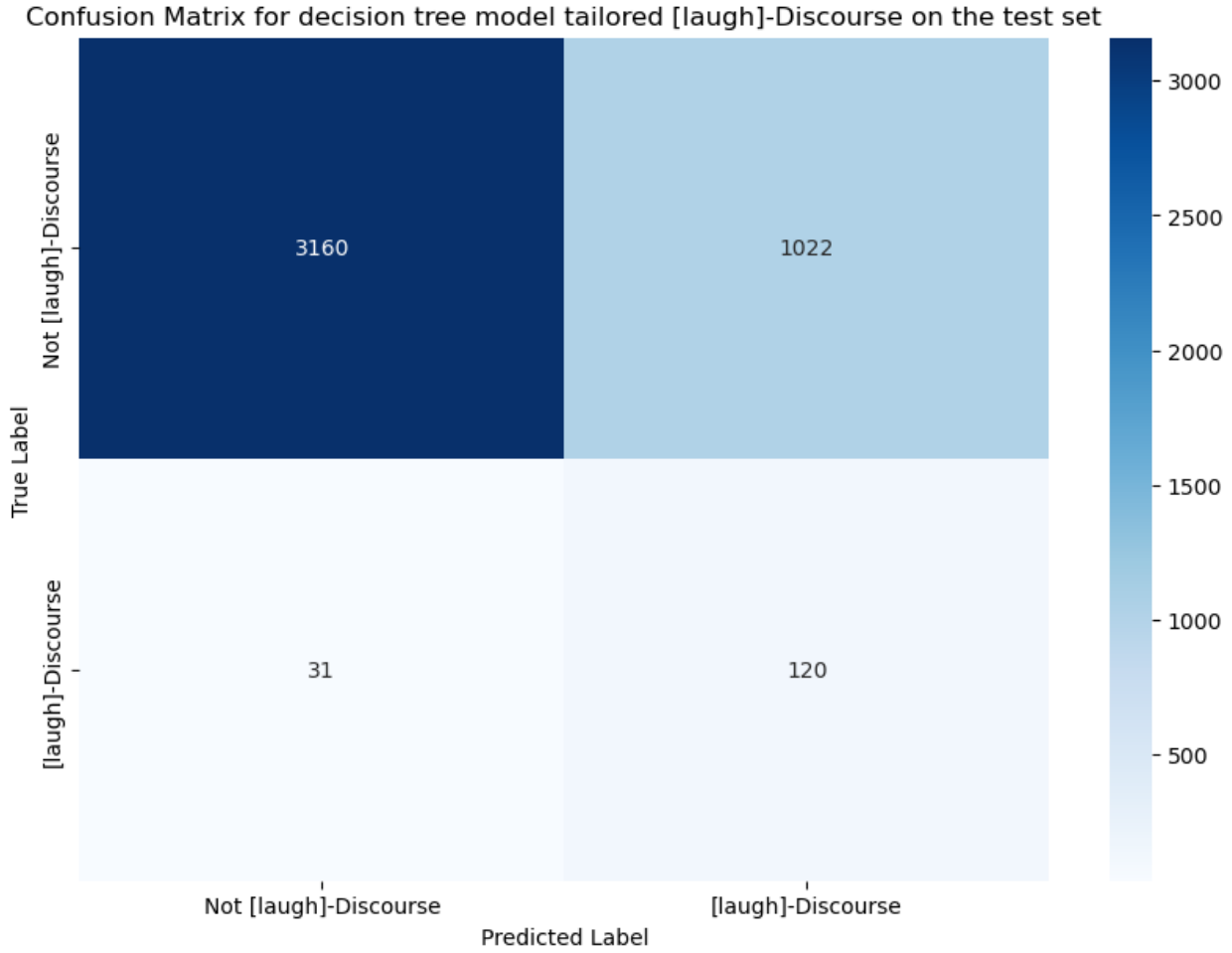


Figure A1.8: The confusion matrix for Decision tree tailed for "[laugh]-Discourse " for the varied duration dataset

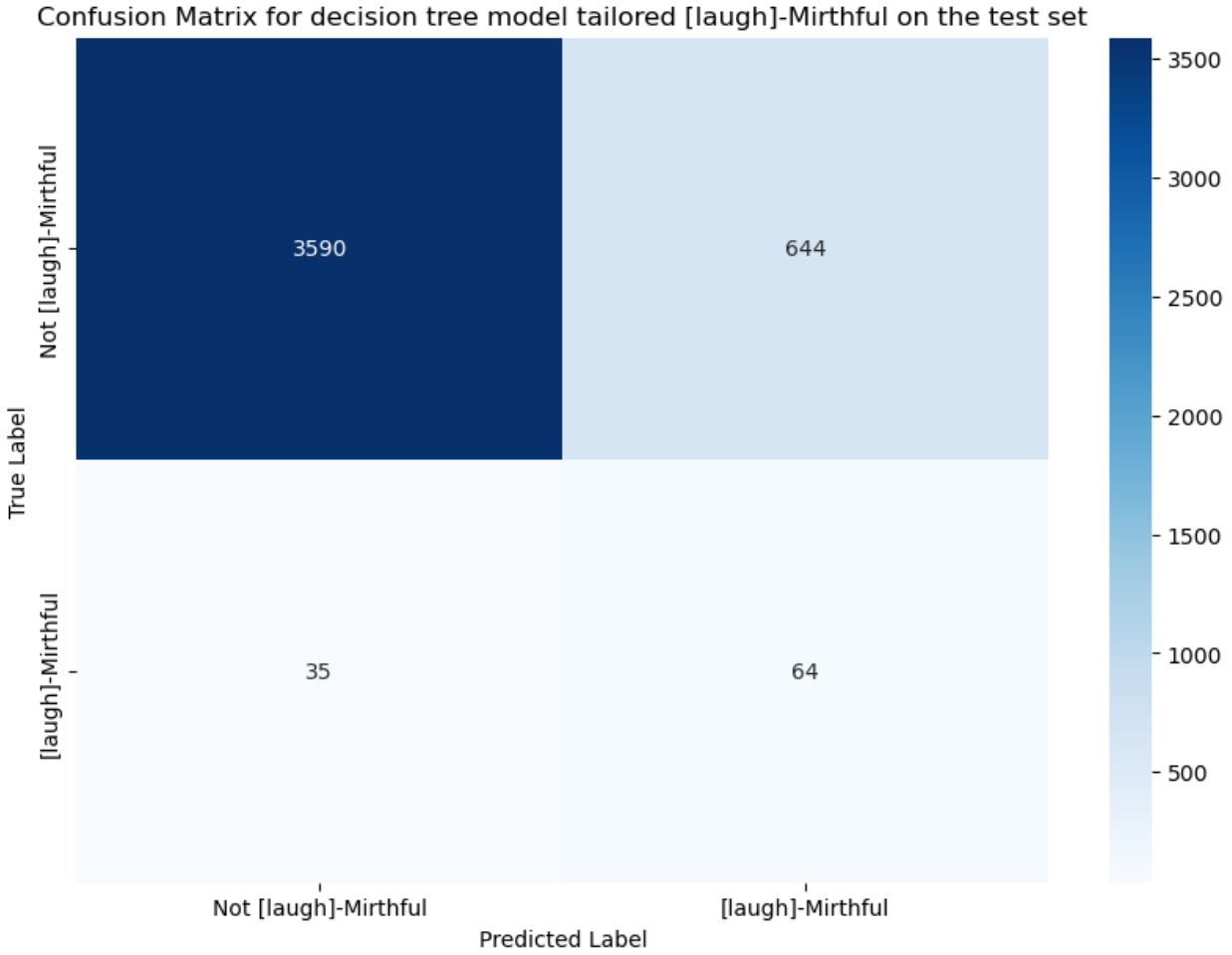


Figure A1.9: The confusion matrix for Decision tree tailed for "[laugh]-Mirthful" for the varied duration dataset

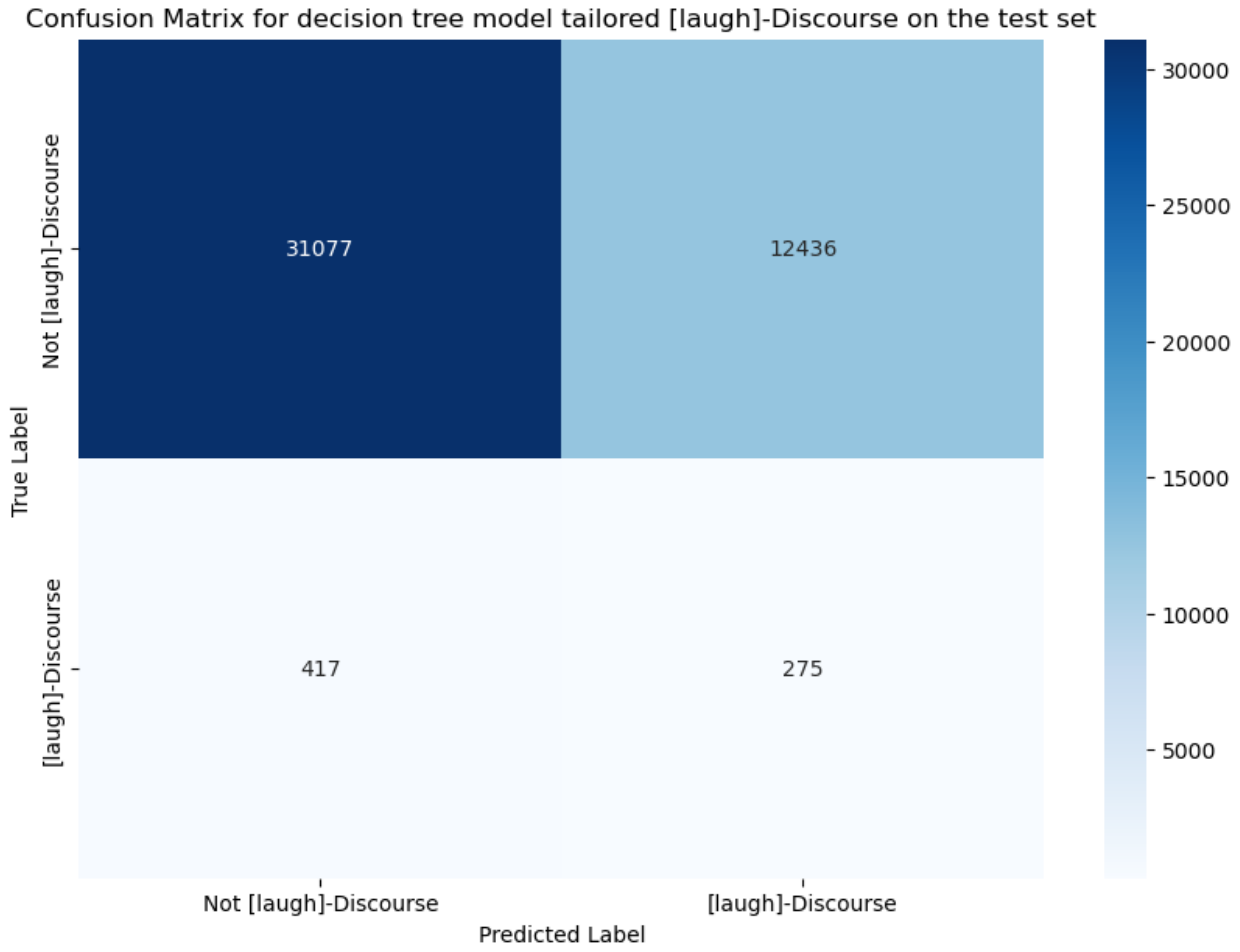


Figure A1.10: The confusion matrix for Decision tree tailed for "[laugh]-Discourse" for the fixed duration dataset

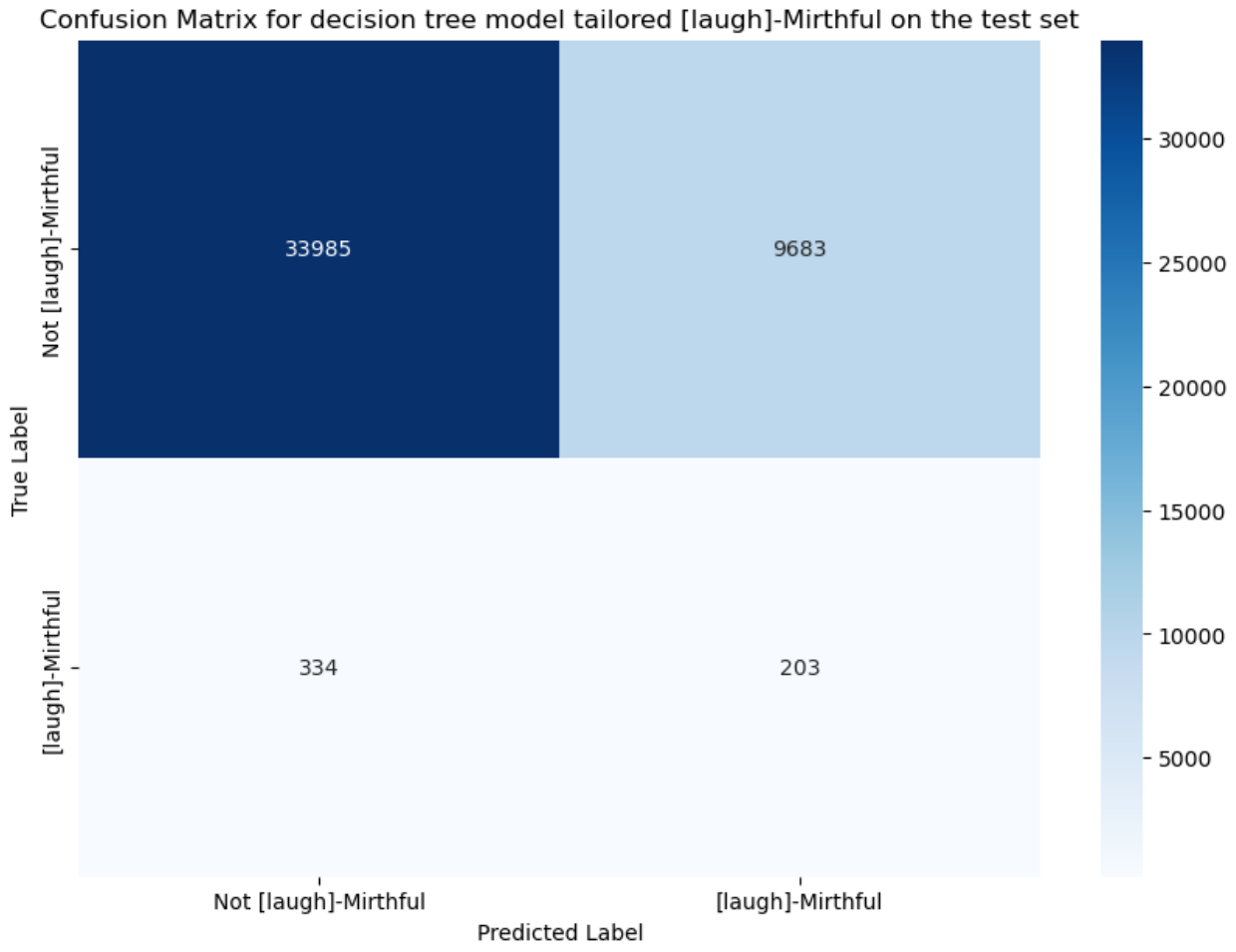


Figure A1.11: The confusion matrix for Decision tree tailed for "[laugh]-Mirthful" for the fixed duration dataset

A1.4.2 Other acoustic properties dynamics for three utterance events

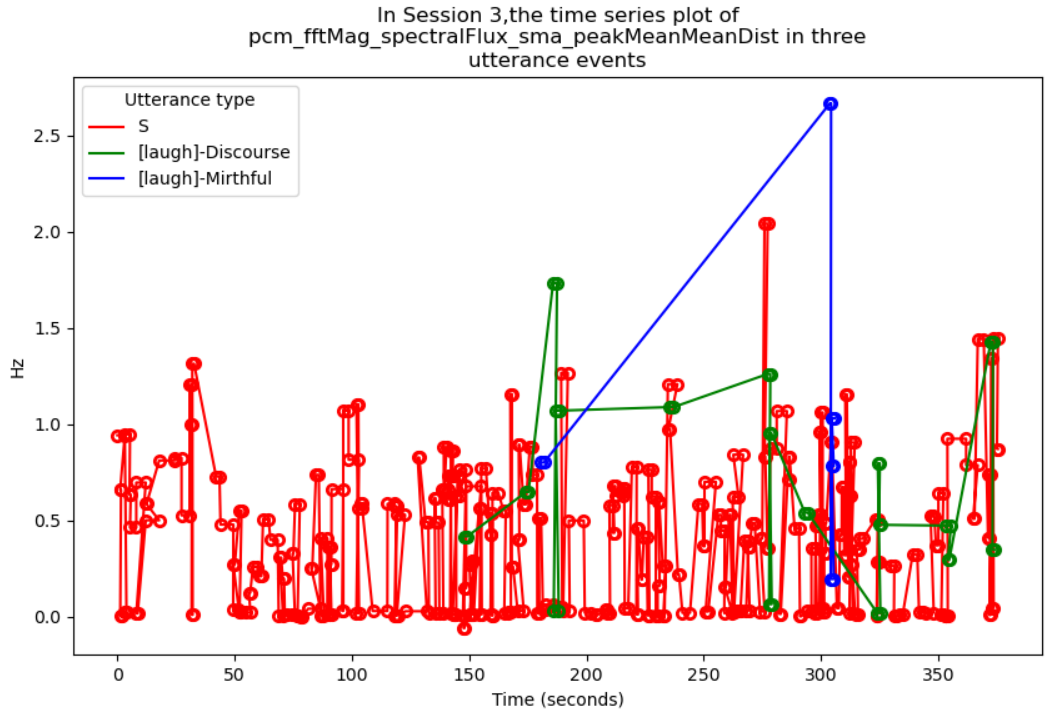


Figure A1.12: The dynamics of “Mean distance of spectral Features ” in the session 3 for three utterance types

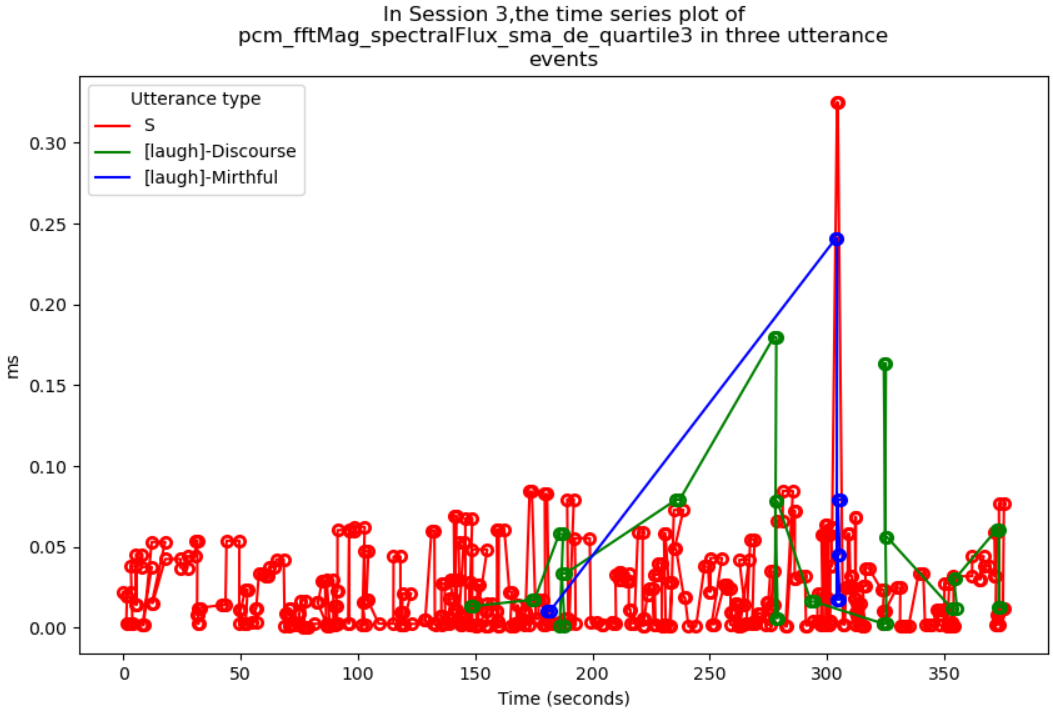


Figure A1.13: The dynamics of “Third quartile of spectral Features” in the session 3 for three utterance types

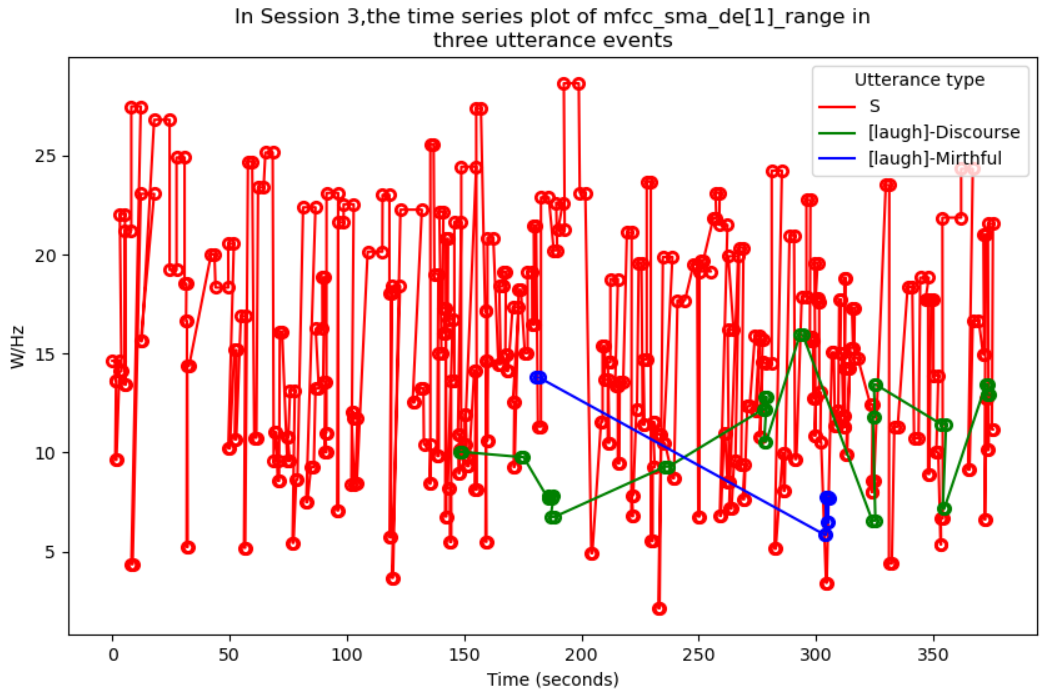


Figure A1.14: The dynamics of “Range of MFCC” in the session 3 for three utterance types

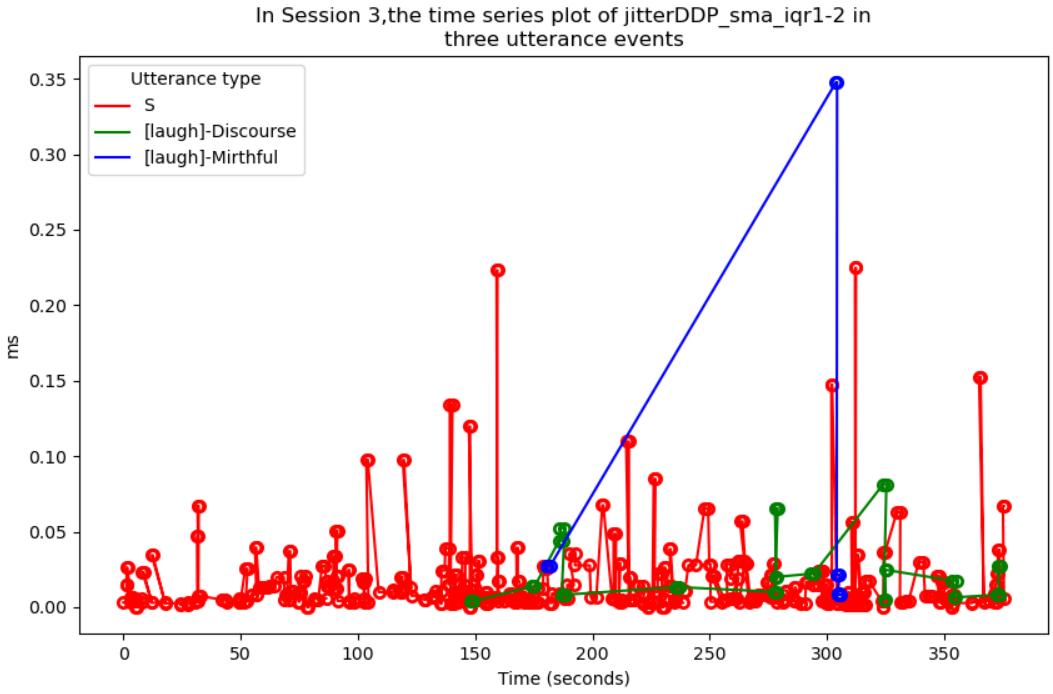


Figure A1.15: The dynamics of “The differential frame-to-frame Jitter” in the session 3 for three utterance types

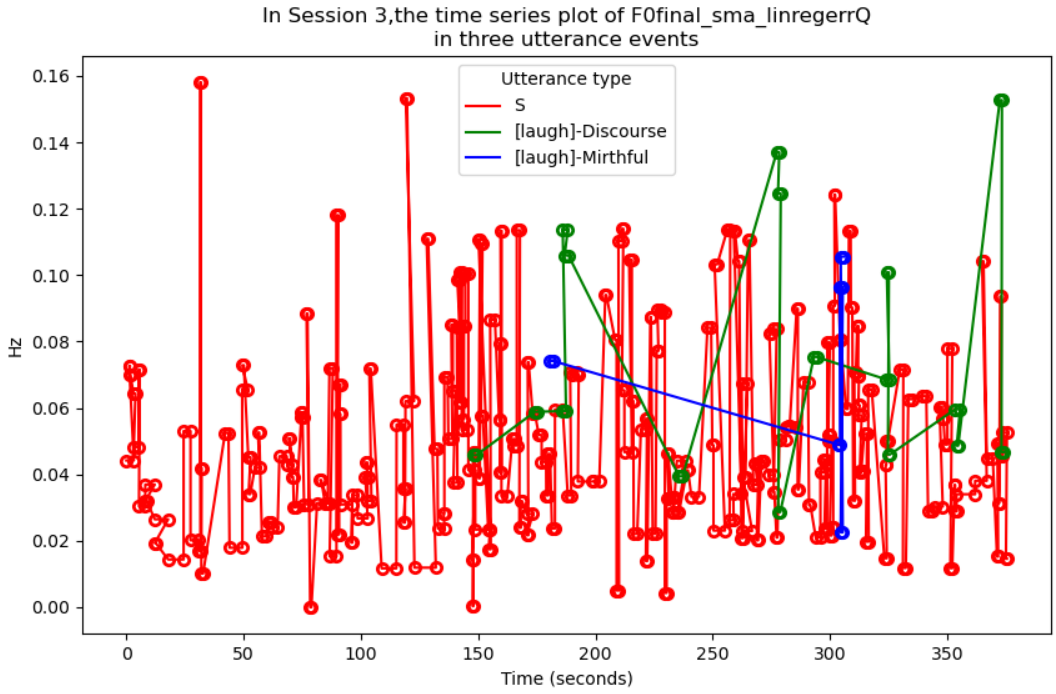


Figure A1.16: The dynamics of “The linear regression error of fundamental frequency” in the session 3 for three utterance types