Audio-based Stylistic Characteristics of Podcasts for Search and Recommendation: A User and Computational Analysis

KATARIINA MARTIKAINEN

Master in Interaction Technology
University of Twente, EIT Digital Master School
University Supervisor(s): Khiet Truong & Roeland Ordelman
Industry Supervisor: Jussi Karlgren
Host company: Spotify, Stockholm
Date: August 18, 2020

Abstract

Even though many search engines already excel when it comes to text document retrieval, retrieving spoken content is an entirely different matter. Furthermore, the paralinguistic dimensions of language have been widely neglected in spoken content retrieval. The focus has been only on the content of *what* is said, ignoring the ways of *how* things are said. This master thesis argues that also the style of spoken content is important to listeners, and that implementing stylistic search and recommendation capabilities could substantially benefit search and recommendation systems of spoken content. This research focuses on searching for and recommending podcasts. We present our results on what kind of stylistic features of podcast content listeners care about, and how these features can form higher level stylistic categories. Furthermore, we report the results of our experimentation on the technical suitability for using the stylistic features as basis for automatic podcast recommendation. Our core findings are that podcast listeners have clear ideas on what kind of stylistic content matters to them, and that many of these stylistic features can be used for automatic podcast recommendation based on the information captured from the podcast audio signal.

Contents

Chapter 1

Introduction

In this section, increase in podcast popularity is discussed and some problems in finding relevant audio content are presented. Additionally, benefits of improving spoken content retrieval systems, and more specifically the benefits of modelling stylistic characteristics of spoken content, are discussed. Finally, the research goals and approach are introduced.

1.1 Increasing Popularity of Podcasts

The amount of new types of digitally stored casual speech in the form of interviews, lectures, debates, radio talk show archives, and educational podcasts is increasingly available [1]. For example, YouTube has around 300 hours of video uploaded to it every minute [2]. One of the forms of audio media which has really gained popularity during the past years are podcasts [3]. As of 2019, there are 90 million monthly podcast listeners in the United Sates, twice as many as in 2015. Podcast usage is driven by a younger generation looking for information, entertainment, and distraction. [4] Podcasts are a particularly appealing form of entertainment when the listener's visual attention is required elsewhere. They are also an attractive option for pastime when for example commuting, cleaning, or exercising [5].

The increasing popularity of podcasts can also be seen in the commercial interest of companies. Many companies are moving towards products that enable making and finding podcasts easier. For example Spotify, a music streaming platform, has recently acquired Gimlet Media and Anchor, two popular podcast creators. Spotify has also re-designed its app to make podcasts more prominent to the user [6]. Google has added functionality to its search engine to facilitate finding podcasts based on the topic of discussion, as well as added support to their home assistant for playing podcasts [7]. On the content supply side, tens of thousands of podcasts are produced on a daily basis [5]. While the popularity of podcasts is increasingly growing, podcasts suffer from the same challenges as other spoken content. When it comes to enabling users to find relevant content to consume and to building scalable spoken content retrieval systems, the industry still has many challenges to tackle.

1.2 Challenges of Spoken Content Analysis

From the end user's perspective, searching and browsing for relevant spoken content can be a challenge. This is true for any kind of spoken content from for example oral history archives, to lectures, or news. However, in our study we concentrate on podcasts. Podcasts are an interesting focus for our study not only because of their increasing popularity, but because from the perspective of audio processing, podcasts are a diverse and a challenging subset of audio media. When it comes to format, podcasts are a more free form of spoken audio than for example traditionally studied broadcast news. Podcasts can come in the form of interviews, informal chats, debates, stories, and much more. In addition to this, podcasts can contain many different kinds of sound effects and music. This rich format makes podcasts particularly interesting for us, and hence in our research we will be focusing on podcasts. However, our research can have implications to a wider range of spoken content. Therefore, we discuss the challenges of spoken content retrieval in a wider context before concentrating on podcasts for the rest of our report.

Spoken content does not afford the same sort of skimming through or jumping back and forth as text does. Additionally, the browsing is limited in terms of speed, as a recording is more limited by the recorded speed of talking. The average reading speed of most adults is around 200 to 250 words per minute. In comparison, an experienced public speaker would deliver his or her speech at a rate of about 160 words per minute. While speaking faster is possible, 160 words per minute is the rate which is generally considered to be comfortable for most listeners [8]. Being able to help users to find relevant content in a way where the users would not have to listen through many irrelevant documents before finding something worth-wile would improve the user experience of the retrieval system. For this, modelling content of the spoken documents as accurately and as richly as possible is required.

Traditionally, in order to describe spoken content, audiovisual data in digital libraries has been labeled manually. However, considering the recent increase in amount, availability, and type of audiovisual data, manual tagging is not a sustainable approach. [9] Information on the spoken documents' content is often mostly limited to metadata, entered manually either by the creators or by the curators of the database. This introduces different problems for the two groups. The content creators are not a unified group, and thus their skill and interest level at entering relevant metadata may vary. Additionally, since they are so disparate, it is unlikely that they would adhere to the same standards, and could use different terms for the same thing, which in turn complicates the information representation schemes in the databases and retrieval systems. Database curators on the other hand, while likely to be more unified in their categorisation, are faced with the problem of having an enormous amount of content, and with the previously mentioned limits on speed which is inherent to audio content, the rate at which this information can be entered is significantly lower than the rate at which it is produced. Therefore, automatic approaches for modelling and describing spoken content media are needed.

One common automatic approach to modelling and describing spoken content is to run the audio through an automatic speech recognition system which transcribes the spoken words to text. In this case the spoken documents, after transcription, could be treated similarly to text documents [1].

However, traditionally this approach has focused solely on the content of what is said ignoring the style of how things are said. Spoken language carries much more information than just the spoken words. For example, it carries information on the speaker's emotion, attitudes, and personality. Additionally, the audio media might contain music and other sounds, which are also ignored by the current text transcript approach. Whereas text transcripts are needed for topical search and recommendation, they do not capture the richness of stylistic information carried by audio-media. This stylistic information might be important to users. Users might want to, for example, search for stylistically similar content, or would benefit from recommendations, which in addition to topic consider the stylistic characteristics of the recommended media. Therefore, there is a need to address "how" spoken content documents are in addition to "what" is said in them. A first step to modelling the "how" is to understand what kind of stylistic characteristics of given spoken content media users care about. Secondly, since manually tagging spoken content is slow, an automatic approach is needed. Some stylistic characteristics of "how" a spoken content document is can be addressed in text transcripts, but also in the audio. We focus on audio based analysis only, which does not require text transcripts. An advantage to this approach is that it does not require use of automatic speech recognition systems which are expensive and slow to scale across languages and different user contexts.

Solving the issues of accurate and rich content descriptors would bring benefit to many different applications of spoken document retrieval. In the case of podcasts, being able to model the podcast content not only in terms of what is said but also in terms of the stylistic characteristics could enable better discovery of relevant podcasts, both in terms of search and recommendation. Consequently, this would bring more relevant traffic to the podcast content creators. On a broader level, finding methods to better describe spoken content would benefit for example media professionals, who often search media from broadcast archives to create new content, such as in news segments or documentaries, as it would bring them more accurate and desirable search results. Furthermore, it would bring similar benefits to researchers going through speech archives, who might be looking for stylistic content of certain kind.

1.3 Research Goals and Approach

In this report we present our research on what kind of stylistic characteristics of podcasts the podcast listeners care about, and how well these characteristics are suited for enriching automatic podcast search and recommendation systems. Therefore, this research focuses on two research questions:

> RQ1: What stylistic characteristics of podcasts do listeners find interesting or important for their podcast listening experience?
>
> RQ2: How suitable are the listener perceived stylistic characteristics of podcasts for automatic podcast recommendation and search based on the attainable information from the podcast audio signal?

The research focuses on podcasts, because we believe that podcasts are an application area where search and recommendation can be made more interesting for the user by adding stylistic character-

istics to the podcast content modelling. As podcasts are spoken content media, which is not only consumed for its informative content, but also for pastime and entertainment [10] we believe that podcasts would benefit from accessing content based on the content's stylistic characteristics more than for example media such as broadcast news which are mainly meant for transferring factual information. Podcasts are also an interesting focus area for the study of spoken content retrieval and speech processing at large because of their unpredictable and unstructured nature. Podcasts mostly consist of natural speech which does not have any pre-determined format and can include many different styles and formats. For example, the format of a podcast can be an interview, a chat, a monologue, or a dialogue. Podcasts might contain a lot of speaker overlap or they might have music which is overlapping with speech. Spoken content retrieval research has only recently started to move towards diverse and unstructured media and as such modelling podcast content is a nearly unexplored research area. Understanding how the stylistic characteristics of podcasts can be modelled could lead to improved user experience when browsing for relevant podcasts. Additionally, it could form a basis for information retrieval approach that can then be applied to other spoken content, for example, to audio books or any other spoken content which could benefit from being accessible by its stylistic characteristics.

This study uses a mixed method approach to answer the research questions RQ1 and RQ2. Because the research consists of two distinct research questions, the research is divided into two phases, one for each question. First, a qualitative approach is used to answer the research question RQ1. A quantitative approach is selected for answering the second research question RQ2. The qualitative approach for answering RQ1 comprised user research with participatory design workshops. The quantitative approach for answering RQ2 consisted of developing a program to detect stylistic characteristics identified in the first phase and observing the value distributions of the characteristics. A framework of stylistic podcast characteristics was devised in order to map out a possible structure of these characteristics. This framework was used to structure and aid the research approach. The framework can be seen in Figure 1.1. The framework together with examples are discussed in more detail below.

The stylistic characteristics of podcasts are broken down to three levels of abstraction namely, stylistic categories, stylistic features and low level acoustic features. From now on these three levels are referred to as discussed below. The stylistic categories are the top level categories podcasts could be categorised to. The stylistic categories can be defined by mid level stylistic features. Finally, the stylistic features could be modelled by low level acoustic features of the podcast audio signal. We envisioned that a high level stylistic category could be, for example, "funny podcasts". The stylistic category consists of different mid level stylistic features. In the case of the example "funny" we envisioned that the mid level stylistic features could be for example "presence of laughter" or "jokes". Then in turn all of these stylistic features would hopefully have some low level acoustic features which could be used for automatic detection of the stylistic features directly from the podcast audio. To refer back to our example, laughter for example could be detected by using Mel-frequency cepstral coefficients of the audio. By building the whole pipeline we hope to ultimately be able to automatically model and detect the stylistic categories ("funniness" in the example) from the podcast audio.

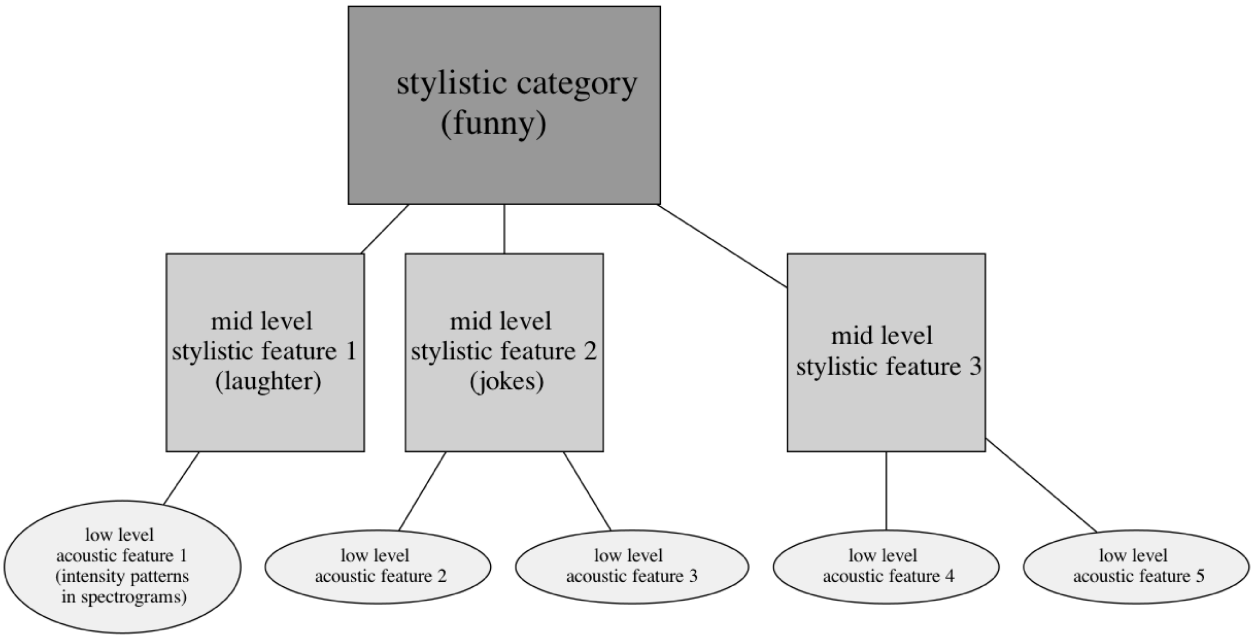

Figure 1.1: Framework for stylistic podcast characteristics

The framework guided our method selection and research design. Additionally, the framework influenced our decision on selecting what stylistic characteristics to focus on during the second research phase. We approached podcast listeners in order to understand their perceptions of the top two levels of the framework. The first step was to find out if podcast listeners perceive any stylistic variations in the podcasts they listen to and if they care about the stylistic variations. If this were to be the case, it was of interest what kind of stylistic characteristics listeners cared about. The results of the fist research phase guided our focus during the second research phase and our work on the two lower levels of the framework: the stylistic features and the underlying low level acoustic features. We planned to work on the two bottom levels of the framework by, if possible, using open source speech processing tools to automatically compute information on the stylistic features. Connecting all the three levels of the framework to one coherent implementation was left for future work.

Chapter 2

Background

In this section the theoretical context and background of the thesis project is presented. More specifically, the research area of spoken content retrieval is described. Then, the field of automatic speech recognition and its challenges in the context of speech content retrieval are discussed. Finally, paralinguistics together with affective content analysis are explained.

2.1 Spoken Content Retrieval

Spoken Content Retrieval (SCR) is defined as the task of returning speech media results that are relevant to an information need expressed as a user query [11]. In contrast, another commonly used term in similar context: "spoken document retrieval" has historically referred to retrieval techniques for content with pre-defined document structure, such as stories in broadcast news. However, as the field has matured it has become clear that there is a need for much wider variety of documents which do not always have a pre-structured format. Therefore, "speech retrieval" [12] was re-adopted to all documents with or without clear document boundaries. Spoken content retrieval concentrates on speech as the returned content, not as the modality of query input. In addition, it can also include retrieval of multimedia documents which contains speech, e.g. video. [11] In this study, the term spoken content retrieval is used as defined above.

2.1.1 Spoken Content Retrieval Systems

Although the architecture of SCR systems might vary depending on the use scenario, the general underlying components remain more or less the same. The SCR architecture can according to Larson [11] be divided in seven abstracted components: *query, retrieval system, ranked results, visualisation and playback, speech collection, speech recognition system, and index*. The user inputs a search *query* which attempts to express the information need to the retrieval system. Often the query is a highly under-specified representation of the information need. Therefore, part of the retrieval system's purpose is to automatically enhance this specification so that it can return useful results to the user. [11]

The *retrieval system* matches the query with the items in the collection. To be able to do this, the retrieval system consults the index. The index contains features which represent the items in the collection and often also time-codes indicating the time points within each item associated with occurrences of these features. Indexing features can be for example terms consisting of words and phrases derived from the spoken content or speaker identities. Additionally, the index often includes weights for each indexing term, indicating how important they are. The exact form of the index depends on the application of the SCR system. For example, some systems return only whole documents. In such cases, providing time-codes of related index terms is not useful. [11]

The indexing features are often generated by the *speech recognition system* that processes the material in the *speech collection* at indexing time. However, metadata and rich transcripts that include for example labels indicating who spoke when are also important sources of information for creating the index. The output of the SCR system is a list of often *ranked results*, a set of results ordered in terms of their likelihood of potential relevance to the query. The results can be of different formats depending on the use scenario. For example, whole documents can be returned, or a dynamically adjusted audio interval, or an entry point to the audio can be provided to the user, depending on the system's purpose. Finally, the SCR system offers the user *visualisation and playback* of the individual results, where the user can listen to the retrieved content. [11]

The method of creating an index and what features the index contains depends on the information retrieval system and its purposes. A basic example of a indexing system is the boolean indexing system. The main idea of boolean indexing is to keep a dictionary of terms (sometimes referred to as a vocabulary). For each term (a normalised word), a list of documents where the term occurs is kept. Each item in the list is called a posting. Thus, this list is often called a posting list. A dictionary is built by looping through each normalised token in the document, organising the tokens in alphabetical order, and merging multiples of the same token. Each document in the document collection is assigned a unique serial number known as document identifier. The input to indexing is then a list of pairs of term and document identifiers. Results to a boolean search query are obtained by finding an intersection of documents where all the specified query terms occur (or not occur). [13]

2.1.2 Using Metadata for Indexing

Podcast retrieval is an example of when adding information from the podcast's metadata to the index and using it together with the transcript can improve retrieval performance. A podcast's metadata is often comprised of information such as the title of the podcast show, the title of the episode, the descriptions of the show, and the description of the episode. Looking into metadata can be useful for example when the user's information need involves speaker identity. This information is more likely to be provided in the metadata than in the spoken content of the podcast. Nevertheless, it is important to note that the quality of the added metadata can vary significantly between podcast creators. It is also worth noting that many internet search engines currently use metadata alone to index podcasts. [11]

2.1.3 Using Automatic Transcripts for Indexing

In spoken content retrieval systems, text transcripts which are used for content modelling of the retrieved documents are often produced by Automatic Speech Recognition (ASR). Automatic Speech Recognition systems computationally map an acoustic signal to a string of words [14], thus converting the spoken language to text. If such a system is used to produce a text transcription of the spoken content, different natural language processing techniques can be used to tag the document with certain indexing labels. For example, word frequencies can be used to understand the topics discussed in the document, and could be used for tagging the document with search keywords in addition to the metadata already existing in the document.

In the context of SCR, ASR has been traditionally used in so called "listening typewriter" paradigm, where the goal of the ASR system is to generate a written transcript of spoken content as accurately and as domain independently as possible. More recently, ASR approaches have shifted towards outputs which can provide indexing terms (words and phrases) for SCR systems. The indexing systems have evolved from earlier approaches of "wordspotting"[15] to "Spoken Term Detection (STD)" and "Spoken Utterance Retrieval (SUR)". Unfortunately, STD and SUR are both blind to the overall context of the document. They return only word and utterance based matches. In other words, these systems do not even attempt to match results with the underlying need for a specific sort of content. [11]

2.1.4 Challenges in Using Automatic Transcripts for Indexing

Podcasts are often characterised by heterogeneous spoken audio which varies widely in quality. The audio can for example contain multiparty conversations, monologues, unpredictable subjects, in addition to background noises and music. Additional variations can be caused by for example the quality of recording, variable speaking styles, and showing emotion or state of mind by laughing, swearing, or crying. These factors are all known sources of ASR errors and hence introduce major challenges when using ASR for creating indexes. [3] Consequently, spoken content retrieval systems may have to suffice with ASR text transcripts with error rate as high as 50% [11]. Additionally, although ASR transcript generation times have decreased in recent years due to hardware improvements, the computation time is still something which should be considered, especially with respect to the huge amounts of available audio data which needs processing [3].

Uncertainty of how accurately an ASR system recognises the spoken words is an additional challenge to using ASR for spoken content retrieval. Furthermore, out-of-vocabulary words add to this challenge. How should the system deal with words which are not in it's language model and how should this problem be solved? Building a larger vocabulary for the ASR system is not a solution, since language is in constant growth and new words enter the vocabulary steadily. On a similar note, ASR systems depend on the language they have been trained on, and training an ASR system for many languages would require significant resources. Finally, it is important to keep in mind that speech media is not fully represented by its transcripts. Other aspects of the media will also influence the relevance of the search result to the user's information need. In the case of video, it is intuitively clear that the user probably has some idea of what kind of visual content their

search query should return. However, even with audio only documents, the user might prefer to attain results of a certain speaking style, certain speaker, a result of certain length, or speech media of certain quality and format. [11]

2.2 Stylistic Dimensions of Speech

The constantly growing collections containing speech data for entertainment, documentation, and research purposes have changed the needs of audiovisual content retrieval from generic categories such as sports and news to something more complex, where for example different kind of nuances in speech or emotions could be used to categorise the media. Therefore, ideas from paralinguistics and affective content analysis could potentially provide ideas to enrich the resolution in which spoken documents can be retrieved. By utilising the paralinguistic and affective levels of speech the documents could potentially be categorised not only in content categories (e.g. news, lectures, talk shows) but also in stylistic level (e.g. fast paced, calm, happy).

2.2.1 Paralinguistics

The term paralinguistics is generally used when referred to non-lexical properties of speech. These properties communicate information such as clues on the speaker's emotions and attitudes well as functional information on whether a spoken utterance is a question or where the focus of the utterance is. In contrast to paralinguistics, the term extralinguistics, refers to other speaker qualities which are informative in the sense of giving clues of the speaker's age, gender, and other physical conditions. For example, Laver [16] uses the term paralinguistic for signals of affective information through tone of voice and conversational interaction regulations, and the term extralinguistic for voice qualities identifying the individual speaker. Prosody, a term within paralinguistics, is a term to group together a collection of speech features such as pitch (fundamental frequency), loudness, and tempo or rhythm of the speech. All of these features can reveal useful information about the speaker's emotional state, personality, state of mind, or for example the speaker's gender and age.

2.2.2 Affective Content Analysis

Affective content is defined by [17] as those video or audio segments which can evoke strong emotional reactions such as cheer or fear in the viewer. Affective content analysis is a research field which aims to automatically detect such segments from various media types such as audio, text, video, or images. In [18] affective content analysis is defined as the interdisciplinary research space of computational linguistics, psycholinguists, consumer psychology, and human-computer interaction which looks at online communication in various forms. Being able to automatically label affective states in spoken content has the potential to increase user experience without the additional personnel costs otherwise required by manual annotation [19]. For example, ability to search for content with certain emotional tone (e.g. sad, happy, excited) could provide the user with increased

control over the retrieval process and increased accuracy of the retrieved results corresponding to the user's information need.

2.2.3 Emotion Theories

The most frequently used emotion categories in the field of video affective content analysis are Ekman's six basic emotions [20]: happiness, sadness, anger, disgust, fear, and surprise [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. These emotions are universal and are considered the building blocks of more complex emotions. They also have distinctive neural and physiological components, distinctive subjective experience, and distinctive regulatory and motivational properties [31]. Ekman reused the idea of universal emotions originally expressed by Darwin. In "The Expression of the Emotions in Man and Animals" [32] Darwin deduced the universal nature of basic emotions. The model assumes that the basic emotions are automatically triggered by objects or situations in the same way everywhere in the world.

Other commonly used emotion theories for affective content analysis are PAD (pleasure, arousal, dominance) emotion model and Component Process model. The PAD emotional state model is a psychological model developed by Albert Mehrabian to describe and measure emotional states. PAD model can be computationally more practical than Ekman's six basic emotions, since it affords quantification of emotions as well as models the dynamic nature in them. Thus, the PAD model is useful in situations where it is not feasible to categorise affect into discrete emotional categories. The PAD model uses three numerical dimensions: pleasure (P), arousal (A) and dominance (D) to represent all emotions on a continuous scale. The pleasure dimension is defined as positive versus negative affective states where higher evaluation of stimuli is being associated with greater pleasure. The arousal-non arousal dimension is defined in terms of level of mental alertness and physical activity. For example boredom and relaxation are defined to be on the lower end of the spectrum. Dominance-submissiveness dimension is defined as a feeling of control and influence over one's surroundings and others versus feeling controlled or influenced by situations and others. Examples of this dimension are anger, power, and boldness versus anxiety, fear, and loneliness. [33]

The Component Process Model developed by Klaus R. Scherer is another common emotion theory. This model emphasises the dynamic and recursive emotion processes which is evoked by an event central to the individual's needs, goals, and values either in a threatening or supporting way. The component process model breaks the emotion process to distinct components: appraisal, motivational effect, psychological and motor response patterns, multi-modal integration area, and categorisation and verbal labelling. [34] The component process model is used in for example studying emotion dimensions and format position because it is the only model of emotion that predicts testable effects of emotion on vocal expression [35].

The PAD model carries similarities to the energy-stress model of Thayer [36]. Thayer's model is based on Russell's model of affect [37]. In Thayer's model, the axes are energy (arousal, y-axis low to high) and stress (valence, x-axis from negative to positive), with four quadrants equating to contentment, depression, exuberance, and anxious/frantic. Thayer's model is especially adapted to analysing mood in music and is often used in studies which classify music according to emotions. Thayer's model is for example used in studies [38] and [39].

Chapter 3

Related Work

In this chapter, the related work on spoken content retrieval, affective content analysis, and on podcast content modelling are presented.

3.1 Spoken Content Retrieval: Applications

Spoken content retrieval (SCR) has been applied in many different application areas. One of the early use cases was gaining better access to broadcast news. First to radio news [40][41], and later to televised news [42]. Something specific to note about broadcast news is that they have a limited scope of topics, as well as a predefined structure. Both of these characteristics make the tasks of SCR easier than for example in the case of completely natural speech. Another important use case in the early phases of SCR was the search of spoken mail messages (voice messages) [43] [44] [45] [46].

Other, more recent research and application area of SCR, involves less planned, more spontaneous speech. This kind of speech could for example occur in the context of a conversation or other less formal settings. Examples of this new application area include search of meetings, [47] [48], call center recordings [49], collections of interviews [50][51], historical archives [52], lectures [53], and political speeches [54]. All of the previously stated works and systems are related to enabling searching and browsing spoken content media based on its topical content, but do not go into approaching spoken content from a stylistic perspective. However, it is important to note that the best methods for indexing audio data can differ depending on the goal of the system [55]. Thus, it is crucial to understand the user scenarios and how they could best translate to a highly effective SCR system. In "Spoken Content Retrieval: A Survey of Techniques and Technologies" [11] it is noted that the differences in user scenarios can comprise both the differences between user needs and the differences between the underlying spoken content collection, for example in terms of language, speaking style, topic, stability, and structure.

3.2 Affective Content Analysis

This section presents a short summary on the often used acoustic features for affective content analysis, briefly presents two examples of studies where music was grouped according to different emotions, and introduces couple of studies on detecting intense moments or highlights from audio- and audiovisual content.

3.2.1 Approaches and Features in Affective Content Analysis

There are many different methods and approaches to automatically extract affective content from auditory data as summarised in the article: "Video Affective Content Analysis: A Survey of State-of-the-Art Methods" [17]. The article presents an overview of methods and features used for affective content analysis of videos. For our study, the visual aspect of videos is ignored. However, the common practises on how to analyse audio for affective content is of interest. The article, for example, explains that audio can be divided into music, speech, and environmental sounds. Performing such division is recommended as a first step for carrying out affective content analysis on the audio. The articles [56] and [57] present work on segmenting audio to speech, music, sounds, and silence. The articles [58] and [59] describe work on separating voiced and unvoiced segments by using for example zero crossing rate together with other audio features.

After the audio has been segmented to these different audio types, audio features correlating with emotions can be extracted. For example, psychological research has shown that psychophysiological characteristics like air intake, intonation, and pitch characteristics vary with emotions in speech [60]. Based on psychological studies many prosodic speech features such as loudness, speech rate, pitch, inflection, rhythm, and voice quality have been used for emotion recognition from speech [17]. The studies [61] and [62] show that inflection, rhythm, voice quality, and pitch are commonly related to valence, while loudness and speech rate relate to arousal. When it comes to discrete emotions features such as pitch levels can potentially indicate feelings such as astonishment or boredom [17]. On the other hand, a study in [63] shows that spectral features, like Mel-Frequency Cepstrum Coefficients (MFCC), are also effective features when estimating the continuous values of emotions in 3D space (PAD model). Speech features are commonly the most used and useful acoustic features for affective content analysis, followed by music and environmental sounds.

3.2.2 Emotion Detection in Music

The area of emotion detection in music is interesting because music can be consumed very similarly to podcasts; it can be listened to for entertainment, and for pastime when for example commuting from one place to another. Therefore, methods from music content modelling could potentially be useful for better understanding of podcast content.

In [38] music is categorised according to its mood clusters, valence and arousal. The study bases its mood clusters on Thayer's emotion theory [36] on two dimensional axis of stress/valence and energy/arousal. The study used 80 music clips where 20 clips represented each of the four mood clusters (contentment, depression, exuberance, anxious/frantic). The labelling of the clusters

was done manually by the authors. The train and test sets consisted of 20 second clips of music. The music files were down sampled to 16kHz, 16 bit mono, and further segmented into frames of 32ms spanning the duration of each clip. The intensity, timbre, and rhythm were extracted and used as input for a Gaussian Mixture Model classifier. In [39] the emotional dimensions of arousal and valence are detected in music for the purpose of enabling search and recommendation of music based on the music's mood. The study used intensity as a feature for detecting arousal, and rhythm regularity as well as tempo as features to detect valence. As in the aforementioned study [38], this study also used Gaussian Mixture Models as the chosen classification tool. This method yielded an average precision and recall of 80%.

3.2.3 "Hot Spots" and Other Highlights in Audio Media

In addition to the direct application of SCR, some studies exist on how a part of SCR could be made more effective, or how the capabilities of SCR could be enriched. For example, a study [64] from 2003 assessed if "hot spots" could be detected from audio recorded meetings by both humans and potentially by automatic computation. The study's motivation stemmed from the desire to find ways to summarize, browse, and retrieve important information from lengthy archives of spoken content. In the study, "hot spots" were defined as "regions in which participants are highly involved in the discussion", for example heated arguments or points of excitement. The results showed that locations of "hot spots" are agreed upon by human listeners. Additionally, inspecting prosody features showed that there are differences in F0 and energy based feature values of voiced segments between of "hot spots" and non-"hot spot" regions of the meeting. These differences could potentially be used for automatic detection of "hot spots". A study [65] from 2008 also concentrated on detecting "hot spots". It modelled vocal interactions for text-independent detection of involvement in multi-party meetings. The study found that laughter is an important component in detecting "hot spots".

Another study focused on detecting sports highlights from audio and video media [66]. This study used features such as pitch, F1–F3 center frequencies, and spectral center of gravity which were extracted from the commentators speech to measure the "level-of-excitement" in the media. According to the same study, previous research on audio-based features for detecting excitement have focused on detecting broad events like cheering, music, applause, or speech characteristics, and employed this information with heuristics to identify exciting plays.

3.3 Related Work on Podcasts

Research on podcast search and recommendation is still a relatively new field. The existing studies have mainly focused on the use of podcasts in different contexts, rather than on browsing, searching or recommending podcasts [67] [10] [68] [69]. These studies have explored the use of podcasting in specific domains such as in education or mobile environments. Additionally, the motivation for listening to podcasts have been studied by for example Chung in [10] where quite general motivations such as "voyeurism/social interaction/companionship", "entertainment/relaxation/arousal", "education/information", "pastime/escape", "habit", and "convenience" were identified. At first,

in this section, a study on genres for internet content is briefly presented. Then, studies on podcast research goals, on predicting podcast popularity features, and on non-textual characteristics of podcasts are presented.

3.3.1 Genres for Internet Content

The study "Assembling a Balanced Corpus from the Internet" from 1998 possesses similar research goals as our research. Therefore, it is interesting to review the methodology of this study. The study aims to understand what text based genres and styles users think exist on the World Wide Web (WWW) in 1998 and develops an automatic system for classifying content to these genres. The study is carried out by first sending out questionnaires to students in Stockholm University and Royal Institute of Technology in Sweden, asking them about which genres they think WWW comprises of. The answers are subsequently grouped to 11 genres, after which the 11 genres are sent to the participants of the questionnaire and checked against their opinion. Next, an automatic classifier is build to sort text to these 11 categories: “Informal/Private”, “Public/Commercial”, “Searchable indices”, “Journalistic materials”, “Reports”, “Other running text”, “FAQs”, “Link Collections”, “Other listings and tables”, “Asynchronous multi-party correspondence”, and “Error messages”. The classifier uses different textual features of language such as word frequencies, certain keywords, and type/token ratios as input features to the classifier. The study concludes that the impressions users have of genre can be elicited and to some extent formalized for genre collection. [70]

3.3.2 Podcast Retrieval Engine Optimisation According To User Search Goals

Besser’s study [71] aimed to discover how podcast retrieval can be better optimized to meet the needs and search goals of users. First, a user study was conducted and then experiments on different podcast retrieval systems were carried out. The user study was designed to identify users’ underlying goals in podcast search, the strategies used to gain access to podcasts, and how currently available tools influence podcast search. First, an online survey was conducted and then it was used as a basis for conducting diary studies and contextual interviews. After this, experiments were conducted on different types of podcast retrieval systems to determine how they performed in respect to the identified user goals and needs.

The user study revealed that podcasts are often seen as rather personal sources of information, often displaying the podcast host’s personal views and opinions instead of stating purely factual information. The study participants were found to favor podcast topics such as technology, news, and entertainment.

A variety of search strategies were used to gain access to podcasts, such as query-based search, directed and undirected browsing, and requested and unrequested recommendations. For query-based search, Besser identified a set of categories classifying the underlying search goals. The categories consisted of searches for person names, searches for podcasts for which the title or a quotation from an episode is known, and searches for information about a general topic or about a current issue or event. The study also found that goals and search strategies for podcast search

may be strongly influenced by the perception of the capabilities of the current search systems, most notably the perceived lack of tools for online audio search.

The experiment phase of this study tested retrieval performance with respect to the newly identified user search goals. The experiments showed that the performance of three different types of podcast retrieval systems, one based on metadata, one based on content indexing, and one on mixed systems, vary depending on the search strategy. For example, metadata indexing outperformed content indexing for title queries.

3.3.3 Features Predicting Podcast Popularity

In [72] Larson addresses the challenge of automatically identifying which podcasts have the highest potential for listener appeal. The study proposes a framework of features which could be used to predict listeners' podcast preference. The features are sorted to categories which are: "Podcast Content", "the Podcaster", "'the Podcast Context", and "the Technical Execution" of the podcast. "Podcast content" consists of features such as whether the podcast has a strong topical focus or appearance of (multiple) on-topic guests. "The Podcaster" category comprises of qualities of the host such as speech rate, use of conversational style, presence of affect, use of humor, and lack of hesitation. "The Podcast Context" includes features such as podcast page or metadata contains links to related material or that the podcast makes references to current events. "The Technical Execution" comprises of qualities of the technical execution of the recordings, such as signature intro/opening jingle, background music, studio quality recording (no unintentional background noise), editing effects (e.g. fades and transitions). In the light of the research goals of our study, the categories of "The Podcaster" and "The Technical Execution" are the most interesting for us, since many of the aforementioned features (for example speech rate) could be automatically detected from the acoustic features only.

3.3.4 Non-textual Characteristics of Podcasts

Extracting non-textual characteristics of podcasts from the podcast audio signal was explored in [5]. They experimented both with different audio features and automatic classification models, as well as two different podcast characteristics; namely seriousness (funny vs serious) and energy (energetic vs non-energetic). The study presents an Adversarial Learning-based Podcast Representation (ALPR) that captures non-textual aspects of podcasts and can predict the seriousness and energy of podcasts. The study also reveals factors which correlate with podcast popularity.

The study collected a podcast dataset of 88,728 episodes and crowd-sourced labels for a randomly sampled subset by using Amazon Mechanical Turk platform. The audio snippets were manually labelled with seriousness and energy scores by using scales which ranged from calm to energetic and from humorous to serious, respectively. The non-textual characteristics, seriousness and energy, were chosen based on an iTunes analysis and based on published literature. 850K iTunes reviews of podcasts were collected and the most mentioned adjectives of the most highly ranked podcasts were examined. The top adjectives were funny, entertaining, and hilarious. This lead the authors of the study to research detecting the non-textual characteristic of seriousness in podcasts.

The non-textual characteristic, energy, was chosen for the study based on a literature review of music recommendation literature [73], [74], which suggested that rhythm, on a scale from fast to slow, and how energetic the sounds are, on a scale from energetic to calm, are important attributes for context-aware music recommendation. Since consuming contexts for music and podcasts are similar, energy was selected as a non-textual podcast characteristic for the study with a prediction that it might be important for a podcast listening experience.

First, the study experimented with existing audio modelling techniques. The used audio features included MFCC [75], IS09 [76], and IS13 [77], and the detection algorithms utilised standard deep neural network based representation learning frameworks [78]. The study found that these methods resulted in sub-optimal prediction performance, because they could not capture complex variations in podcasts. To address this limitation the study experimented with adversal learning [79] and investigated an unsupervised learning algorithm that progressively builds podcast representations from fine-graded spectrogram details. Only the maximum of first 10 minutes of each podcast episode were used for the experimentation. The study found that their adversarial learning-based podcast representation captures subtleties of complex audio spectrograms and achieves significantly better performance in predicting non-textual attributes of energy and seriousness than the first explored methods.

The study also conducted some podcast popularity prediction experiments and found that incorporating the adversarial learning-based podcast representation into topic based popularity prediction lead to a significant performance gain. The study revealed some factors that correlate with podcast popularity. For example, the perceived energy correlated positively with popularity. Also topics related to family, politics, crime, and food were positively correlated. Extensive use of functional words was negatively correlated with popularity.

3.4 Conclusions of Chapter 3

Only recently the field of spoken content retrieval has moved to research retrieval and applications of audio media which contains unstructured natural speech, of which podcasts are an example. However, research on unstructured natural speech has concentrated on enabling searching and browsing spoken content media based on its topical content, neglecting the other stylistic dimensions of spoken content such as paralinguistic information. The best indexing methods can differ depending on the goal of the retrieval system. Therefore, it is important to consider additional ways of enriching podcast representations beyond the topical content.

In affective content analysis a common practise is to divide audio into music, speech and environmental sounds before doing any other analysis. Psychological research has shown that prosodic features like speech rate, intonation, and pitch characteristics vary with emotions in speech. This had lead to many studies using prosodic features for emotion recognition from audio. Additionally, spectral features like MFCCs are effective for affective content analysis. Work exists on emotion detection in music. Also other kind of affective content, “hot spots” and other highlights have been detected in audio media. These regions of high engagement have been detected in meetings and in sports’ audio and video media. However, work on emotion detection in podcasts is close to non-

existent.

Related work on podcasts has mainly concentrated on the use of podcasts in different contexts rather than on browsing, searching or recommending podcasts. However, in [72] a set of podcast characteristics for predicting listeners' podcast preference were compiled. These characteristics could potentially be useful for personalised podcast search and recommendation experiences. In [5] non-textual characteristics, seriousness and energy, of podcasts were automatically detected from spectral features of the first 10 minutes of podcast episodes' audio. Both of these studies [72], [5] have researched podcasts from the perspective of popularity. Our study builds on the stylistic characteristics identified in [72] and [5] but does not limit itself to popularity prediction. Instead, it aims to understand stylistic characteristics of podcasts in a wider context from the perspective of podcast listeners' listening experiences.

Chapter 4

Research Question 1: User's Perspective

In this chapter the research question RQ1 and the selected method are presented in more detail. Additionally, the results and the analysis of the results of the first phase of the research are presented.

4.1 Objectives

The goal of this research phase was to understand if podcast listeners care about the stylistic characteristics of podcasts, and if so, what stylistic characteristics listeners find interesting or important for their podcast listening experience. Hence, this chapter answers to the research question **RQ1: "What stylistic characteristics of podcasts do listeners find interesting or important for their podcast listening experience?"** We narrow down the scope of stylistic features to features which can be observed by listening to the podcast audio and which can be verbalised by the podcast listeners. We broke Research questions RQ1 down to sub-questions, which can be seen below.

> RQ1.1: What stylistic features can listeners observe and verbalise in podcasts? Both in terms of what features they like and do not like in podcasts.
>
> RQ1.2: What are listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about?
>
> RQ1.3: How well do stylistic features identified from literature and by experts match listeners' perceptions on stylistic features and categories?

RQ1.1 refers to the middle level stylistic features from our framework. RQ1.2 refers to the top level stylistic categories from our framework and RQ1.3 refers to both, the stylistic features and the stylistic categories from our framework.

4.2 Methodology

The study employs a qualitative user research approach in order to answer the above research questions. This choice was motivated by the research aim of understanding podcast listeners' perspective on stylistic podcast characteristics on a detailed level. Consequently, we needed as many ideas and opinions on stylistic characteristics of podcasts from the study participants as possible. One of the better ways to do this is via participatory design workshops. However, we are aware of the limitations of this approach, in particular that the results of this approach cannot be generalised to a wider population, but that they need a further quantitative analysis for validating their potential for generalisation beyond the study sample. The selected approach for the qualitative user study was to carry out three participatory design workshops with three participants for each workshop. This lead to altogether nine participants for the study. Participatory design is an approach to engineering technological systems that seeks to improve them by including users in the design process. It is motivated primarily by an interest in empowering users, but also by a concern for building systems better suited to user needs. [80] The workshops were carried out in May 2020.

4.2.1 Participants

We recruited nine podcast listeners for our user research. The participants listened to podcasts at least once a month and came from various cultural and professional backgrounds. The represented nations were Finland, Germany, India, Latvia and Spain. The professional backgrounds were dental care, electrical engineering, interaction technology, law, corporate communication and event organization, economy, human computer interaction, and medicine. Two of the participants were male and the rest female. The participants were 20-29 years old and recruited from the social network of the main author. The sample age was motivated by podcast usage being typically driven by a younger generation [4].

The frequency of podcast listening among participants varied from daily listening to the said minimum of once a month. Most of the participants listened to podcasts at least once a week. The types of podcasts the participants usually listened to varied from science, news, politics, economics, documentaries, true crime, and culture to humour and gossip podcasts, football, advice shows, entertainment podcasts (e.g. celebrity gossip, funny stories), and to lifestyle podcasts related to meditation and philosophy. The participants reported that they listened to podcasts from Spotify, YouTube, Google Podcasts and Yle Areena. The demographic and podcast listening information were collected from the participants anonymously by using a short online survey.

Another participant recruitment criteria was that the participants were able to express themselves in English without limitations due to the study being carried out in English. The participants were also required to have sufficient knowledge in IT and access to a personal computer in order to join and participate in the online workshops.

4.2.2 Tools and Materials

The workshops were carried out in an online format due to the outbreak of the COVID-19 pandemic which lead to a work from home policy of the host organisation. The workshops were carried out by using Google Hangouts, Google Drive, and Mural. Google Hangouts is an online conference tool which allows real time video calls with a group of several participants, as well as sharing a participant's screen and recording the meeting's audio and video. Google Hangouts was used for meeting the participants and for orally guiding them through the workshop exercises. Google Drive is a cloud storage solution which users can use to upload, store, and share content. Google Drive was also used to make selected podcast episodes available for the participants so that they could listen to the episodes during the workshop. Mural is an online tool which people can use to collaborate in real time and to write and arrange notes and text. It has many functionalities for facilitating a brainstorming session such as a timer, ways to hide and show content, and to "summon" all the participants to the part of the board where the facilitator is. Mural was used to organise and facilitate writing down and organising sticky notes during the workshops. Mural was also used later on in the analysis of the workshop results as a tool to organise and summarise the workshop results.

Fourteen podcast episodes were selected for the workshops. Participants were instructed to listen to these podcast episodes during the first workshop exercise (workshop exercises are described in more detail in Section 4.2.3) in order to observe what stylistic features they cared about in podcasts. This set of episodes was selected for the workshops as opposed to the participants freely browsing and listening to any podcasts they find in order to guarantee that the participants listened to many different types of podcasts from different podcast genres. The selection of the podcast episodes was done in the following way. First, a set of interesting stylistic features was gathered. These features (Table 4.1) were selected based on a existing literature on a framework for predicting podcast preference [72]. Additionally, we added some of our own expert ideas (Table 4.2) on potentially interesting stylistic features to the set. The stylistic features were later introduced to the workshop participants in the third workshop exercise to test which of these stylistic features participants found relevant for their podcast listening experience. During the workshop podcast selection and the third workshop exercise all of these stylistic features were handled together as one set. In other words, any distinction between stylistic features from existing literature and our expert ideas was not made.

Table 4.1: Stylistic features from existing literature.

FROM EXISTING LITERATURE		
Audio quality (low, high, varying)	Signature intro / opening jingle	Background music (with/without simultaneous talking)
Atmospheric sound/Sound effects	Editing effects (e.g. fades, transitions, music to signal change of topic or situation)	Ads present
Showing emotion	Speech rate (slow, medium, fast)	Different personalities/attitudes
Multiple people	One person	

Table 4.2: Expert ideas on interesting stylistic features.

EXPERT IDEAS		
Excitement	Clapping	Laughter
Swearing	Other vocalisations "wowow", "gggshh", "yea", "mm", "ha", "arrr"	Social multi-party conversation
Interruptions	A lot of silence	Conversational style
Serious	Light mood	News reading/reporting style
Explainer/Lecturing	Informative	Storytelling
Monologue	Interview	Chat
Calm	Boring	Lively
Discussion	Gender (females, males, mixed gender)	Age

The workshop podcast episodes were collected by looking for podcasts while paying attention to the merged set of stylistic features (Tables 4.1 and 4.2). The goal was to have a small collection of podcasts which vary a lot in style. The small size of the podcast set was motivated by the need to be able to browse through it during a workshop. The podcasts were collected by browsing different podcast genres in the music and podcast service Spotify. A visual feature matrix was created to represent which podcast episodes had which stylistic features in them. This way it was easy to verify visually that all the interesting stylistic features were covered by the selected podcasts and that the podcasts contained different stylistic features from each other. Both the main researcher and the academic thesis supervisor listened to the podcasts individually and marked down the stylistic features they observed in the podcast episodes in order to create this matrix. It should be noted that the feature matrix does not represent the absolute truth on what stylistic features each of the selected podcasts contain, but presents the opinions of the main researcher and the thesis supervisor. This method is justified by the purpose of simply collecting a set of varying styles of podcasts without making any attempts to present precise claims on the stylistic features for each podcast. Thus, we argue that the carried out expert analysis was sufficient for this purpose.

Once the podcast episodes were selected for the workshops, they were renamed with random numbers so that the name of the podcast episode would not influence the participants' perceptions of the podcast. The selected podcast episodes, the podcast shows they belong to, their genres on Spotify, and the random numbers (rand) used for re-naming the files can be seen in Table 4.3

Table 4.3: Podcast episodes selected for the workshops.

EPISODE	SHOW	GENRE	RAND
Editor's Picks: March 16th 2020	Economist Radio	Politics	52
Checks and Balance: Getting a grip	Economist Radio	Politics	69
How Grassoline Works	Stuff you should know	Educational	11
385- Shade	99% Invisible	Educational	42
#406: Bob Iger - CEO and Chairman of Disney	The Tim Ferriss Show	Business & Technology	45
Estee Lauder - "Dedication"	Great Women of Business	Business & Technology	23
Taking Her on a First Date	VIEWS with David Dobrik and Jason Nash	Comedy	50
Dealing Drugs in Miami	VIEWS with David Dobrik and Jason Nash	Comedy	43
A Hawaiian Hookup culture & A Wedding Showdown	Anna Faris Is Unqualified	Comedy	55
ICYMI - "Porch Pirates" Steal Holiday Packages	The Daily Show With Trevor Noah: Ears Edition	Comedy	90
Emmy - Pregnancy Test	Everything Is Alive	Comedy	70
Sebastian, Alex and Alex, Russian Dolls	Everything Is Alive	Comedy	22
859 - Sleepy Ticket to Ride | Bored Game Unboxing	Sleep with me	NA	13
bargain	Knifepoint horror	NA	4

4.2.3 Procedure

The workshops were conducted in English. Each workshop took two to three hours and were conducted in groups of three participants. The workshops focused on answering the questions: RQ1.1. What stylistic features can listeners observe and verbalise in podcasts? Both in terms of what features they like and do not like in podcasts. RQ1.2. What are listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about? RQ1.3. How well do stylistic features identified from literature and by experts match listeners' perceptions on stylistic features and categories? Three corresponding exercises were designed to produce answers to these three questions, respectively. The participants were guided through the exercises one exercise at a time, and the following exercises were revealed only after the previous exercise was completed. The first half of the workshop (~1hour) was spent on introduction, ice breaker, the first exercise, and a break. The second half of the workshop (~1hour) was spent on the second and third exercise and wrap up of the workshop. Below are descriptions of each exercise.

Exercise 1: Stylistic Features

The first exercise was designed to answer What kind of stylistic features listeners can observe and verbalise in podcasts. In the first exercise the participants were instructed to individually listen to the given 14 podcast episodes. The participants were instructed to make note of things they liked or did not like in the podcasts other than the discussed topic or content and write these observations down while listening to the 14 podcasts. The observations which were pleasant to the participants were written on pink sticky notes and the unpleasant observations were written on blue sticky notes. The participants were instructed to write down one observation per sticky note and also write the name of the podcast audio file on the sticky note, so that the observations could be mapped back to the podcast episodes. After around 10 minutes of listening and writing observations the partici-

pants were instructed to share their observations and write down clarifications on the sticky notes if needed. After that the participants had another round of listening to the 14 podcasts, writing down more observations and discussion on their observations. After the first exercise the participants had a break for about 15 minutes.

Exercise 2: Stylistic Categories

The second exercise was designed to answer what are listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about. In the second exercise the participants were asked to look at the sticky notes they had created during the first exercise, discuss the sticky notes, and together group the sticky notes to categories which made sense to the participants. An empty scale on which to place the sticky notes was provided in order to help the participants to reflect on and group the sticky notes, as well as to hopefully provide an additional level of detail through which to understand the stylistic categories. The use of a scale as part of the structure of the stylistic categories was inspired by the use of scales "funny - serious" and "energetic - non-energetic" in [5] for non-textual podcast characteristics of seriousness and energy. The participants were instructed to freely name the dimensions of the scale as well as assign a name to the category. The participants had the freedom to not place all the sticky notes to the categories, if the placement did not seem suitable for them. They also had the freedom to add things to the sticky notes or create new sticky notes if they felt like something was missing from the categories they created. The participants were instructed to create as many categories out of the sticky notes as they saw fit. The participants were also encouraged to discuss the groupings with each other as much as possible. Around 30 minutes of time was reserved for this exercise, but if the participants needed more time to discuss the sticky notes' placement, more time was granted.

Exercise 3: Stylistic Features from Existing Literature and from Experts

The third exercise was designed to answer how well do stylistic podcast features identified from existing literature and the stylistic features suggested by us fit to listeners' perceptions on stylistic podcast categories and their perceptions on which stylistic features are important for their podcast listening experience. In the third exercise, new sticky notes on yellow background were introduced to the participants. The yellow background was chosen in order to differentiate these sticky notes from the participant generated ones. The yellow sticky notes contained stylistic features picked from previous research on podcasts [72] as well as the exert ideas on what might be interesting features. This set of stylistic features were the same set as which guided the selection of the podcast episodes for the workshops (Tables 4.1 and 4.2).

The participants were told to go through the yellow sticky notes, discuss them, and if the yellow sticky notes fit to the categories from the second exercise, group them to these categories. The participants were given the freedom to create new categories for the yellow sticky notes if they thought that the sticky notes were relevant for their podcast listening experience, but that the sticky notes did not fit to any of the existing categories. The participants were also instructed to move aside any yellow sticky notes which they found irrelevant for their podcast listening experience.

During the grouping of the sticky notes to the categories participants were encouraged to discuss the groupings with each other as much as possible. Around 30 minutes were used for this exercise. Again, if the participants required more time to group and discuss, it was granted.

4.2.4 Analysis

The workshop results from the first exercise were analysed with the help of Mural by gathering all the sticky notes together, grouping the sticky notes based on the stylistic observation described on them, and counting the number of sticky notes in each formed group. For example, all sticky notes which mentioned music were organised to a group called "music". The frequency of sticky notes in each such group were calculated. This number was then divided over the total number of sticky notes to calculate the percentages of the sticky notes in each group. The percentages were plotted into a bar plot to visualise which type of observations were most frequent. The grouping was done independently by two experts, the author and the author's industry supervisor, who had not seen the author's groupings prior to his grouping.

The results from the second and third exercise were analysed by manually going through all the participant generated categories and by observing what kind of categories were created, and what kind of sticky notes were placed under the category. After going through the categories, similar and overlapping categories were combined by the main researcher. This was done in order to summarise the results to a comprehensive form. For example, the categories from the second and third exercise called "Ways of Speaking" and "Speaking" had sticky notes in them which were all related to speaking style and voice qualities. Examples of such sticky notes were the speech rate of the podcasters, clear speaking style or unclear voice, swearing, and other speech related observations. Furthermore, both of these categories had similar scales. The "Ways of Speaking" scale was labeled "annoying - enjoyable" and "Speaking" was labeled "unpleasant - pleasant". Hence, these two categories were combined to one category, by the main researcher, called "Speaking Style and Voice Qualities". The other categories were analysed by following a similar prosedure.

Finally, an example of a filled in framework of stylistic podcast characteristics was given by completing it with the workshop results of one of the summarised stylistic categories. This was done in order to demonstrate how the workshop findings fit to the framework defined by this study.

4.2.5 Original Plan and Influence of COVID-19

The workshops were originally planned to be carried out face-to-face in a conference room with physical sticky notes. The idea was that the hands on exercises where the participants would be writing physical sticky notes and moving around a board to place the sticky notes, would help the participants stay engaged and stimulate creativity. However, a couple of weeks before the planned workshop dates, the COVID-19 virus spread to Europe and caused a global lock-down. Due to world-wide governmental guidelines to social distance oneself and a ban on conducting on-site user research, the workshops were forced to move to an online setting. This decision was also affected by the relatively tight schedule of the thesis project, since it was estimated that the lock-down situation would last past the set end date for the research and report writing.

However, the online format was not detrimental to the workshops to any large degree. The online tool used for writing down and grouping sticky notes, Mural, fulfilled the same function as writing sticky notes and moving them around on a physical board, expect without the need of being in the same room. Furthermore, we found that the online format had many benefits. It was easier to recruit people from various backgrounds due to not having to limit our recruitment to Stockholm where the main researcher was situated. It was also easier to make the podcasts available to the participants for listening, as they provided their own computers and audio equipment, and to save and process the results of the workshops in a digital format. Additionally, it was easier to facilitate conversation on the generated sticky notes due to the researcher having a constant overview available on all the sticky notes, as well as the ability to zoom in and out of the sticky notes.

4.3 Results

This section reports the results of the first phase of the research. It walks through the results of the three workshop exercises and by doing so answers the three sub-questions of the research question RQ1.

4.3.1 Exercise 1 Results: Stylistic Features

In the first exercise the participants were asked to write down stylistic features they observed in the podcasts they listened to during the workshop. They were asked to write the observations they liked on pink sticky notes and the observations they did not like on blue sticky notes. Example of what kind of observations the participants made can be seen in Figure 4.1. The sticky notes from the workshops were gathered together and grouped based on the observations written on the sticky notes. This was done in order to analyse the frequency of each observation. An example of the grouped sticky notes can be seen in Figure 4.2. The percentage of the sticky notes in each group over the total number of sticky notes was calculated and visualised in Figure 4.3. This helped us to gain a quick overview of which kind of stylistic features were mentioned most often by the participants. In the next section, we report the stylistic features in order from the most mentions to the least mentions.

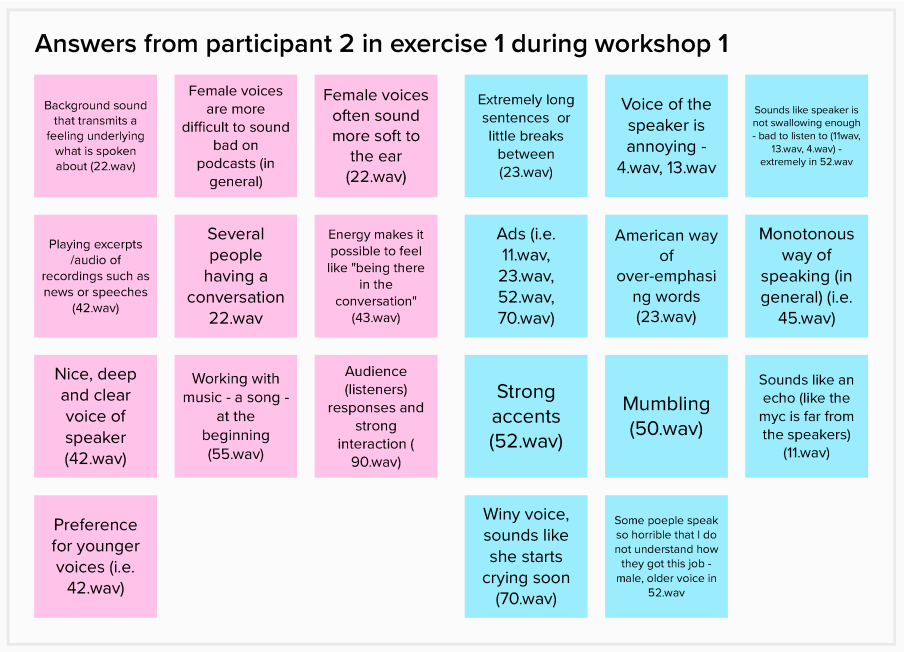

Figure 4.1: Workshop answers of one participant from the first exercise. The liked features are written on pink, disliked features on blue notes.

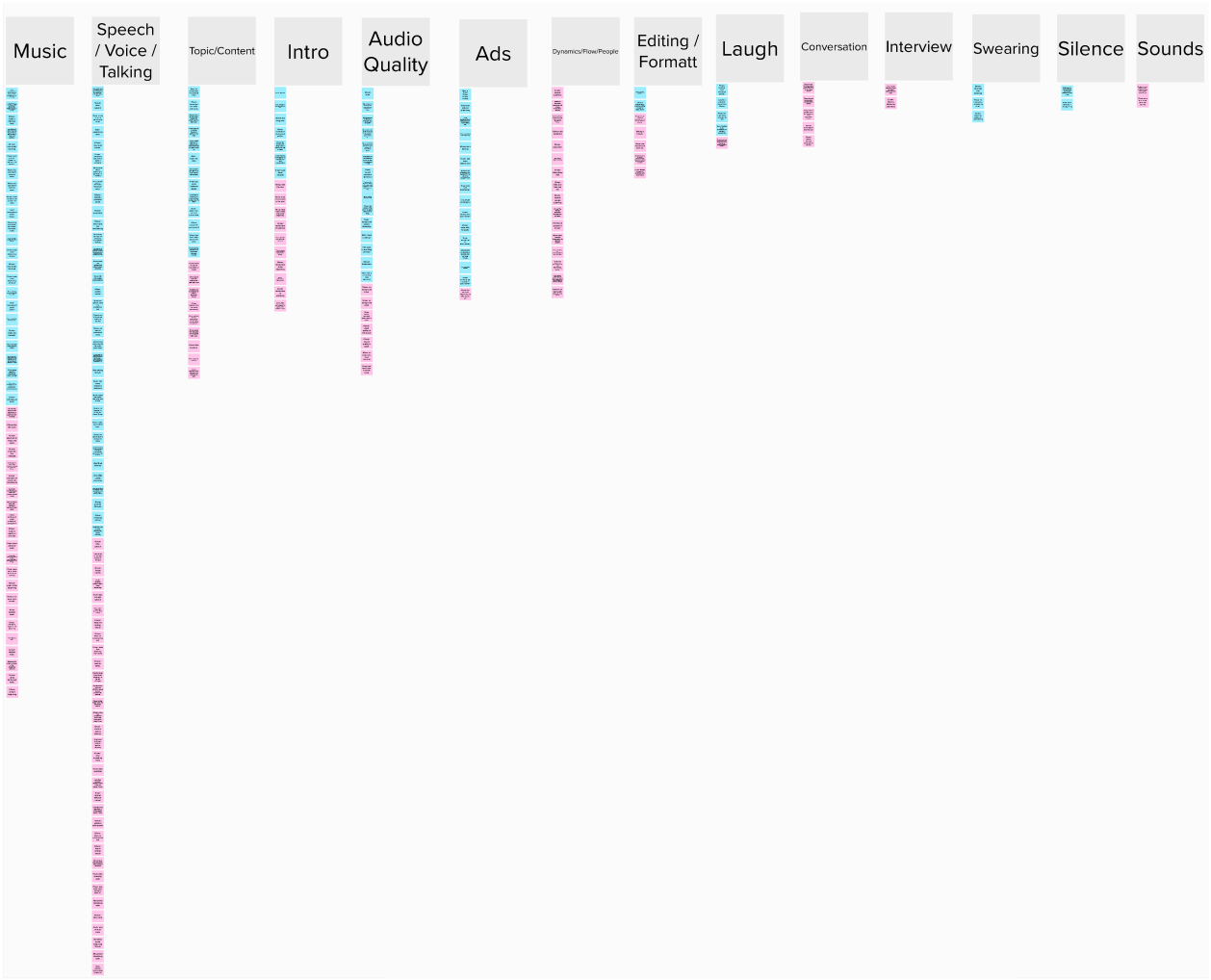

Figure 4.2: Sticky notes from the first exercise grouped by the type of observation written on the sticky note.

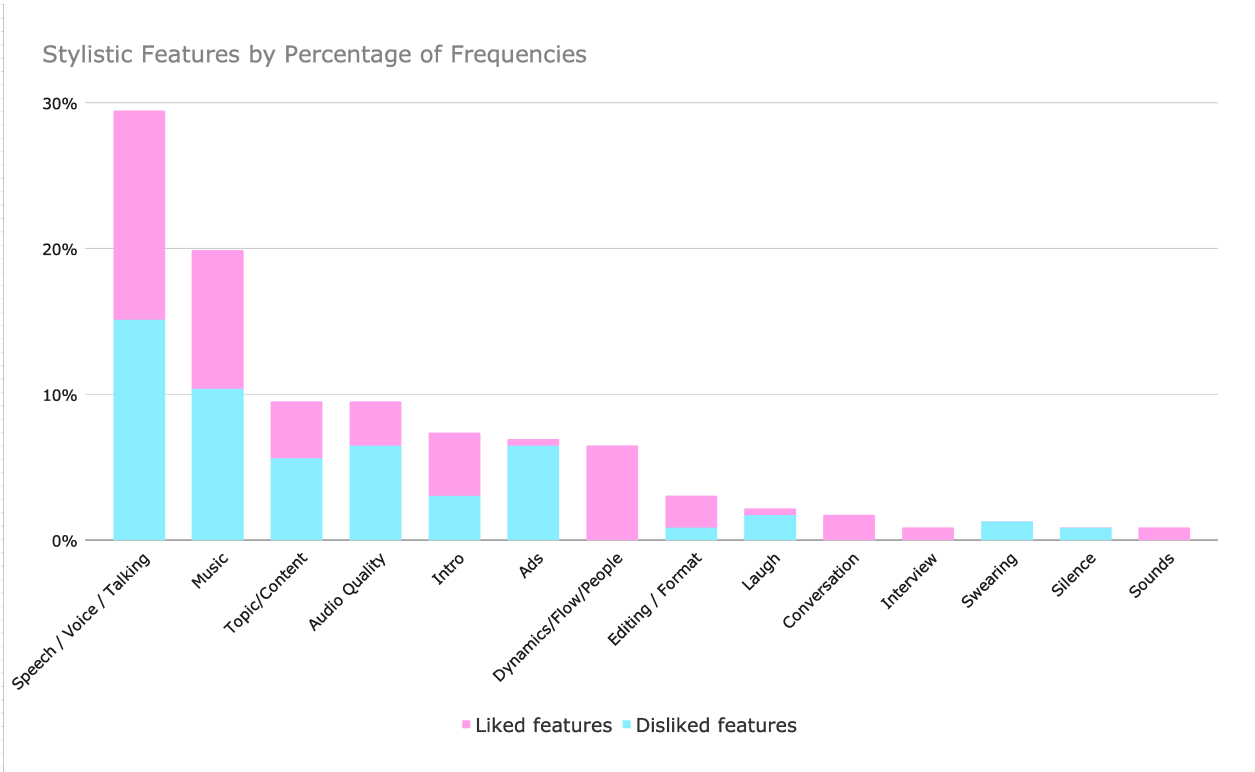

Figure 4.3: Stylistic features from the exercise 1 by their percentages over the total amount of notes.

General Features

All together 231 sticky notes were generated during all the three workshops. Most of the stylistic features observed by the participants were related to the **speech, voice, or talking style** of the people in the podcasts. 68 sticky notes, i.e. 29% of the sticky notes, were written about things related to this category. Some examples of sticky notes in this category were: "clear speech", "happy voices", "hosts are pronouncing well", "female voices are more difficult to sound bad on podcasts (in general)", "melodic voice is easy to listen to", "slow speech", "monotonous way of speaking", "whiny voice", and "talk with unclear pronunciation".

The second most observations were made on the presence of **music** in the podcasts. 20% of the sticky notes were about music and contained comments like: "music fits the message", "background music not overbearing", "background music creates an atmosphere", "energetic music in the beginning", "suitable music", "Music/voice mix; the music becomes quieter before the presenter speaks and rises afterwards. Creates unnatural feeling. There are better techniques to combine them.", "music a little too loud", "intro music too long", "loud background music", and "using repetitive background music too long".

Despite of the repeated instructions on not concentrating on the topic or content discussed about the participants generated many topic related sticky notes. **Topic and content** related sticky notes made up 10% of all the generated sticky notes. Observations grouped to this category were such as: "explaining the background for topic of discussion", "speaker tells the topics beforehand", "good

that they give out the genre right away", "cannot get the topic right away", "they sort of rumble instead of getting to the topic", "would like them to indicate what the podcast and its topic is early on", "too many irrelevant details", and "they don't really stick to the point".

Audio quality related observations also reached to (10%) of the total sticky notes. These mentions were for example as such: "no background noise", "quality sounded good, easy to listen", "sound quality is good", "well mixed voice loudness levels", "mic placement is improper, also the quality", "sounds like an echo (like the myc is far from the speakers)", "sound sounds a bit echo-y", "background noise is distracting", "static noise in ", and "there is a humming noise background".

Also the **introduction** of the podcasts was important to the listeners and they paid attention to it relatively much. 7% of the total sticky notes were podcast introduction related observations, for example: "good intro, straight to the point", "introduction of speakers", "liked the intro, looked like news in tv", "spontaneous, different intro", "too long intro", "introduction too loud and sudden", "the intro is ambiguous for too long, so I lose concentration", and "it just starts abruptly".

Another somewhat content related feature the participants paid attention to was the presence of **ads**. 7% of the sticky notes were ad related and whereas almost all the other features had about equal distribution of positive and negative mentions, ads had predominantly negative mentions. The only positive mention about ads was "the ad in the beginning: it's clear it's an ad". Some of the negative notes were: "the ads are quite long", "advertising at the beginning", "annoying advertising at the end of the podcast", and "commercials at the start of podcast are irritating".

Other stylistic feature related sticky notes were about the dynamics of the people or flow of the podcast(6%), about the editing or format of the podcast(3%), about the presence of laughter(2%), whether the podcast was mainly in conversational style(2%) or if it was like an interview(1%), whether the podcasters' language contained a lot of profanity(1%), the silent breaks in the podcasts(1%), and other sounds(1%).

All the notes made about the **dynamic of the people or flow of the podcast** were positive. Some of them were: "several speakers", "audience (listeners) responses and strong interaction", "dynamic between 2 people (seem like friends, relaxed atmosphere)", "chill atmosfere", "interactive", "dialog flows well", "several people speaking", "voices sounded interested in the opponent speaker", "liked the excitement of the presenter. She seemed to be engaged with the topic and making the "interviewees involved", and "It felt like bunch of friends messing around and then talking about stuff".

Some examples of the **editing or format of the podcast** are: "playing excerpts /audio of recordings such as news or speeches", "well recorded and good mix", and "quite long podcast". Examples of mentions of **laughter** are "laughing track overbears speaker", "laughing in the background was not ideal, seemed fake", and "hosts seem to know each other and can joke around in an authentic way (indicated by laughter and jokes)". A reason for relatively many negative mentions on laughter was that most of the comments regarding laughter were written about the same podcast episode, which was a stand up comedy show with a sitcom laughter track in the background. This was something the participants seemed to not like very much.

Comments related to **conversational style** and **interviews** were both only positive. The comments were the following: "feels that they just have a conversation about topic and not focused on

only making podcast", "natural conversation, people joking around", "energy makes it possible to feel like "being there in the conversation"", "natural conversation style", "several people having a conversation", "easy to follow the interviews", and "in the interviews, hosts are talking during their own turns".

Notes related to **swearing**, **silence**, and other **sounds** were the following: "do not like the swearing", "not crazy about the swearing either", "swearing; gave a bit trashy feeling", "long pause at the beginning, was unsure whether it starts", "silent seconds in the beginning", "starts with a nice set of bell sounds", and "background sound that transmits a feeling underlying what is spoken about".

Speech, Voice, & Talking Style Related Features

Since speech, voice, and talking style related stylistic features were clearly important for the participants and accounted for almost 30% of all the mentioned features, this group was further inspected in order to understand what kind of stylistic features within speech, voice, and talking style were important to the participants. The breakdown to sub-groups within speech related features can be seen in Figure 4.4 with some examples answers highlighted from each sub-group.

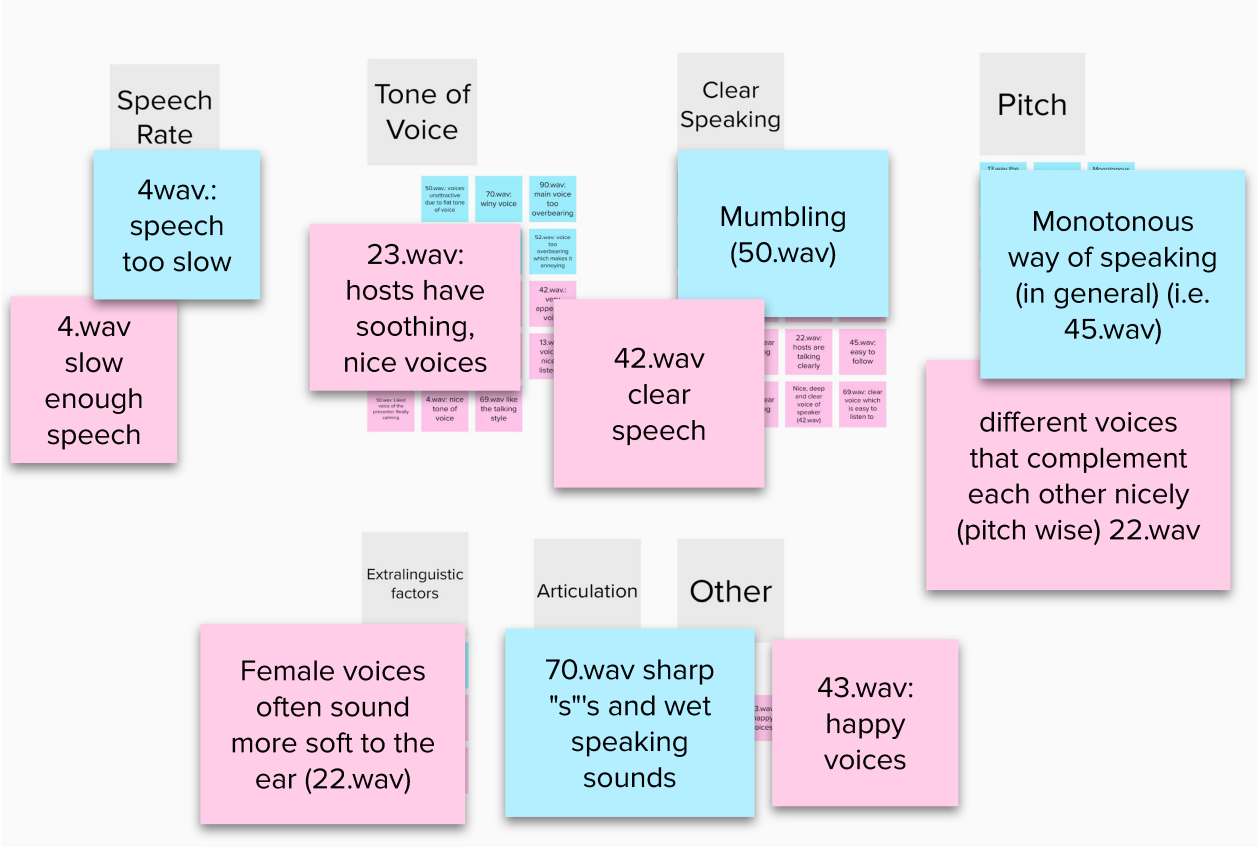

Figure 4.4: Separately inspected speech features from the group "Speech/Voice/Talking".

Mentions about **clear speaking** made up a sub-category of 21 notes and contained sticky notes such as: "clear voice", "proper talking, keeps you in the discussion", "clear pronunciations and structured sentences", "easy to follow", "hosts are pronouncing well", "clear voice which is easy to listen to", "clear speech", "reader has stuffed nose", "unclear voices", "talking over one another - hard to understand individual people", and "mumbling".

Tone of voice accounted for 18 of the sticky notes in the category "speech, voice, and talking style" and contained sticky notes such as: "very appealing voice", "hosts have soothing, nice voices", "melodic voice is easy to listen to", "liked voice of the presenter, really calming", "nice tone of voice", "voices unattractive due to flat tone of voice", "winy voice", "winy voice, sounds like she starts crying soon", "speaker sounded bored", and "quite dull. The content must be good but the tone of voice was quite boring and neutral, without any kind of excitement in the words".

There were eight mentions related to **pitch**. Pitch mentions contained sticky notes such as: "monotonous voice, sounds like audio book", "Monotonous way of speaking (in general)", "a long intro in a monotonous voice", "different voices that complement each other nicely (pitch wise)", "several different voices", and "different voices make it more interesting".

Speech rate related comments made up a sub-category of six notes with mentions such as: "too many breaks in sentence", "slow speech", "speaking is too fast - listening to it makes me feel rushed and agitated", "slow talking", and "slow enough speech".

Extralinguistic factors such as age, gender or accent consisted of 8 sticky notes with mentions such as: "preference for younger voices", "female voices often sound more soft to the ear", "some people speak so horrible that I do not understand how they got this job - male, older voice", "annoying accent", "strong accents", and "American way of over-emphasising words".

Articulation together with **other** features made up the rest of six notes. These categories contained sticky notes such as: "speaker breathing /inhaling/exhaling while talking", "sharp "s"'s and wet speaking sounds", "Sounds like speaker is not swallowing enough - bad to listen to", "very loud voices", and "happy voices".

Reflection on Colour of the Sticky Notes

The colour of the sticky notes generated in the first exercise reflect on the participants' preference for the stylistic feature written on the sticky note where pink notes indicate the liked features and blue notes indicate the disliked features. Figure 4.3 gives an overview of the distribution of the liked and disliked features. One can see from Figure 4.3 that almost all of the stylistic feature groups contain almost equal amount of mentions of both liked and disliked features with only a few exceptions. The exceptions were the following. Ads related features were seen almost entirely as negative. Stylistic features related to the dynamics, flow, or people were only seen as positive. Laughter in the workshop podcasts was mainly perceived as negative due to one specific podcast which contained sitcom type of laughter. This type of laughter seemed to not be preferable to our participants. Stylistic features related to the podcast containing conversations or interviews were seen as positive whereas swearing and silence were seen as something negative by our participants. Additionally, presence of sounds got only positive mentions.

We found that in general preference for certain kinds of features was highly individual among

the workshop participants. For example, as seen in Figure 4.4 two participants had made the same observation on the podcast "4.wav" that the speech rate is slow, but one of the participants liked the slow speech rate, the other participant disliked it.

4.3.2 Exercise 2 and 3 Results: Stylistic Categories and Stylistic Features from Existing Literature and from Experts

In the second exercise the participants were instructed to group the stylistic feature sticky notes they had generated during the first exercise to categories which made sense to them. The participants were asked to organise the sticky notes to empty scales and label the opposing ends of the scales. The participants were also asked to name the categories they had created.

In the third exercise stylistic features from existing literature as well as the expert ideas were introduced on yellow sticky notes. The participants were asked to place the yellow sticky notes to their categories as they saw fit, make new categories if needed, and move aside any notes they found irrelevant for their podcast listening experience. Because the grouping exercises from both the second and third exercise lead to one set of categories which contained sticky notes from both the second and third exercises, we discuss the results of both of the exercises together in this section.

Summing up all the categories created in the three workshops, the participants created altogether 19 categories. These categories were formed by both the participant generated stylistic features from the first exercise and the stylistic features from existing literature as well as the expert ideas on potentially interesting stylistic features. Categories the participants of the first workshop created were "Use of Music or Sound", "Elicited Emotions in Listener", "Design of Podcast", "Engagement of Podcast", and "Ways of Speaking". Categories the participants of the second workshop created were "Non-dialogue (music/commercials)", "How Gripping/Entertaining a Podcast is", "Type of Podcast", and "Sound Quality". Categories the participants of the third workshop created were "Music (background and intro)", "Sound and Audio", "Speaking", "Ads", "Style", "Host", "Participants", "Intro", "Endings", and "Other Random Factors". The 19 participant created stylistic categories had some similarities and overlaps among themselves. These similarities and overlaps were used to group the 19 categories to seven final stylistic categories, which are discussed in more detail in Section 4.3.3.

Below each one of the 19 participant generated categories are described. It is marked into each category name in which workshop they were created. For example, "WS1: Ways of Speaking" is a category called "Ways of Speaking" which was created in the first workshop.

WS1: Use of Music or Sound

This category had the scale "bad use of music - good use of music" and was very similar to the category "Music (background and intro)" from the third workshop. It is worth noting that these categories were created independently in two different workshops. On the "bad use of music" end of the scale were sticky notes like "background noise is distracting"', "too aggressive opening music", and "using repetitive background music too long" whereas on the "good use of music" end of the scale were sticky notes like "music is used in a nice way - calming" and "background sound that

transmits a feeling underlying what is spoken about". From the third exercise's sticky notes the ones mentioning the background music, signature intro, and audio quality were placed in this category. "A lot of silence" was also placed in this category.

WS1: Elicited Emotions in Listener

This category had the scale "negative - positive". A lot of sticky notes were placed under this category, both sticky notes which were created during the exercise 1 and a lot of the third exercise's sticky notes. The participants wrote on the sticky notes which kind of emotions they thought the feature on the post-it would elicit in them. They wrote for example in the "negative" end of the scale: "starts with advertisement –> annoyed", "speech too slow –> annoyed", "sharp "s"'s and wet speaking sounds –> disgusted", "whiny voice, sounds like she starts crying soon –> sad, irritated" and so on. In the "positive" end of the scale they wrote: "happy voices –> positive mood, engaged", "very appealing voice –> engaged, calm", "playing excerpts/audio of recordings such as news or speeches –> calm, relaxed", "audience(listeners) response and strong interactions –> sense of belonging (to the audience)", and "energy makes it possible to feel like "being there in the conversation" –> entertained". Some of the third exercise's sticky notes placed into this category with their emotion tags were: "boring", "audio quality low –> irritated", "a lot of silence –> confused, annoyed", "ads present –> aggressive, annoyed", "monologue –> bored", "interruptions –> hard to follow", "interview –> engaged", "audio quality high –> engaged", "excitement", and "lively".

WS1: Design of Podcast

This category had the scale "bad - good". In the "bad" side of the scale were notes such as: "topic not clear", "starts with advertisement", "it just starts abruptly", "intro music too long", and "extremely long sentences or little breaks between". In the "good" end of the scale were notes such as: "starts with a nice set of bell sounds", "song in the beginning", "speaker tells the topic beforehand", "introduction of podcaster", and "several people having a conversation". This category seemed to mainly reflect the subjective ideas of the participants who created this category had on what is a good or a bad podcast.

WS1: Engagement of Podcast

This category had the scale "un-engaging - engaging". In the "un-engaging" side of the scale were notes like: "doesn't start with the topic straight away", "speech too slow", "background noise distracting", "intro music too long", "mumbling", "ads", "annoying accent", "monotonous way of speaking", and "talking over one another –> hard to understand individual people". In the "engaging" side of the scale were things like: "very appealing voice", "good use of music and audio", "several different voices", "introduction of podcaster", "speaker tells the topic beforehand", and "nice, deep and clear voice of speaker". The third exercise's sticky notes placed to this group were related to the audio quality, to speech rate, to the presence of ads, to mixed genders and different set of personalities and attitudes, as well as to keywords like "boring" versus "excitement" and "lively". In some ways, although named differently, this category bears similarities to the category "Elicited

Emotions in Listener" from the first workshop. In the emotion category the positive emotions were often about engagement and feeling of entertainment. The post-it mentioned on the scales of both categories were also similar.

WS1: Ways of Speaking

This category had the scale "annoying - enjoyable". It contained sticky notes related to the speaking style; "whiny voice", "speech too slow", "unclear voices", "strong accent", "mumbling", "very appealing voice", "nice, deep and clear voice of speaker", "female voices often sound more soft to the ear". A lot of the third exercise's sticky notes were added to this category. The third exercise's sticky notes were for example: "multiple people", "calm", "excitement", mentions on the gender or age, mentions about showing emotion, and about speech rate. All the sticky notes placed into this category can be seen in figure 4.5.

Figure 4.5: The category "Ways of Speaking" and the sticky notes placed to this category.

WS2: Non-dialogue (music/commercials)

This category had the scale "bad - good" and was very similar to the three other categories about music, sounds and other audio: "Music (background and intro)" and "Sound and Audio" from the third workshop, and "Use of Music or Sound" from the first workshop. This category contained notes such as: "music is a bit off. Does not fit well with the content.", "loud background music", and " laughter/audience being there" in the "bad" end of the scale and "music fits the message", "background music creates an atmosphere", and "well-mixed voice loudness levels" in the "good" end of the scale. Additionally, the category included mentions about commercials: "the ads are quite long", "commercials at the start of the podcast are irritating". The third exercise's sticky notes which were placed into this category were about ads, amount of silence, laughter, clapping, editing effects, use of music and other sounds.

WS2: How Gripping/Entertaining a Podcast is

This category is similar to the category "Engagement of Podcast" from the first workshop. It had the scale "not at all - can't stop listening to it". Things mentioned in the "not at all" end of the scale were about loud or disturbing background music, about slow or monotonous speech or otherwise annoying pronunciation or a background noise/buzz. The sticky notes in the "can't stop listening to it" end of the scale were about the flow of the podcast, about the discussion between host and quests, about natural conversations where people were joking around, about non-distracting background music, and about the podcast sounding like "a bunch of friends messing around and then talking about stuff". The third exercise's sticky notes placed to this category were for example: "boring", "one person", "monologue", "audio quality low", "interruptions", "calm", "speech rate fast", "storytelling", "multiple people", "social multi-party conversation", "audio quality high", "showing emotion", "excitement", and "discussion". All the sticky notes placed into this category can be seen in figure 4.6.

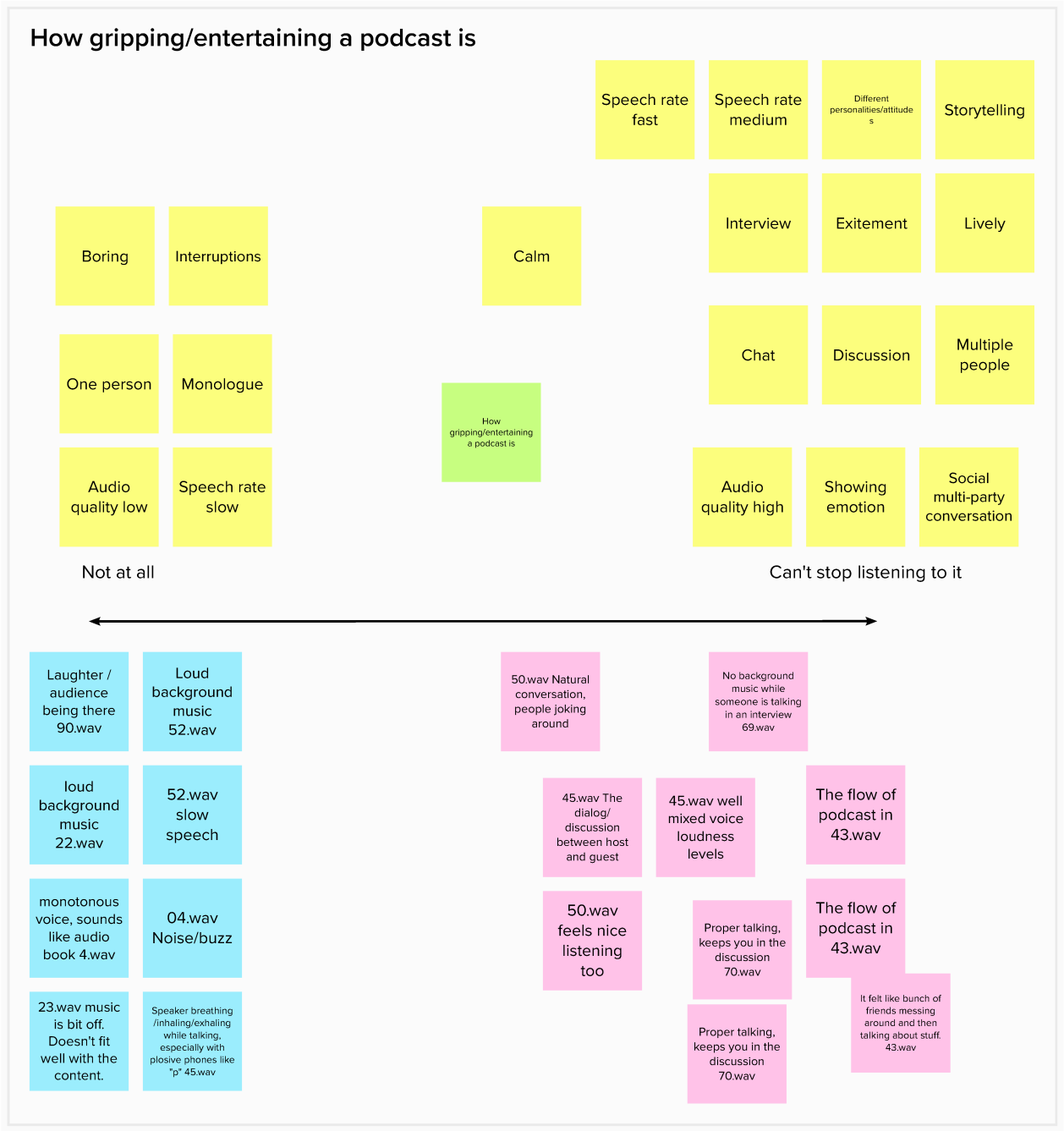

Figure 4.6: The category "How Gripping/Entertaining a Podcast is" and the sticky notes placed to this category.

WS2: Type of Podcast

This category had the scale "serious - leisure". Whereas this category and the scale are interesting, the sticky notes placed in this category seem to be at least partially arbitrary. The "serious" end of the scale contains notes such as "editing", "background music creates atmosphere", "music is a bit

off. Does not fit well with the content". It is not obvious from these notes why they are placed to this category and to this end of the scale. On the other hand in the "leisure" end of the scale notes such as "hosts seem to know each other and can joke around in an authentic way(indicated by laughter and jokes)", and "Dynamic between 2 people (seem like friends, relaxed atmosphere)" are easier to understand how they fit to this category. From the third exercise's sticky notes things like were placed to this category.

WS2: Sound Quality

This category had the scale "bad - good". This category had sticky notes such as "echo" and "noise in recording" in the "bad" side of the scale and things like "well mixed voice loudness levels" and "well recorded and good mix" in the "good" side of the scale. The third exercise's sticky notes related to audio quality were placed in this category.

WS3: Music (background and intro)

This category had the scale "suitable - unsuitable". Participants placed sticky notes like "intro music short enough" and "suitable music" to the "suitable" end of the scale. Things like "music is chaotic and all over the place" and "music in the beginning a bit too long" were placed on the "unsuitable" end of the scale. From the third exercise's sticky notes the ones with background music and the one about signature intro/opening jingle were placed to this category.

WS3: Sound and Audio

This category had the scale "suitable - unsuitable". This category consisted of sticky notes mentioning the audio quality. It also included sound effects and some other sound related things like laughter and clapping.

WS3: Speaking

This category had the scale "pleasant - unpleasant". The category bears similarities to the category "Ways of Speaking" from the first workshop. In the "Speaking" category things like "host is talking clearly", "hosts are pronouncing well", and "hosts have soothing, nice voices" are in the "pleasant" side of the scale. In the "unpleasant" side of the scale are things like "speaker talks in a lazy/sleepy way", "talk with unclear pronunciation", and "speak too unstructured". From the third exercise's sticky notes the ones about speech rate were added to this category. Swearing was also included to this category, to the "unpleasant" end of the scale.

WS3: Ads

This category had the scale "suitable - unsuitable". This category contained only notes about the presence of ads.

WS3: Style

This category had the scale "pleasant - unpleasant". This category was created in the third workshop to exclusively describe the third exercise's sticky notes. The third exercise's sticky notes placed into this category were for example: "light mood", "serious", "lively", "interview", "monologue", "discussion", "informative", "calm", "excitement", and "social multi-party conversation". The sticky notes placed under this category seem partially arbitrary.

WS3: Host

This category had the scale "pleasant - unpleasant" and contained only one post-it which said "liked the excitement of the presenter. She seemed to be engaged with the topic and making the interviewees involved".

WS3: Participants

This category had the scale "suitable - unsuitable" and contained only two sticky notes which were both from the third exercise. These sticky notes were "different personalities/attitudes" in the "suitable" part of the scale, and a note about both female and male speakers in the unsuitable part of the scale. It is unclear why this sticky note was placed there.

WS3: Intro

This category had the scale "suitable - unsuitable". The category contained notes such as: "spontaneous, different intro", "really liked the intro:introduced the show, the topic and a question to the listeners!!", "good intros and examples on topics", "advertising at beginning", "starts a bit slow, don't understand which topic it is right away", "too long intro, didn't like the music intro". No sticky notes from the third exercise were placed in this category.

WS3: Endings

This category had the scale "suitable - unsuitable" and contained only one post-it: "good and sweet end". The lack of sticky notes might be a symptom from the fact that participants mostly listened to the beginning and to the middle of the podcast episodes in the exercise 1, and often left the end of the podcast without attention.

WS3: Other Random Factors

This category had the scale "suitable - unsuitable". Miscellaneous notes to do with audio quality, different people having an interview, comments about the voices, or the length of the podcast were placed in this category. This category seemed to function as a leftover "bucket" for the notes the participants did not manage to place elsewhere but still found important. The notes in this category seemed to have nothing in common with each other.

WS1-3: Redundant Features

The third exercise proposed stylistic features which we had picked from existing literature to the participants, and also included some of the expert ideas on stylistic features. The participants were asked to set aside the proposed stylistic features which they did not find relevant for their listening experience. When comparing the workshop results is became apparent that each one of the disregarded proposed stylistic features were used in at least one category across all the workshops. Therefore, none of them can be seen as completely irrelevant. Nevertheless, the disregarded proposed stylistic features mainly contained stylistic features related to extralinguistic factors, such as the speaker's age or gender. The second and the third workshop both resulted in disregarded the stylistic features "Age", "Male only", "Female only" and "Mixed genders". In addition to extralinguistic factors, a few other types of stylistic features were set aside. The second workshop's participants set aside "Other vocalisations" and "Swearing". The third workshop's participants set aside "Multiple people", "One person" and "Showing emotion".

Participants of the first workshop set aside "Clapping", "Laughter" and "Social multi-party conversation". It is of note that the first workshop ran out of time and had to rush to complete the third exercise. This may have lead to the participants disregarding the sticky notes "clapping" and "laughter". Furthermore, for example "laughter" was observed and written down by the participants themselves during the first exercise so we have a reason to believe that it was an important feature for some of the participants even though it was put aside in the third exercise. The features the participants considered redundant for their podcast listening experience and thus moved aside can be seen in Figure 4.7.

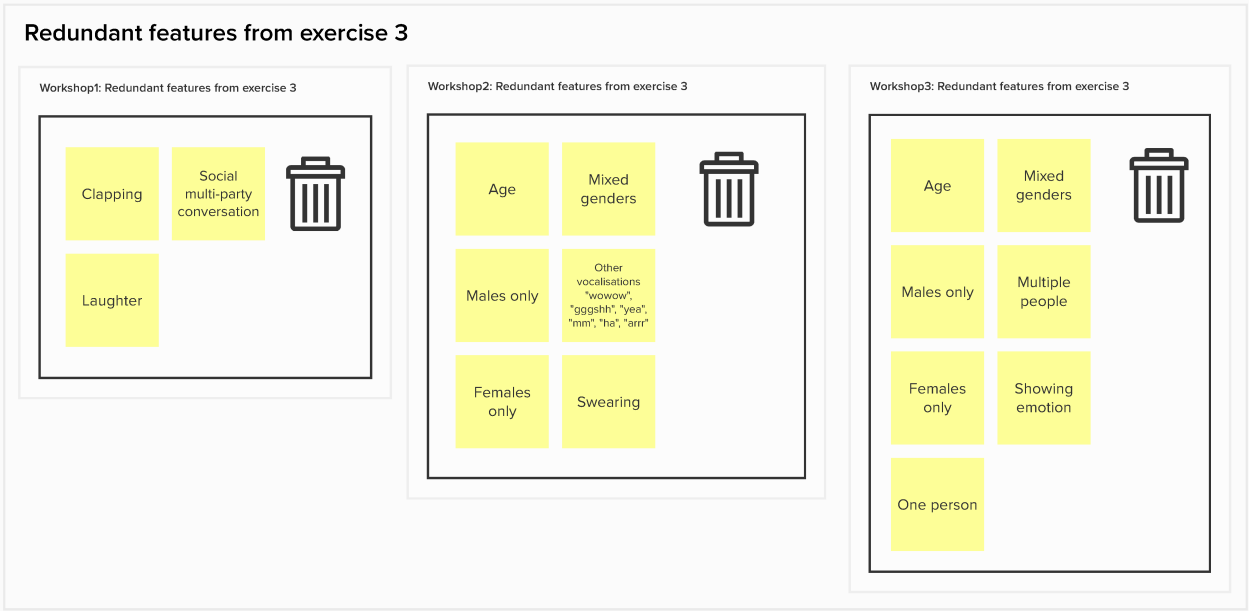

Figure 4.7: Redundant features from the third exercise for all the three workshops where the results are displayed from the first workshop to the third workshop (left to right).

4.3.3 Summarised Stylistic Categories and Framework of Stylistic Podcast Characteristics

In order to further summarise the results of the second and third exercise we removed overlap from the participant generated categories and grouped the categories based on their similarities both in terms of the category name, the category scale, and the sticky notes under the category. This created seven final stylistic categories, one of the main contributions of the study. The seven categories are: "Music and Sounds", "Ads/Commercials", "Audio Quality", "Speaking Style and Voice Qualities", "Formality Level", "Engagement Level", and "Format". The participant generated category "other random factors" was left out from the seven final categories because of its miscellaneous content. The seven final stylistic categories, their scales, the participant generated categories they consist of, and examples of participant generated sticky notes indicating the stylistic features which make up the stylistic categories can be seen in Table 4.4.

To demonstrate how the findings fit to our framework of stylistic podcast characteristics, an example based on the results is given in figure 4.8. In the example, we have taken one of the seven final stylistic categories "engagement level", placed stylistic features from that category to the middle level, and inserted examples of acoustic features for detecting these mid level features to the lowest level of the framework.

Table 4.4: Seven final stylistic categories.

CATEGORY NAME	**SCALE**	**INCLUDES PARTICIPANT GENERATED CATEGORIES**	**EXAMPLES OF WORKSHOP STICKY NOTES INDICATING STYLISTIC FEATURES**
Music and Sounds	pleasant - unpleasant	"Music (background and intro)" "Use of Music or Sound" "Sound and Audio" "Non-dialogue (music/commercials)"	"background music creates an atmosphere" "music is chaotic and all over the place" "using repetitive background music too long"' "no sound effects/music"
Ads/Commercials	suitable - unsuitable	"Ads" "Non-dialogue (music/commercials)"	"ads present" "in the beginning, cannot be sure if it's an ad or the program itself"' "commercials at the start of the podcast are irritating"
Audio Quality	good - bad	"Sound Quality"	"audio quality" "echo" "noise in recording" "well recorded and good mix"
Speaking Style and Voice Qualities	pleasant - unpleasant	"Ways of Speaking" "Speaking" "Host"	"nice, deep and clear voice of speaker" "whiny voice" "speech too slow" "mumbling" "speaker talks in a lazy/sleepy way" "excitement of the presenter" "host seemed to be engaged with the topic making the interviewees involved."
Formality Level	serious - leisure	"Style" "Type of Podcast"	"light mood" "dynamic between 2 people (seem like friends, relaxed atmosphere)" "conversational style" "serious" "interview" "explainer/lecturing" "news reading/reporting style"
Engagement Level	engaging - not engaging	"Engagement of Podcast" "How Gripping/Entertaining a Podcast is" "Elicited Emotions in Listener"	"monotonous way of speaking" "talking over one another –>hard to understand individual people" "excitement" "showing emotion" "speech too slow –>annoyed" "happy voices –>positive mood, engaged" "energy makes it possible to feel like "being there in the conversation" –>entertained"
Format	suitable - unsuitable	"Design of Podcast" "Participants" "Intro" "Endings"	"it just starts abruptly" "intro" "topic not clear" "introduction of podcaster" "different personalities/attitudes" "good and sweet end"

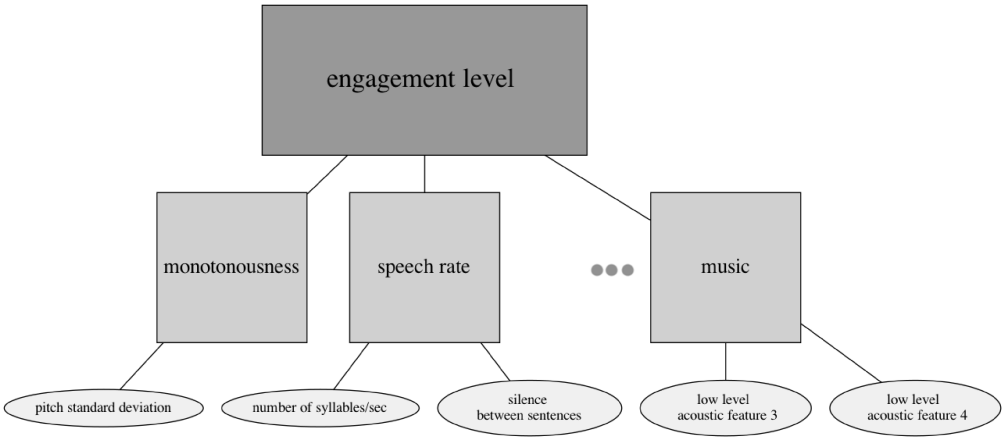

Figure 4.8: Partially filled framework of stylistic characteristics for category "Engagement Level".

4.4 Conclusions and Discussion of Chapter 4

In this section the conclusions and discussion of chapter 4 are presented. In addition, some limitations of the first phase of the study are discussed.

4.4.1 Conclusions

In this chapter we have presented our work on answering to the research question RQ1: "What stylistic characteristics of podcasts do listeners find interesting or important for their podcast listening experience?" We have broken down this research question to three sub-questions, namely RQ1.1: What stylistic features can listeners observe and verbalise in podcasts? Both in terms of what features they like and do not like in podcasts. RQ1.2: What are listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about? RQ1.3: How well do stylistic features identified from literature and by experts match listeners' perceptions on stylistic features and categories? Three qualitative participatory design workshops were carried out in small groups in order to collect data to answer these questions. We summarise the answers to the sub-questions in the following sections.

RQ1.1: What stylistic features can listeners observe and verbalise in podcasts? Both in terms of what features they like and do not like in podcasts.

The results of the research show that podcast listeners care substantially about different kinds of stylistic features in podcasts. Listeners can observe and verbalise stylistic features in podcasts which generally have to do with speech, voice, or talking style in the podcasts, presence and type of music, the topic or content related features such as topic clarity, the audio quality of podcasts, introduction related features, and presence of ads. Additionally, listeners can observe stylistic features related to the dynamics of the participants in the podcast and the flow of the podcast as well as the presence of

laughter or swearing. Podcast listeners also care about stylistic features such as whether the podcast contains conversations or interviews, or whether the podcast has a lot of silence or other sounds.

The participants of the study paid most attention to stylistic features related to speech, voice, and talking style. Further inspection of these features revealed that the following types of stylistic features were interesting to the participants. The participants cared about speech rate, tone of voice, clear speaking, pitch related features like monotonousness of speech, extralinguistic factors such as gender, articulation, and other features such as affect in the voice of the speaker.

When it comes to participants' preference for the aforementioned stylistic features, all participants had their own individual preferences. In general, almost every kind of stylistic feature varied nearly equally between being preferable or disliked with a few exceptions, e.g. ads being consistently disliked and stylistic features related to dynamics, flow, and people in the podcasts were always preferable.

RQ1.2: What are listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about?

Listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about, according to our results, are the following. Categories the participants of the first workshop created were "Use of Music or Sound", "Elicited Emotions in Listener", "Design of Podcast", "Engagement of Podcast", and "Ways of Speaking". Categories the participants of the second workshop identified were "Non-dialogue (music/commercials)", "How Gripping/Entertaining a Podcast is", "Type of Podcast", and "Sound Quality". Categories the participants of the third workshop came up with were "Music (background and intro)", "Sound and Audio", "Speaking", "Ads", "Style", "Host", "Participants", "Intro", "Endings", and "Other Random Factors". Altogether, the participants of our study created 19 categories where some of the categories were similar to each other in terms of the name of the category, the dimensions of the category indicated by the category scale, and the stylistic features which were grouped under the category. By using these similarities, similar categories were merged together. This formed final seven stylistic categories, one of the main contributions of our study. The categories are "Music and Sound", "Ads/Commercials", "Audio Quality", "Speaking Style and Voice Qualities", "Formality Level", "Engagement Level", and "Format".

RQ1.3: How well do stylistic features identified from literature and by experts match listeners' perceptions on stylistic features and categories?

The stylistic features from existing literature and the expert ideas on stylistic features fit well to the participants' perceptions on stylistic podcast categories. All of the proposed stylistic features were used as part of the participant created stylistic categories at least once in at least one of the workshops. Moreover, most of the proposed stylistic features were used in the stylistic categories in every workshop. This demonstrated not only that the proposed features from the existing literature and the expert ideas on stylistic features fit to the participants' perceptions on stylistic podcast categories, but that the proposed features were important to the participants' listening experience.

Despite the fact that all the proposed stylistic features were placed to a participant created category at least once, some proposed stylistic features were seen as less relevant for the listening experience than others. Extralinguistic factors, namely age and gender, seemed to be less relevant for our participants.

4.4.2 Discussion

The results of this research phase provide an insight into listeners' perceptions on possible stylistic characteristics of podcasts in a general context meant to serve not only popularity prediction, but modelling style across any kinds of podcasts. A way of organising stylistic characteristics was proposed in order to approach them in a structured way. The identified stylistic characteristics were divided into stylistic categories which comprise of more detailed stylistic features. Based on the fact that the participants of the study had no trouble placing the observed stylistic features on empty category scales and naming the scales, it seems plausible that the stylistic categories should not be treated as only collections of stylistic features, but that they possess a dimensional component. For example, a dimension for describing the category "Engagement Level" is "not engaging - engaging". Whereas, most of the stylistic categories' dimensions reflected subjective ideas on what is, for example, a pleasant or unpleasant use of music in the podcast, the categories "Formality Level" and "Engagement Level" could be described with presumably more objective scales of "serious - leisure" and "not engaging - engaging", respectively.

Both the stylistic categories and the stylistic features identified in the study can be used as starting points for understanding and modelling the stylistic characteristics of podcasts. For our study, the workshops yielded stylistic features which will be further experimented with in the next research phase. In general, the results can serve as direction indicators for further studies which aim to dig deeper into any of the identified stylistic characteristics from the user's perspective, or which might want to address automatic detection of any of the stylistic categories or features. For example, it might be worth while researching what kind of consumer personas could be build around the stylistic categories or features. Such research could be to investigate if patterns exist among users in terms of the different stylistic categories or features. For example, what kind and how much of music and sounds should be included in podcasts for a consumer group to find it pleasant. In terms of ads, one could study the optimal amount and placement of ads in podcasts in order to optimise the ad revenue without annoying listeners away from the product.

The results could also be used as a tool to reflect on podcasts' stylistic dimensions both during creation, in market positioning, and in distribution. For example, podcast creators could use the seven identified stylistic categories and their scales when designing their podcasts. These seven categories could be used to either identify a stylistic niche within some content based podcast genre, or to design podcasts which conform with the given genre's style. Such approach could aid in both the design of the podcast and in the market positioning of the podcast. Similarly, viewing a streaming platform's podcast catalogue from the perspective of the seven stylistic categories could help to identify stylistic gaps in the podcast offering.

The results of this study have the potential to impact both podcast search and recommendation, given that the podcasts can be tagged with the categories and features in a feasible way. We hypoth-

esise that including the identified stylistic features to podcast representations and enabling users to search for content based on style can create new, improved search experiences. Similarly, podcasts could be recommended not only based on the topic and content of what is said, but also based on their stylistic features. Including the stylistic features identified in this study to podcast recommendation algorithms would take into consideration the paralinguistic and musical dimensions of podcasts, types of information widely ignored in spoken content retrieval until today. All in all, we believe that utilising stylistic characteristics of podcasts could lead to improvements in the quality of recommendations and returned search results, thus improving used experiences when browsing for relevant podcasts.

4.4.3 Limitations

In this section we discuss the limitation of the work on users' perceptions on what kind of stylistic characteristics of podcasts are important for their listening experience. The limitations of this phase of the research are related to the user study set up.

On the selection of the workshop participants, there was a slightly unequal gender distribution of seven females and two males. The results of the user study showed that one of the aspects related to the likability of the podcasts was softness of voice. Softness of voice can most likely be perceived easier in a female voice than in a male voice. Typically when there is a feature under study that is fully connected to the characteristics of the participants, it is recommended to make sure the participants are fairly represented as it could affect the findings. Since it was not anticipated that gender dependent stylistic features to surface in the user study, the participants' gender were not controlled for. However, now being aware of some of the user study results, if we would repeat the study, we would aim for an equal gender distribution in our participants.

Another limitation of our user study lies in the set of podcasts we selected for the participants to listen to during the workshops. Even though we made an effort to collect the set of podcasts in such a way that they would represent a wide range of different stylistic characteristics, we might have missed some entirely different podcast episodes from the ones we selected. This means that another kind of a podcast set could yield different user study results.

A third point of discussion is what kind of results would have surfaced, especially in terms of the most mentioned stylistic features and the stylistic categories, if we would have carried out the user study workshops individually instead of in small groups. It is left to possible further studies to examine if the most mentioned features would still be related to music or speech, voice, or talking style and what kind of stylistic categories would emerge from such set up.

Chapter 5

Research Question 2: Automatic Extraction of Stylistic Features

Chapter 4 reported on a user study about listeners' perceptions of stylistic podcast dimensions. The main motivation for the study was to gain insights into users' perceptions on what kind of stylistic features they care about, and how these stylistic features form higher level stylistic categories. The purpose of these results was to inform the development of stylistic podcast modelling technologies for search and recommendation. As a step towards this, we conducted experiments to assess how well suited, from a computational perspective, the stylistic features, which listeners care about, are for automatic podcast recommendation and search functionalities. In particular, we examined the stylistic features which can be extracted from the podcast audio signal.

The chapter is organised as follows. Section 5.1 states the objectives of this part of the research. The reasoning for our focus, the data, the feature extraction, and the data analysis plan are presented in the methodology section 5.2. Section 5.3 presents the results of our experimentation. Finally, section 5.4 provides a summary of answers to the addressed research question.

5.1 Objectives

The goal of the second phase of the research was to answer the research question **RQ2: "How suitable are the listener perceived stylistic characteristics of podcasts for automatic podcast recommendation and search based on the attainable information from podcast audio signal?"** A bottom-up strategy was applied in order to approach the stylistic characteristics of podcasts, by starting from the simpler sub-elements of what constitutes a stylistic characteristic. As such, we limited our experimentation on stylistic characteristics to the middle level stylistic features and their low level acoustic features as described in our framework. We left the stylistic categories out of the scope of our experimentation, thus proposing their assessment in future studies.

We approached answering RQ2 step-wise by first reflecting on what stylistic features were a sensible starting point for modelling stylistic podcast characteristics and narrowing down to these stylistic features for our experimentation. Secondly, we ask whether the selected stylistic features

can be used to discriminate between podcasts and if so, which ones of the selected stylistic features can be used to do so.

5.2 Methodology

The experiments were carried out in the following steps. First, a set of stylistic features were selected, followed by the development of a program which automatically extracted selected stylistic features from podcast audio. Finally, the results of the program were plotted, and a statistical analysis was conducted on them. A more detailed overview of the methodology is given below.

5.2.1 Methodology Overview

A set of stylistic features was selected by first defining which features were a sensible starting point for modelling stylistic podcast characteristics. This narrowing down was made due to the scope of the project. Implementing extraction of the complete set of stylistic features, which emerged from the user study, was left for future studies. We reflected on the sensibility from four perspectives: what listeners care about according to our the study results, what is generally sensible from the perspective of podcast production and consumption, how does the feature align with our research focus, and how easily can the feature be automatically detected from podcast audio by using available open source tools. More on the selection of stylistic features in section 5.2.2.

After selecting sensible stylistic features to start with, the focus was to answer the question: "Can stylistic features be used to discriminate between podcasts and if so, what are the stylistic features which can be used to do so?" Discriminative power worked as an evaluation criterion for a feature's suitability for search and recommendation. If a feature could not discriminate between podcasts it would not be useful, since the features should be able to be used to differentiate between different podcasts.

In order to answer the aforementioned questions a program to automatically extract selected stylistic features from podcast audio was developed and the resulting feature value distributions were examined. Podcast episodes were chosen as the unit for extracting stylistic features. In practise, this meant computing a single value for each stylistic feature for each episode. This choice was motivated by episodes being a natural unit of podcast access and consumption. Therefore, the interest was in comparing podcasts with each other, starting from the episode level. The developed program is described in detail in 5.2.4 Experimental Setup .

After computing values of the selected stylistic features for each episode in the dataset, the distributions of these feature values were examined. This was done by first conducting an exploratory analysis where the feature value distributions was plotted in multiple ways. Then, statistical tests were carried out in order to confirm hypotheses regarding the feature value distributions. The hypotheses were set based on emerged information from the exploratory analysis. Details of the data analysis are found in section 5.2.5.

5.2.2 Selecting Stylistic Features

In this section, selecting stylistic features for experimentation is discussed. We go through stylistic features which emerged in the user study and reflect on them from four perspectives: what listeners care about according to the user study results, what is generally sensible from the perspective of podcast production and consumption, how does the feature align with our research focus, and how easily can the stylistic feature be automatically detected from podcast audio by using available open source tools.

We start by discussing the stylistic features we chose for the experimentation, namely, “pitch”, “monotonousness”, “speech rate”, proportion of “music”, “noise”, “silence”, “female speech”, “male speech” and “speech” in general. After this we discuss the stylistic features we decided to disregard from our experimentation, namely, “topic/content”, “introduction of the podcast”, “ads”, “swearing”, “audio quality”, “dynamics/flow/people", “editing/format”, “interviews”, “conversations”, “clear speaking”, “tone of voice” and “laughter”.

Chosen Features

The decision was made to concentrate mainly on “pitch”, “monotonousness”, “speech rate”, and proportion of “music” for the implementation. Pitch, monotonousness, and speech rate were selected, since the most interesting features to the workshop participants were related to speech, voice, and talking style. Additionally, pitch and speech rate are prosodic features, which contribute to many different paralinguistic aspects of speech. For example, pitch and speech rate can be connected to emotions, speaker’s personality, and speaker’s state of mind and are thus interesting basic features to describe speaker’s speaking style. Pitch standard deviation is a direct indicator of monotonousness in speech [81]. The lower the pitch standard deviation is, the more monotonous the speech is. Speech rate is indicating if the speech is slow, fast or stays within the limits of average speech rate. Both of these features, monotonousness and speech rate, were repeatedly mentioned in the workshops, therefore including them in the second study phase was sensible from the user research perspective. In addition to user’s perspective, the availability of promising open source tools for computing pitch [82] and speech rate [83] values from audio signal was ascertained.

We selected proportion of “music” to be one of the stylistic features for the experimentation. This stylistic feature was selected because music related features were the second most mentioned in our the study. From this we can draw the conclusion that listeners pay attention to the presence of music. Research has already been done on music detection and analysis [84], [85], [86], [87] and open source tools used to segment music from the other parts of audio [88] were available.

In addition to the above stylistic features, “silence”, “other sounds” like noise, and the perceived sex of the speaker was also chosen for detection. With any speech processing task, the common first step is to separate speech from other sounds and silence. Analysing speech related features from audio full of non-speech sounds can add significant amount of noise or bias to the results. Therefore, it was important to separate the non-speech sounds from speech for the analysis. It was also important to separate “female speech” and “male speech” from the speech in general so that the variations in pitch could be calculated without the change in speakers of one gender to another

influencing the variation by increasing it. Hence, these features became part the set of stylistic features for the experimentation as well.

Disregarded Features

The decision was made to disregard the stylistic features related to “topic/content”, “introduction of the podcast”, “ads”, and “swearing”. This decision was made based on that to the best of our knowledge detecting these features requires transcripts from an automatic speech recognition system. Therefore, these features are outside the scope of this study.

Detecting “audio quality” was not relevant for this study because the assumption was that the better the audio quality is, the more listeners will enjoy the podcast. In our opinion podcasts should not be searched or recommended based on audio quality, since this would most likely favour the professional podcasts over amateur ones and decrease the likelihood of the amateur podcasts being found by the user. Furthermore, whereas audio quality is a property that depends on the equipment, the focus of this study was in modelling the stylistic features that are produced by humans and those which characterise the speaker. Therefore, it would not add any value to this study to detect audio quality of podcasts.

Next, we discuss the stylistic features we think might be harder to operationalise. We start by discussing the stylistic features grouped into “Dynamics/Flow/People". Things like “chill atmosphere”, “dialog flows well”, and “interactive” are very undefined concepts which might be subject to huge subjective variations. For example, what does it mean to people that a “dialog flows well”, what in the podcast indicates a good flow of dialogue? The decision was made to exclude these higher-level concepts from this study and leave them for future studies.

Participants wrote several notes related to the engagement of the speakers. They wrote notes such as: “ liked the excitement of the presenter. She seemed to be engaged with the topic and making the interviewees involved”, and “voices sounded interested in the opponent speaker”. Some research has been done on engagement detection. For example, detecting "Hot Spots", involvement in multi-party conversations, and other highlights in audio media, have been studied by [64], [65], and [66]. Prosodic features such as pitch and F0 were found to be useful for detecting this kind of involvement from audio. Despite of existing studies, detecting engagement does not yet seem to be a trivial task in computer science. No open source, out of the box tools to detect engagement were available for this study. Training an engagement detector from zero would require good, quality human labelled training data and training a machine learning model on the data. Additionally, the ways of showing engagement in a conversation might vary across cultures. This would mean that a detector trained on data from some culture could possibly perform poorly in the context of another culture. However, detecting engagement could prove to be an interesting task for enriching stylistic podcast representation, but with the above considerations it is left for a future study.

Next we look into the group “Editing/Format”. This group contained notes such as “the loudness of all audio components in the podcast are matched so they compliment each other” or just “editing”. The spread of these notes make it unclear how to implement a detector which detects “good editing”. Because the ambiguity of the stylistic features in this group, detecting the stylistic feature related to this group is left for future studies.

All the comments from the user research which mentioned "interviews" or "conversations" were positive. This makes detecting if a podcast is an interview or just a casual chat interesting to this study. However, it might be difficult to operationalise what defines an interview and what sets it apart from a conversation. We speculated characteristics like amount of speaker overlaps, frequency in the change of speaker turn, and amount of questions asked by one of the speakers, marked by intonation rise in the end of the sentence. However, this seemed to require further investigation. Therefore, these features were disregarded from the experimentation.

Three other stylistic features, which were left out of the scope of the experiments were "clear speaking", "tone of voice" and "laughter". "Clear speaking" and "tone of voice" were mentioned relatively frequently in the user study. We speculated that clear speaking could be marked by how distinctly phones are pronounced and how close they are to the corresponding phoneme. Applying this logic one could model the distance between vowels (area of the vowel triangle) in speech and compare different speakers' vowel difference to each other to find out who speaks clearer. Additionally, things like speech rate, length of pauses between sentences, and use of intonation and volume of voice can attribute to what is perceived as clear speech. What the participants of the user study probably meant by "tone of voice" is perhaps related to prosody and voice quality. Voice quality has been studied in multiple occasions in speech processing research field. For example, a study from 2004 reviewed the experimental derivation of voice quality markers. The study suggests that voice quality is best investigated as a multi-dimensional parameter space involving a combination of factors involving individual prosody, temporally structured speech characteristics, spectral divergence and voice source features [89]. Due to the time constrains of this research project, these features were left to further studies.

Laughter in audio and visual media has been widely studied [90], [91], [92], [93], [94], [95] and some laughter detection projects are available online [1] [2]. Laughter can for example be detected from the high intensity peak patterns of spectrograms or by using MFCCs and pitch related features. However, laughter detectors can be sensitive to performing badly in contexts which have not been represent in their training data. For example, a laughter detector trained with purely speech and laugh might perform badly in the presence of noise and background sounds, detecting for example clapping and other sudden high intensity sounds as laughter. Based on the above and the fact that laughter got relatively little mentions in our the study, the decision was made to leave it out of the experiments.

5.2.3 Podcast Dataset

In order to be able to detect selected stylistic features from podcast audio and in order to examine the value distributions of these features, we needed to construct a dataset with not only the podcast audio data, but also metadata such as information on the podcast episode, the podcast show, and the genre the podcast belongs to.

"The Spotify Podcasts Dataset"[96] of about 100 000 podcasts was used in this study. The

[1] https://github.com/ideo/LaughDetection

[2] https://github.com/jrgillick/laughter-detection

dataset was chosen because of its diversity in topics, because of its size, and because it was freely available for use in this study. Next, the main characteristics of this dataset will be described. "The Spotify Podcasts Dataset" contains podcast episodes and their metadata from wide variety of different types of podcasts from professional production to amateur podcasts. The dataset covers a wide range of topics, including lifestyle and culture, storytelling, sports and recreation, news, health, documentary, and commentary. In addition, the content is delivered in a variety of structural formats, number of speakers, and levels of formality, whether scripted or improvised, or in the form of narrative, conversation, or debate. "The Spotify Podcasts Dataset" mainly contains episodes with duration less than 90minutes and has primarily English content. The episodes were sampled randomly from episodes which were published between January 1, 2019 and March 1, 2020, according to the above criteria. The 100 000 episode collection consists of 50,000 hours of audio and accompanying transcripts and metadata. The metadata includes the following information: show URI (Uniform Resource Identifier), episode URI, show name, episode name, show description, episode description, publisher, language, RSS link, and duration. Show URIs and episode URIs are strings of characters and numbers which uniquely identify the corresponding shows and episodes.

The data for this study was sampled from "The Spotify Podcasts Dataset" by first joining "The Spotify Podcasts Dataset" with other metadata tables from Spotify. These tables contained information on higher level categories/genres the podcast episodes belong to. A test set of 911 episodes was sampled from all of the episodes by taking all episodes which fulfilled the criteria of episode duration being between 7 and 67 minutes, and that there were at least 30 episodes available for each show. The duration was limited to 7-67 minutes because the average duration of all the episodes was 37 minutes. Very short episodes were often just trailers or ads to advertise other content and very long episodes lead to much longer computation time of the feature extraction program. Hence, excluding these made sense for the purpose of this study. The requirement of at least 30 episodes per show was set in order to be able to compare different shows' feature values to each other. The intention was to avoid situations where a show would have only a few feature values representing them. In such cases, no strong conclusions could be drawn. Apart from controlling for the duration or for the amount of episodes per show, the selection of the episodes, for example in terms of content or topic, was not controlled for.

5.2.4 Feature Extraction

In this section we describe the program which was developed for automatic extraction of the selected stylistic features "pitch", "monotonousness", "speech rate", and proportion of "music", "noise", "silence", "female speech", "male speech", and "speech". The feature values were calculated for each podcast episode. An episode was selected as the precision level for inspecting the selected stylistic features because an episode is a natural unit for podcast consumption, therefore also being a natural unit for search and recommendation. Thus, the interest lies in comparing feature values with each other starting from the episode level. Limitations of this approach are discussed in section 5.4.2.

The decision was made to use an episode as the basic level in which the stylistic features would be explored, meaning that each episode would get assigned numerical values describing the given

features. Episode level inspection was decided on because episodes are the natural unit of consumption, the listener listens to an episode at a time and might listen to episodes across many shows and genres. An episode is also a natural unit for recommending relevant content. Next, the program overview is described, and the parts of the program which are responsible for extracting the selected stylistic features are presented in more detail.

Program Overview

The program was developed in Python programming language. Each episode_uri (unique episode identifier) was read from the metadata .csv file and used to download the corresponding audio file. Each episode was processed separately and parallel to each other in order to to enable utilization of multiple cores and thus decreasing the total computation time. First, the audio file was converted to .wav format and downsampled to 16kHz in order to have the right format for various Python speech processing tools. Next, the audio file was segmented to music, noise, silence, female speech, and male speech segments by using an open source Python tool inaSpeechSegmenter [88]. The percentage of each type of segments was calculated over the whole podcast episode and stored in a .csv file.

The segmenter returned timestamps of the different audio segments (music, start_time, end_time), therefore for further analysis of the speech the audio was sliced around these timestamps and the speech segments were temporarily saved to .wav files. Subsequently, the speech rate was calculated from the speech segments by using a Praat script [83]. The average speech rate over the whole episode was saved to a .csv file.

After computing the speech rate, the female and the male speech segments were fed to the part of the program which calculates the average and standard deviation of pitch. These were calculated separately from female speech and male speech to avoid biased high standard deviations caused by the potential speaker changes from a lower male speech to higher female speech or vise versa. The results were saved to the same .csv file as the results of the segmenter and the speech rate calculations. This way the program computed stylistic feature vectors for each of the episodes, where percentage of music, speech rate, pitch average, pitch standard deviation, and others were the dimensions of the vector. After all the above steps had been completed the temporary audio files and the cache of the program were deleted and the operating system process which took care of the current episode started over with a new episode. A rough overview of the program can be seen in Figure 5.1.

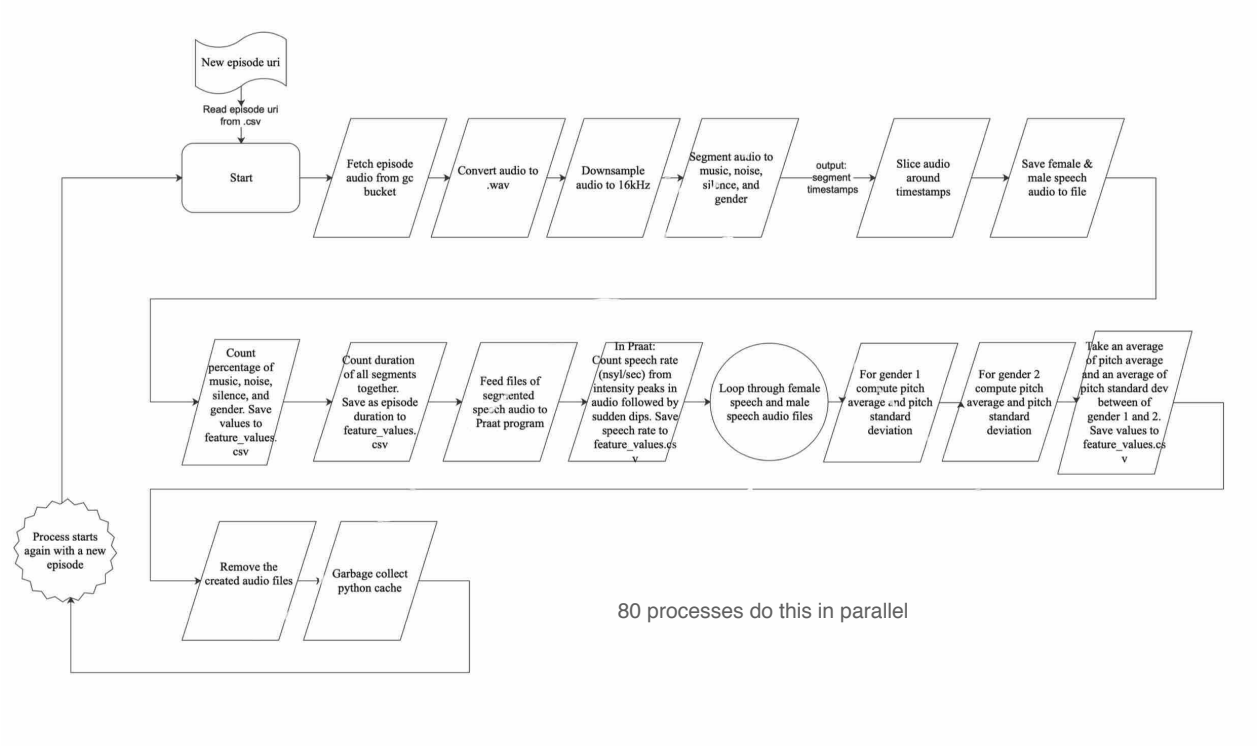

Figure 5.1: Rough overview of the program flow.

In the following sections the tools to detect the selected stylistic features are presented in more detail.

Speaker Segmentation

We used convolutional neural network based audio segmentation toolkit "inaSpeechSegmenter" [88] to segment the podcast episode audio. The segmenter splits audio signals into homogeneous zones of speech, music, noise and silence. Speech zones are split into segments tagged using speaker gender (male or female). Zones corresponding to speech over music or speech over noise are tagged as speech. The segmenter returns timestamps of audio together with the corresponding sound label. The segmenter was originally designed for large scale gender equality studies based on speech time per gender.

The segmenter was used to return time stamps around music, noise, silence, female speech, and male speech segments. These time stamps were used to split the podcast episode's audio to smaller segments tagged with music, noise, silence, female speech, and male speech.

Speech Rate

In order to calculate the speech rate from audio, the Praat program developed at Amsterdam university was used. Praat is a software developed specifically for analysis of speech. The developed speech rate program automatically detects syllable nuclei in order to measure speech rate without the need of transcription.

Peaks in intensity (measured in dB) that are preceded and followed by dips in intensity are considered as potential syllable nuclei. The script first considers all peaks above a certain threshold (determined with respect to the median intensity of the speech file) as possible syllable nuclei. However, this results in many peaks within each syllable. Therefore, the script then discards peaks that are not followed and preceded by dips of at least 2 dB in intensity. To discard loud fricatives as syllable nuclei, the script finally additionally discards all peaks that are not voiced. The number of syllable nuclei is divided by the total time of the given speech segment to give the speech rate of the segment. [83]

Speech rate was calculated only from speech audio segments. The speech rate of all the small speech audio segments which resulted from the output of the segmentation algorithm (inaSpeechSegmenter and splitting the audio according to the inaSpeechSegmenter outputted timestamps) were averaged to form one value to describe the whole podcast episode. It should be noted that this method presents only a rough summary of the speech rate across the episode and does not capture the potential variations within an episode. Nevertheless, using the average speech rate can be used to summarise this feature for the episodes and as such is useful for comparing speech rate between episodes.

Pitch

We calculated the average pitch and the monotonousness of speech for each of the podcast episodes in our dataset. The Python library "pYAAPT" from "amfm_decompy" was used for extracting the needed pitch information from the podcast episode audio signal. The original version of "YAAPT" was published in [82]. First, we mapped monotonousness onto pitch standard deviation because by definition monotonousness is described as "uttered or sounded in one unvarying tone : marked by a sameness of pitch and intensity" [81]. Accordingly, the smaller the pitch values' variability (standard deviation) is, the more monotonous the speech is. Below, instead of talking about monotonousness, we talk about calculating the pitch standard deviation for the podcast episodes. We also describe how the average pitch across a whole podcast episode was calculated.

Pitch was extracted from small speech audio segments, the output of our segmentation algorithm (inaSpeechSegmenter and splitting audio around time stamps from the output of inaSpeechSegmenter). Pitch was extracted separately from female speech and male speech. Pitch was extracted by inputting the audio signal of these female and male speech segments to the method pYAAPT.yaapt(signal). For simplicity, we used the default settings for the additional options of this method. For example, the length of each analysis frame was kept in the default of 35 milliseconds. The method pYAAPT.yaapt(signal) returned a pitch object from which the pitch values were separately saved. The pitch values for the given speech audio segment were an array of floats. The

array of pitch values for the given speech file was then passed on to two different functions, one which averaged all of the pitch values in the array, and one which calculated the standard deviation of the pitch values in the array. It was kept track of whether the resulted pitch standard deviation was calculated from a female or from a male speech segment. The average pitch of each of the small speech audio segments was saved an eventually an average of the averages was calculated and saved as the final overall pitch average over the whole episode. In case of the pitch standard deviation, and average of the pitch standard deviation of the small speech audio segments was calculated separately for all the female speech and separately for all the male speech after which these two values of the average of standard deviation of all the female speech and the average of standard deviation of all the male speech in the episode were added together and divided by two in order to get the final pitch standard deviation value for the whole episode. The aforementioned steps for calculating the pitch standard deviation for the whole podcast episode were used in order to keep the possible leaps in pitch values between speaker change from female to male speech and vice versa from affecting the magnitude of the overall pitch standard deviation. Finally, the pitch average over the whole podcast episode and the pitch standard deviation over the whole podcast episode were saved to the same .csv file where the percentages of music, noise, silence, female speech, and male speech and the speech rate were saved.

5.2.5 Analysis

After computing stylistic feature values for the podcast dataset with the help of the developed program, The focus was to answer the question: “Can stylistic features be used to discriminate between podcasts and if so, what are the stylistic features which can be used to do so?” Hence, the interest was in potential differences in the stylistic feature values between podcasts, and more specifically in how the stylistic features varied between episodes, since an episode is by design a natural unit of podcasts consumption.

Podcast shows are another natural unit for podcasts. A show consists of multiple episodes which are often either related to the same topic or are hosted by the same people through the whole show. Therefore, we were interested in exploring if episodes within a show would bear stylistic similarities to each other, and if the episodes within a show would be more similar to each other than to episodes in other shows.

Lastly, in many streaming services podcast shows are grouped to genres based on the topic of the shows. These genres are often used as means of organising content on the streaming services’ interfaces. We were interested to find out if differences in stylistic features between genres could be observed or if all the genres would have similar feature value distributions to one another. If the feature value distributions would visibly vary between episodes, shows, or genres, the conclusion could be made that podcasts can be searched for and recommended on the corresponding level based on the tested stylistic feature.

In this section the plan for analysing the results of the automatic feature extraction program is presented. Firstly, the approach for the exploratory data analysis is described. Secondly, the decisions on the statistical hypothesis testing are explained.

Data Exploration

The stylistic feature values were plotted separately for each selected stylistic feature (“monotonousness”, “speech rate”, “music”, “noise”, “silence”, “female speech”, “male speech”, “pitch average” and “speech”). We started by generating histograms for each of the selected features in order to gain an overview of the shape and spread of the episode level distributions. We were interested in whether all the feature values would accumulate around some specific value or whether a wide spread of values could be observed. If variations in the feature value would appear large, we could conclude that the feature shows strong potential for being an useful feature for search and recommendation on an episode level, since the feature would show discriminative power.

Secondly, we were interested in seeing whether episodes within a show would bear more stylistic similarities to each other than to episodes in other shows. In order to explore whether this was the case we plotted the feature values in a following way. The values were visualised in a swarm plot, one dot marking the value of one episode. The episodes were grouped by their shows in order to compare different shows' feature value distributions to each other. In order to effectively visualise different key elements of the feature value distribution over a show, a box plot was drawn over the swarm plot. The box plots show a “minimum”, a first quartile (Q1), a median, a third quartile (Q3), and a “maximum” value of the distribution.

Lastly, we were interested in finding out if any differences in stylistic features between genres could be observed or if all the genres would have similar feature value distributions to one another. In order to explore the genre distributions of the feature values we plotted the feature values similarly to the show plots. This time with swarm plots and box plots were grouped by podcast genres.

Statistical Hypotheses Testing

Since the results of the data exploration showed some differences in the stylistic features' distributions, statistical significance tests were performed to test the hypotheses: "There is difference in the feature value distributions between different shows" and "There is difference in the feature value distributions between different genres".

Since the goal was to compare distributions of different groups to each other (shows or genres) Kruskal-Wallis and Welch's ANOVA tests were used to test the H0 hypothesis that “The means of all the different group's distributions are equal”. The alternative H1 hypothesis was: “At least one of the means is different”. Kruskal-Wallis and Welch's ANOVA were selected instead of any standard t-tests, because of the need to compare more than 2 groups with each other. Both Kruskal-Wallis and Welch's ANOVA were selected because in terms of the prerequisite assumptions of any available test there was no perfect fit for the data.

For example, Kruskal-Wallis does not assume a normal distribution of a group's values, but is said to require somewhat equal spread of values between the compared groups. In contrast to this, Welch's ANOVA does not require equal spread between of the compared groups, but assumes an underlying normal distribution. In the case of this study, neither of these assumptions could safely be made regarding neither normal distribution of the feature values, nor equal spread's between the different groups. This conclusion was drawn based on observations made from the histograms and

box plots of the data exploration. Considering the above, a decision was made to perform both of these statistical tests, and if they would result in a similar conclusion on rejecting or not rejecting the H0 hypothesis, it was to be assumed that the results of the statistical tests could be trusted.

Comparing 17 shows or 12 genres to each other would have included too many groups to the comparison. This is because the more comparisons are made, the larger the possibility increases that we find at least one group where the mean of the distribution is different from the rest. Thus, comparing this many groups with each other can increase the risk of type 1 error where the H0 hypothesis is wrongly rejected. In order to take this into consideration the statistical tests were performed only on the six most popular podcast shows and genres. Popularity was selected as the selection criteria for the six podcast shows and genres, because finding useful results for the most popular podcasts would achieve the highest impact of the results. The more the podcasts were listened to, the wider the influence of being able to recommend podcasts based on stylistic features would be.

The popularity of a show was approximated by looking at the number of times each episode within the show had been streamed for over 60 seconds of time during each subsequent day from the publishing date of the episode. We looked at seven days, including the publishing day and the sixth day after publishing. The stream count of each episode was summed together and an average count for streams per day was calculated. We examined which shows had the most streams during the publication day, publication day + one day after the publication, publication day + two days after publication, and so on. We found that the list of top six shows with most streams stabilised after two days. This stabilised list was selected as the top six shows for the statistical tests. The top six shows were: “Gynning & Berg - Hitta sig själva”, “The Hottest Take”, “Today in True Crime”, “Famous Fates”, “Heavyweight”, “Parcast Presents: March Mysteries”. Similar procedure was used to find the most popular genres, which were: “Music”, “Sports & Recreation”, “Stories”, “Arts & Entertainment”, “News & Politics”, “Society & Culture” respectively.

5.3 Results

In this section the results of the exploratory data analysis and the statistical hypotheses testing are presented.

5.3.1 Data Exploration Results

We plotted the stylistic features' distributions in three different ways to explore the feature distributions of each selected stylistic feature on an episode level, on a show level, and on a genre level. The selected stylistic features were "monotonousness", "speech rate", "music", "noise", "silence", "female speech", "male speech", "pitch average" and "speech". Next, the stylistic feature value distributions for each of the inspected levels are presented.

Episode Level

First, the stylistic feature value distributions were plotted to histogram plots in order to examine the shape and spread of the distributions and compare episodes to each other. The histograms can be seen in Figures 5.2 - 5.10.

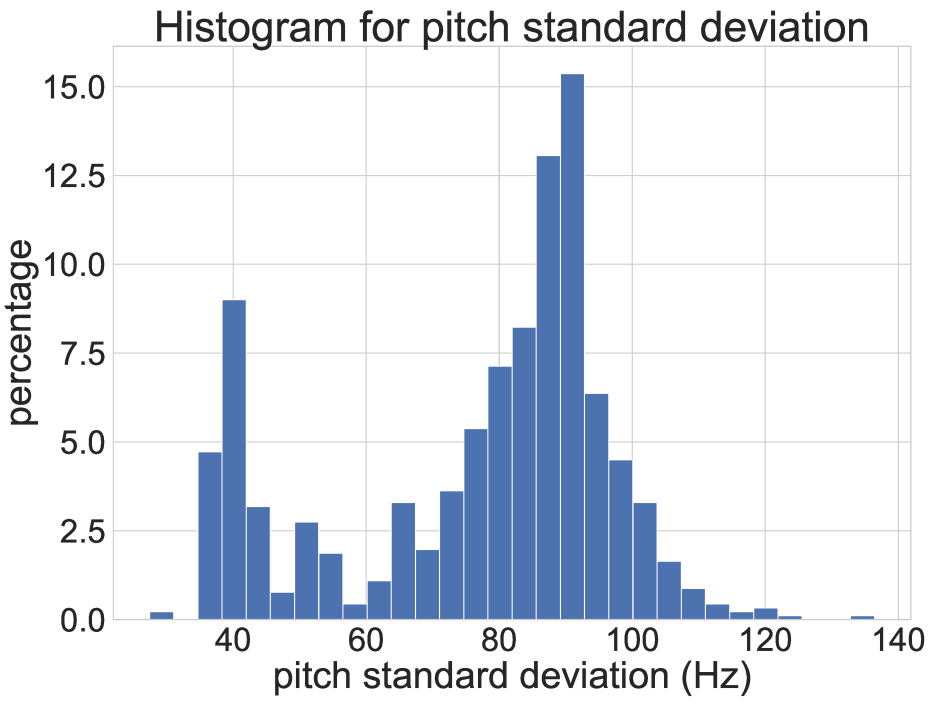

Figure 5.2: Episode level inspection of monotonousness.

The histogram for monotonousness, expressed by standard deviation of pitch, can be seen in Figure 5.2. By definition, the lower the pitch standard deviation is, the more monotonous the speech is. The histogram shows that pitch standard deviation in the sample varies roughly between 30Hz

and 120Hz. Moreover, two peaks in frequency can be observed around 40Hz and 90Hz. This shows that episodes could be referred to as monotonous or less monotonous depending on where they belong to on this scale.

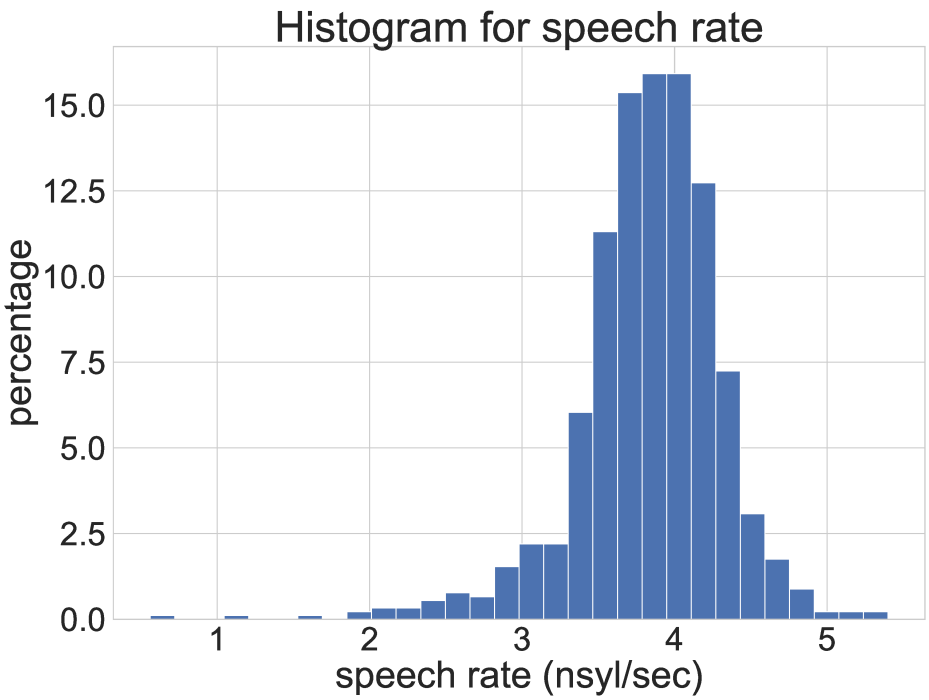

Figure 5.3: Episode level inspection of speech rate.

The distribution of speech rate can be seen in Figure 5.3. One can see from the plot that speech rate is nearly normally distributed across episodes with a slight tail to the left. The values are distributed mainly between 2-5 syllables per second with most of the values lying around 4 syllables per second. Because some spread of speech rate values can be observed between the 2-5 syllables per second, speech rate shows some potential for being used as a basis for search and recommendation on an episode level.

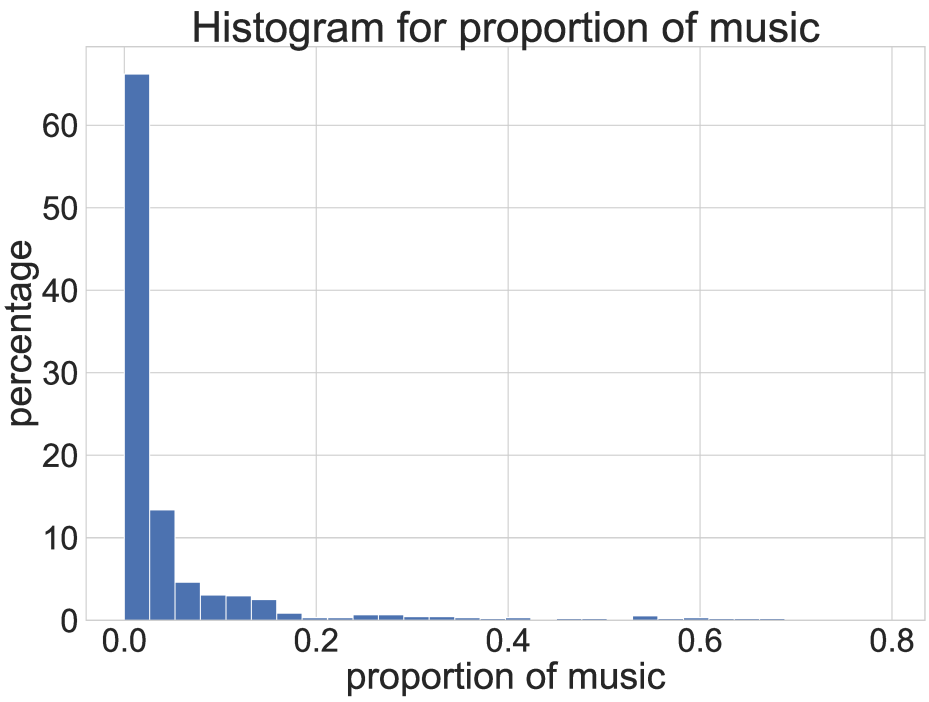

Figure 5.4: Episode level inspection of percentage of music.

The histogram for proportion of music in an episode can be seen in Figure 5.4. From the plot, one can see that most of the episodes in the sample consist of 0-20% of music with a frequency peak close to 0% for music. According to this plot episodes could be for example searched and recommended with the criteria "contains no music" and "contains some music". Hence, it is plausible that music could be used to discriminate between podcast episodes.

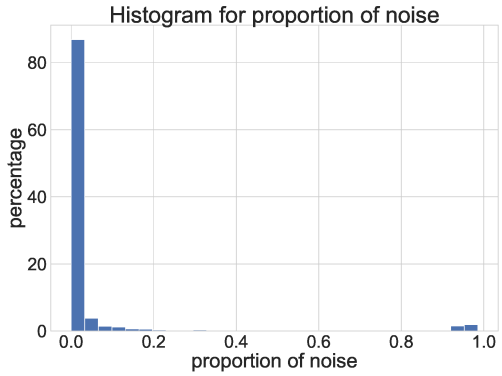

Figure 5.5: Episode level inspection of noise.

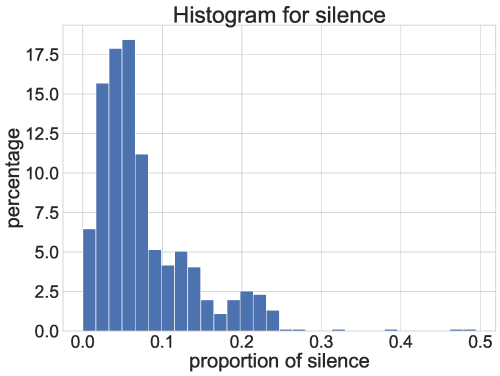

Figure 5.6: Episode level inspection of silence.

The histogram 5.5 shows the proportion of noise in our sample episodes. Noise here can be any other sounds than music, speech or silence, e.g. the sound of rain or wind. The histogram shows that

most of the episodes accumulate between 0-20%, with an exception of some episodes appearing near 100% of noise. Based on this it seems plausible that noise could be used to discriminate between podcast episodes based on this observation.

The histogram 5.6 shows the proportion of silence in our sample episodes. The proportion of silence varies mainly between 0-25% of the episode duration. There seems to be a peak in amount of silence in around 5% of the podcast episode duration. Because a variance between the values of percentage of silence between podcast episodes can be observed, it seems promising that silence could potentially be used to discriminate between podcast episodes.

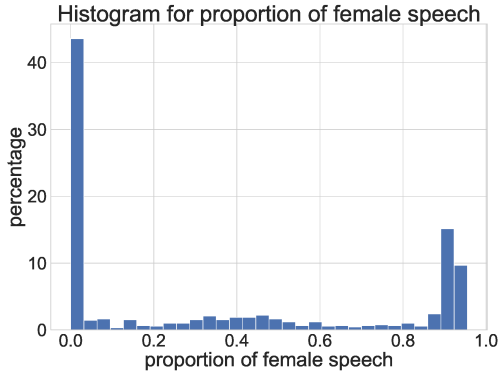

Figure 5.7: Episode level inspection of female speech.

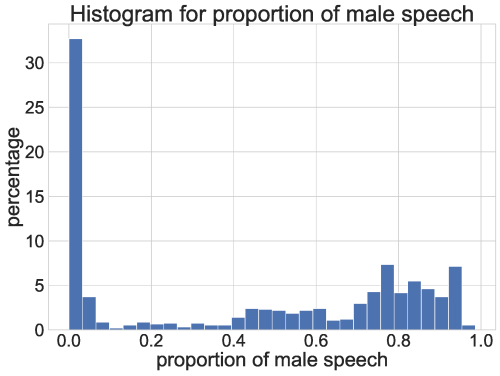

Figure 5.8: Episode level inspection of male speech.

The proportion of female speech in an episode can be seen in Figure 5.7 and the proportion of male speech can be seen in Figure 5.8. The plot in Figure 5.7 shows that most of the podcast episodes have either around 0% of female speech in them or around 90% of female speech in them. Similarly, Figure 5.8 shows a frequency peak around 0%, whereas the rest of the episodes are more evenly distributed between 40% and 100%. Since both of these features, female speech and male speech, show a wide variance between 0-100% we can conclude that these features show potential for discriminating between podcast episodes.

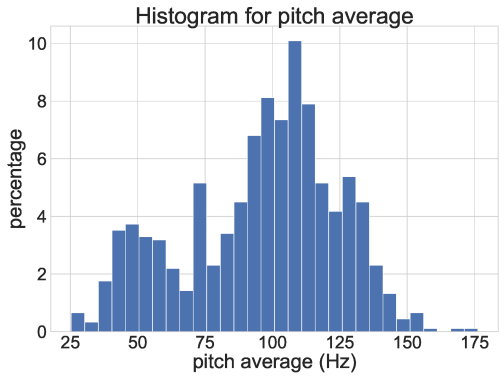

Figure 5.9: Episode level inspection of pitch average.

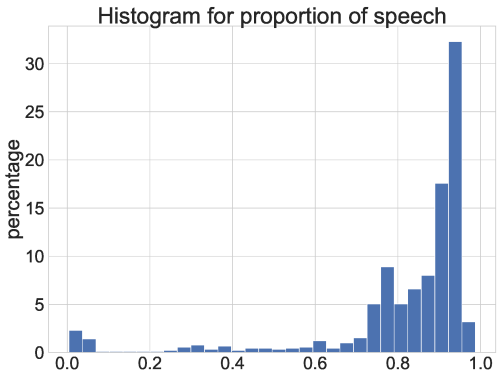

Figure 5.10: Episode level inspection of speech.

Figure 5.9 shows the histogram of pitch average across episodes. The distribution is spread between around 20Hz and 160Hz. This shows potential for using pitch average for differentiating between podcast episodes.

Figure 5.10 shows the histogram of proportion of speech in a podcast episode across the sample episodes. The distribution is left skewed with a peak around 90%. All in all the distribution spreads from 0% to 100%. Based on this we can conclude that proportion of speech within an episode shows potential for discriminating between podcast episodes.

Show Level

Next, we were interested in seeing whether episodes within a show would bear more stylistic similarities to each other than to episodes in other shows. If this was the case, it could be concluded that the feature in question could be used to discriminate between podcast shows. In order to explore this the stylistic feature values were plotted to swarm plots and box plots grouped by shows. In this section, the results of the exploration on the show level potential for discrimination is presented. For clarity, one can see the explored shows and short descriptions of them in Table 5.1 to get an idea of what kind of shows were explored.

Table 5.1: Explored shows and their short descriptions.

SHOW NAME	DESCRIPTION
Optimal Living Daily: Personal Development & Minimalism	"I read you the best content on personal development, minimalism, productivity, and more, with author permission. Think of Optimal Living Daily as an audioblog or blogcast where the best blogs are narrated for you for free. "
Today in True Crime	"Every day, we flip back the calendar and examine a true crime event from the same date."
Bore You To Sleep - Sleep Stories for Adults	"The Sleep Podcast that reads you a short story to try and help you sleep. The stories are read in slow English so they can also help listeners everywhere improve their English and improve their listening, while getting a good night's rest. "
Alpha Male Strategies	"Teaching men real alpha male qualities."
Parcast Presents: March Mysteries	"In Parcast Presents: March Mysteries, we try to uncover the circumstances surrounding some of history's strangest murders, disappearances, and sightings exposing more about the unknown than ever before."
The Receipts Podcast	"#TheReceiptsPodcast is a fun, honest podcast fronted by three girls who are willing to talk about anything and everything. From relationships to situationships to everyday life experiences, you can expect unadulterated girl talk with no filter."
Motivation and Inspiration for Ambitious Achiever	"Motivation and inspiration for who wants to feel inspired and motivated during the day."
Purely Being Guided Meditations	"Weekly guided meditations to relax your mind, soothe your soul and awaken the LIGHT within."
Sleep and Relax ASMR	"Sleep and Relax ASMR is a weekly podcast that creates audio experiences designed to help people sleep and relax. The show uses various ASMR triggers including whispers, gentle speaking, relaxing background noise, and general ambiance to help people unwind and relax from their busy lives."
The Hottest Take	"Bill Simmons and his friends from The Ringer will debate, defend, and parse a controversial opinion on a pressing topic of the day."
Famous Fates	"Discover the stories of incredible people whose grandiose lives were matched only by their shocking demise."
Coach Corey Wayne	"Life & Peak Performance Coach. I Teach Self-Reliance."
Crimetown	"Each season, we investigate the culture of crime in a different city."
Heavyweight	"Join Jonathan Goldstein for road trips, thorny reunions, and difficult conversations as he backpedals his way into the past like a therapist with a time machine."
Gynning & Berg - Hittar sig sjalva	"Meet Carolina Gynning and Carina Berg, two who are definitely not afraid to do away with themselves. Both love to immerse themselves in human behavior, both their own and others'. Exclusively for Spotify, they now delve into the small and big issues of life from a highly personal perspective."
English Speeches | Learn English	"Here you can listen to all the speeches provided on our channel or website. The main idea in this podcast is helping English students to practice their listening and reading."
Big Little Life with The Dashleys	"Listen to Dallin and Ashley as they go behind the scenes and chat candidly about marriage, family, parenting, business, faith, and more."

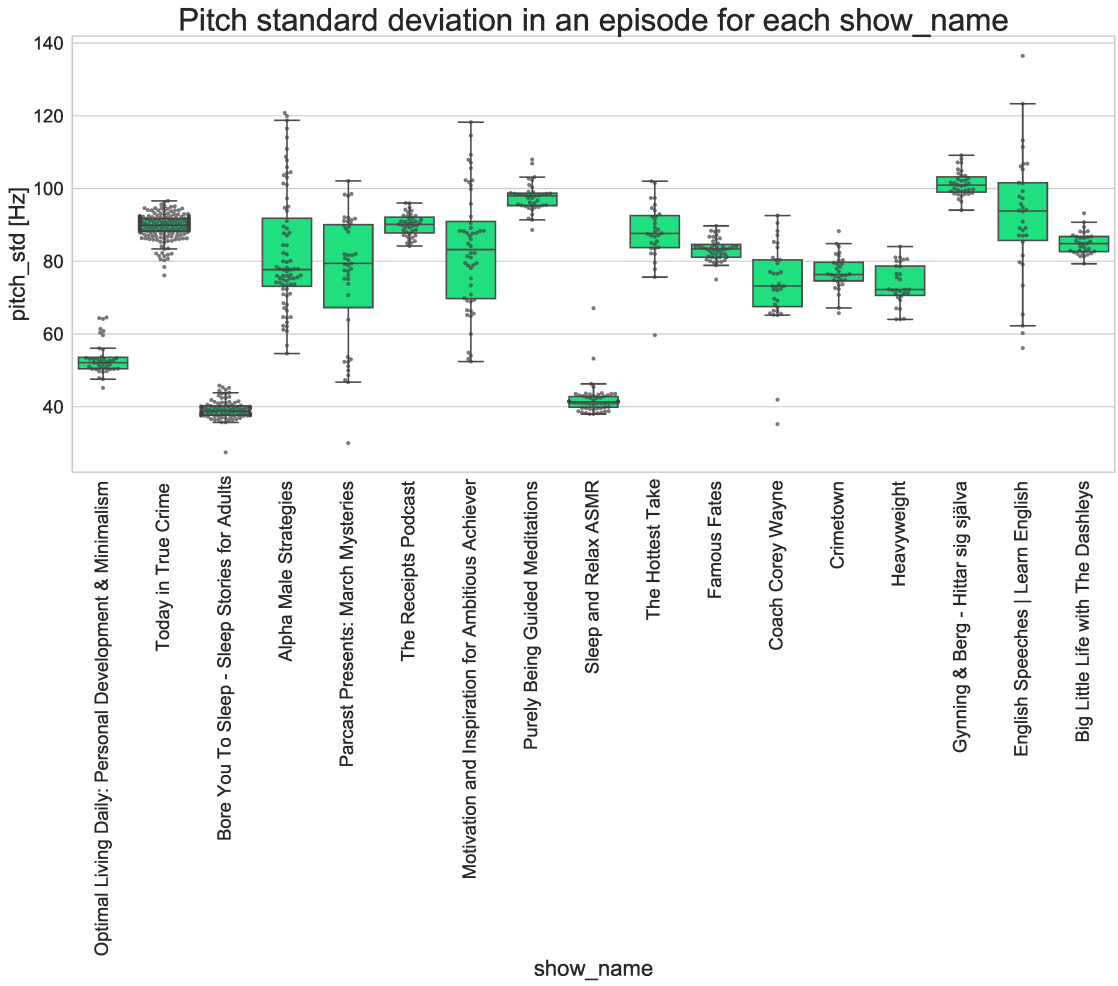

Figure 5.11: Show level inspection of monotonousness.

Figure 5.11 shows the stylistic feature “monotonousness”, as discussed earlier, in the form of pitch standard deviation, grouped by podcast shows. In other words, the lower the pitch standard deviation is in the plot, the higher the monotonousness of the speech in the episode is. Some differences in terms of monotonousness can be observed. For example, according to the plot, the shows “Bore You To Sleep - Sleep Stories for Adults” and “Sleep and Relax ASMR” contain very monotonous speech. On the other hand the shows “Today in True Crime”, “The Receipts Podcast”, “Purely Being Guided Meditations” and “Gynning & Berg - Hittar sig själva” contain speech which is much less monotonous compared to the aforementioned podcast shows. Therefore, by looking at the plot, one can conclude that monotonousness shows potential for be used to discriminate between podcast shows.

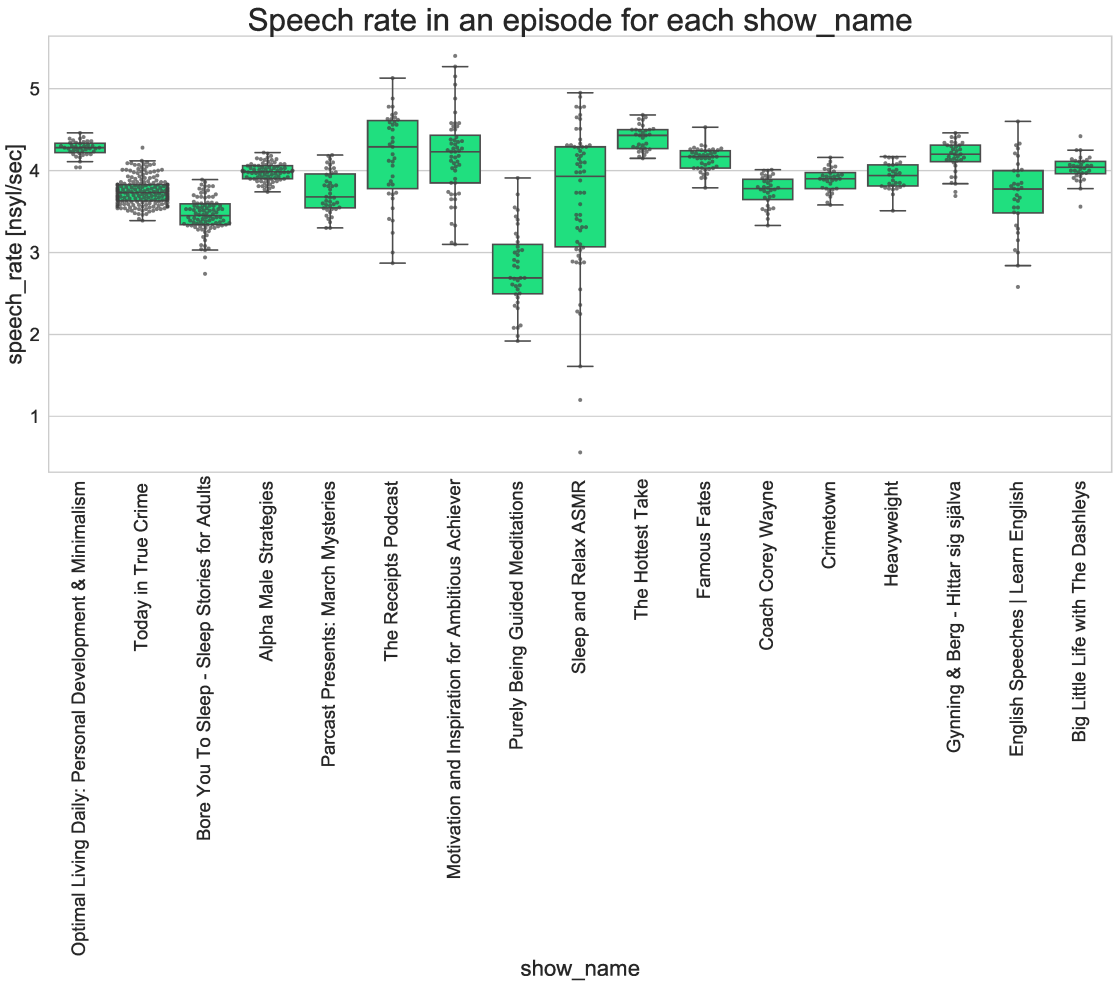

Figure 5.12: Show level inspection of speech rate.

The distribution of speech rate across different podcast shows can be seen in Figure 5.12. One can see that for example the shows “Optimal Living Daily: Personal Development & Minimalism” and “The Hottest Take” have a much faster speech rate compared to the show “Purely Being Guided Meditations”. Additionally, different shows have different spreads of speech rate values. For example, “Today in True Crime” has a small spread of the speech rate values whereas “Sleep and Relax ASMR” has a large degree of variation between its episodes. All in all, based on the fact that variations in the distributions of speech rate between shows can be observed, it can be concluded that speech rate shows potential for be used to discriminate between podcasts on a show level.

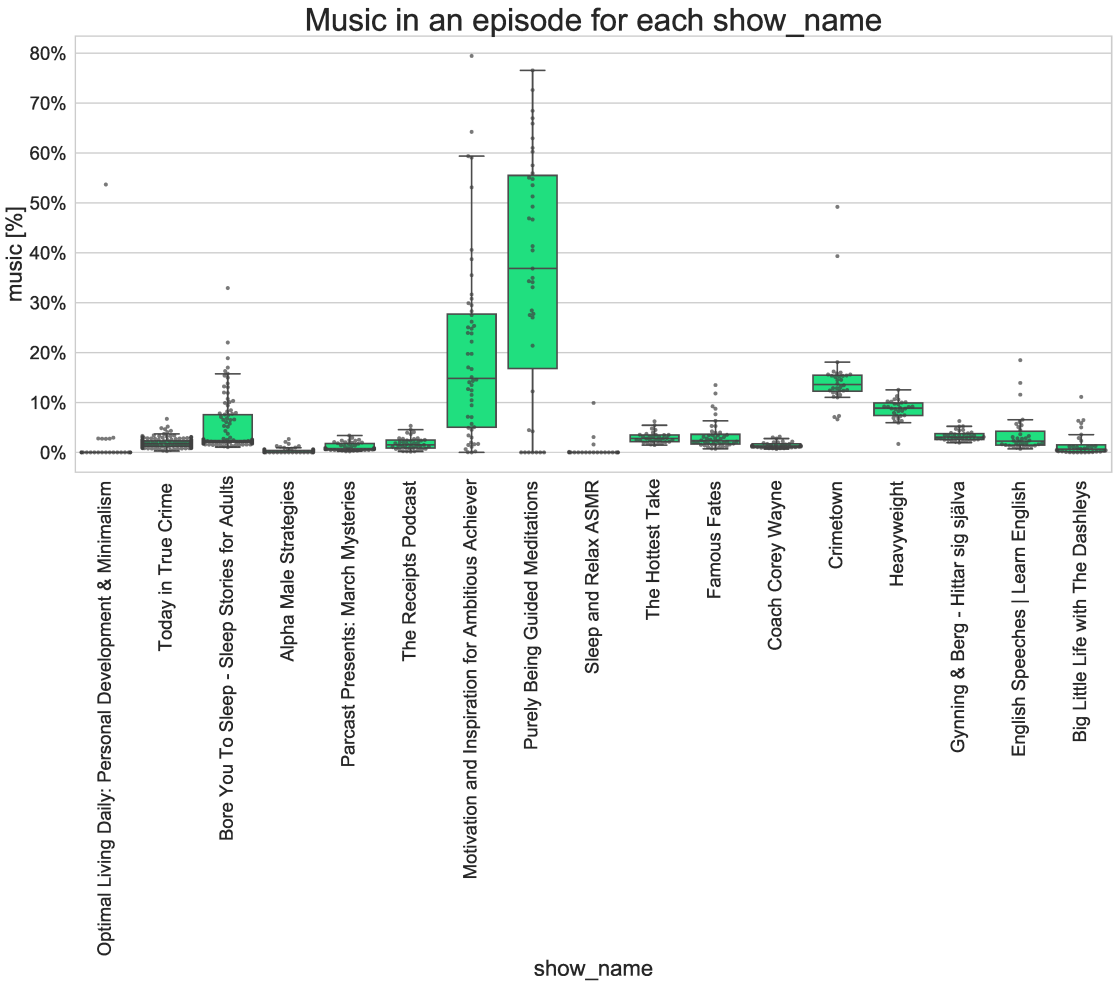

Figure 5.13: Show level inspection of percentage of music.

Like monotonousness and speech rate, percentage of music seems to also vary between of shows. The distribution of percentage of music between shows can be seen in Figure 5.13. For some shows, episodes have a larger spread than in other shows. For example, the shows "Motivation and Inspiration for Ambitious Achiever" and "Purely Being Guided Meditations" exhibit a larger spread of music percentage to any other show in our sample. "Motivation and Inspiration for Ambitious Achiever" contains mostly 5-28% and "Purely Being Guided Meditations" 20-55% of music whereas the other shows seem to have a spread width of around 5% or less and have mostly less than 10% music in them. This larger spread between some shows than other shows might indicate a difference between discriminative power on an episode level between shows. In general, the variations show that percentage of music might be plausible for being used to discriminate podcasts from each other on a show level.

The rest of the stylistic features, namely, noise, silence, female speech, male speech, pitch average, as well as speech in general also showed a lot of variation between shows as shown by the different spreads and locations of the box plots. In other words, the distributions do not vary only

by the position of the mean value, but also in the size of the spread of the show distributions. This might indicate that where there is a difference in the size of the spread, there might be a difference between the discriminative power on an episode level between the shows. At large, it can be concluded that all these features (noise, silence, female speech, male speech, pitch average, as well as speech independent of sex) show promise for being used to discriminate podcasts on a show level. The rest of the plots can be found in appendix E.

Genre Level

The value distribution of the stylistic features were also grouped and plotted by genres in order to explore if any visible differences in the distributions between genres could be observed. The distribution of monotonousness can be seen in Figure 5.14 where the pitch standard deviation is plotted according to the different podcast genres, called "show_cat_name" in the graph. Each episode (a dot in the swarm plot) is marked with a colour of the show the episode belongs to in order to visualise how the different shows make up the distribution of a genre. One can for example see that "Lifestyle & Health" genre is represented by 4 shows, which together create quite a large spread between of pitch standard deviation of 40Hz and 100Hz. On the other hand the genre "Stories", represented by three shows has a relatively small spread of pitch standard deviation. Most of the episodes in this genre fall between values of 80-95Hz of pitch standard deviation. Generally, the spread size of the feature values varies between genres, and so does the location of the median value. Based on this we can conclude that monotonousness could be used to discriminate podcasts between podcast genres.

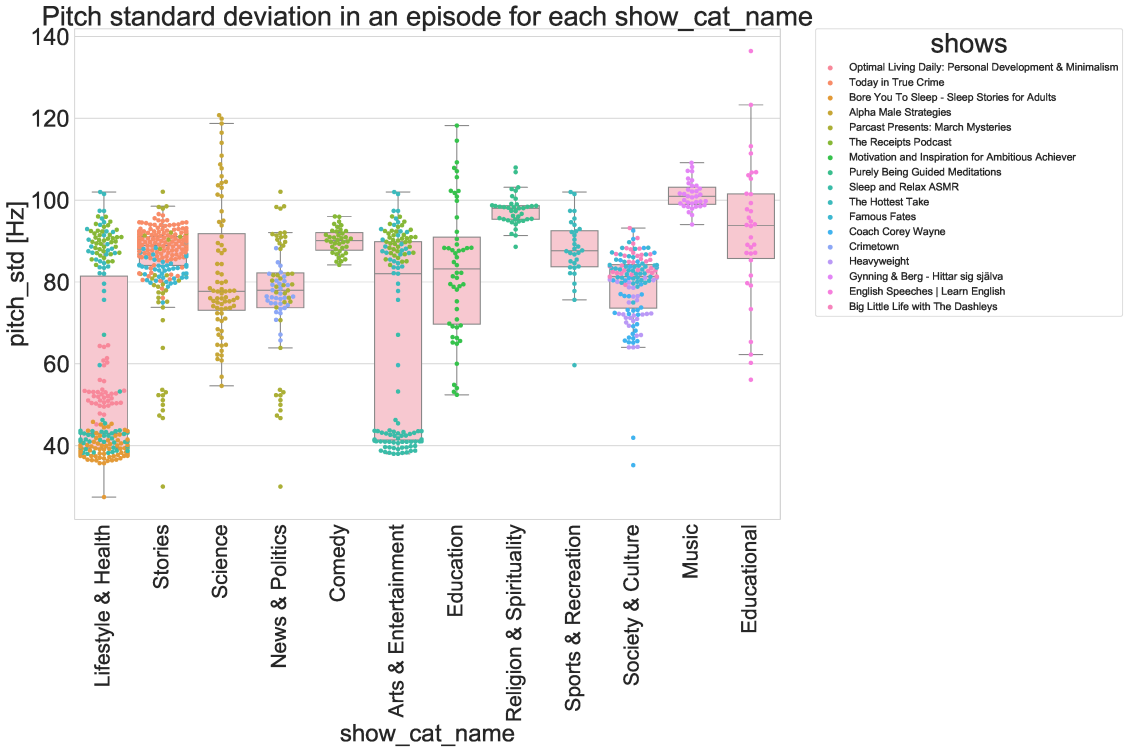

Figure 5.14: Genre level inspection of monotonousness.

The Figure 5.15 shows the distributions of speech rate across different podcast genres. The plot shows that the genre "Religion & Spirituality" has a relatively slow speech rate whereas "Sports & Recreation" has a relatively high speech rate. Variations in the spread of the distributions also exist. For example, the spread of "Lifestyle & Health" appears to be much wider than the spread of "Stories". In general, variations exist between the values of speech rate across podcast genres. Therefore, speech rate seems to be useful for discriminating between podcast genres.

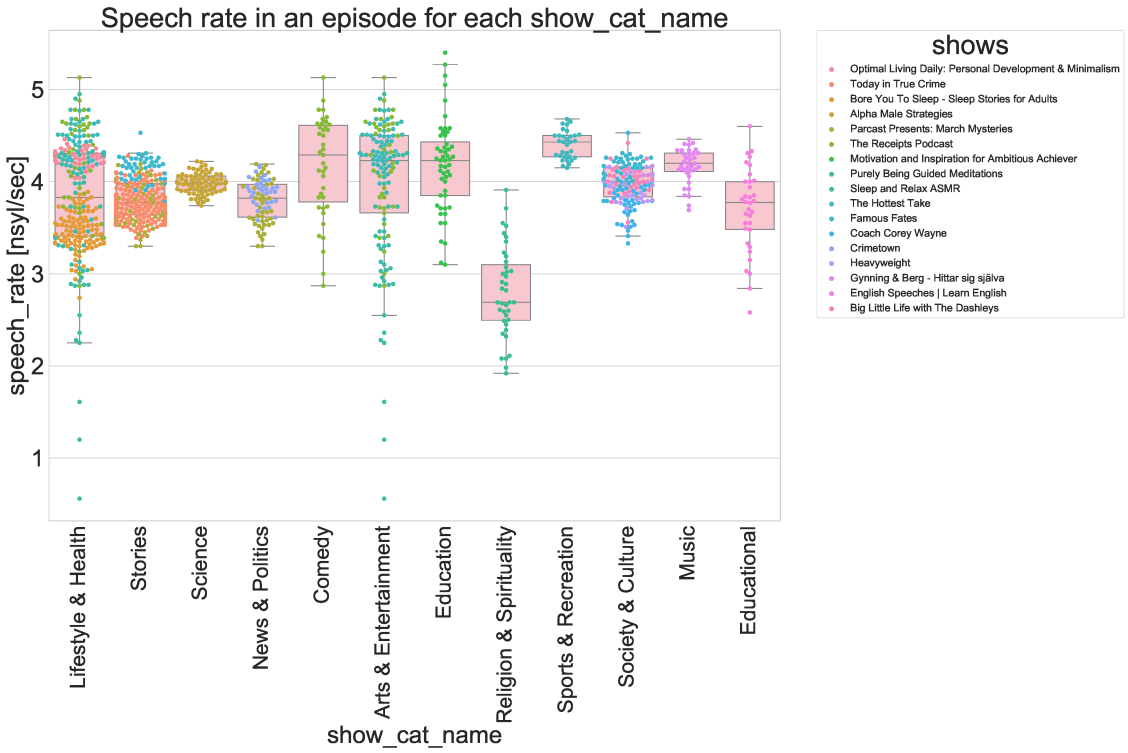

Figure 5.15: Genre level inspection of speech rate.

Distribution of the percentage of music across different podcast genres can be seen in Figure 5.16. Most of the genres' distributions have quite a narrow spread. For example, "Science" appears to have nearly no music in it's podcast episodes. On the other hand, "News & Politics" and "Society & Culture" appear to have roughly up to 20% of music in them. Moreover, "Education" and "Religion & Spirituality" appear to contain up to around 60-80% music in their episodes. Based on these observable differences, we can conclude that music can be used to discriminate between podcast genres.

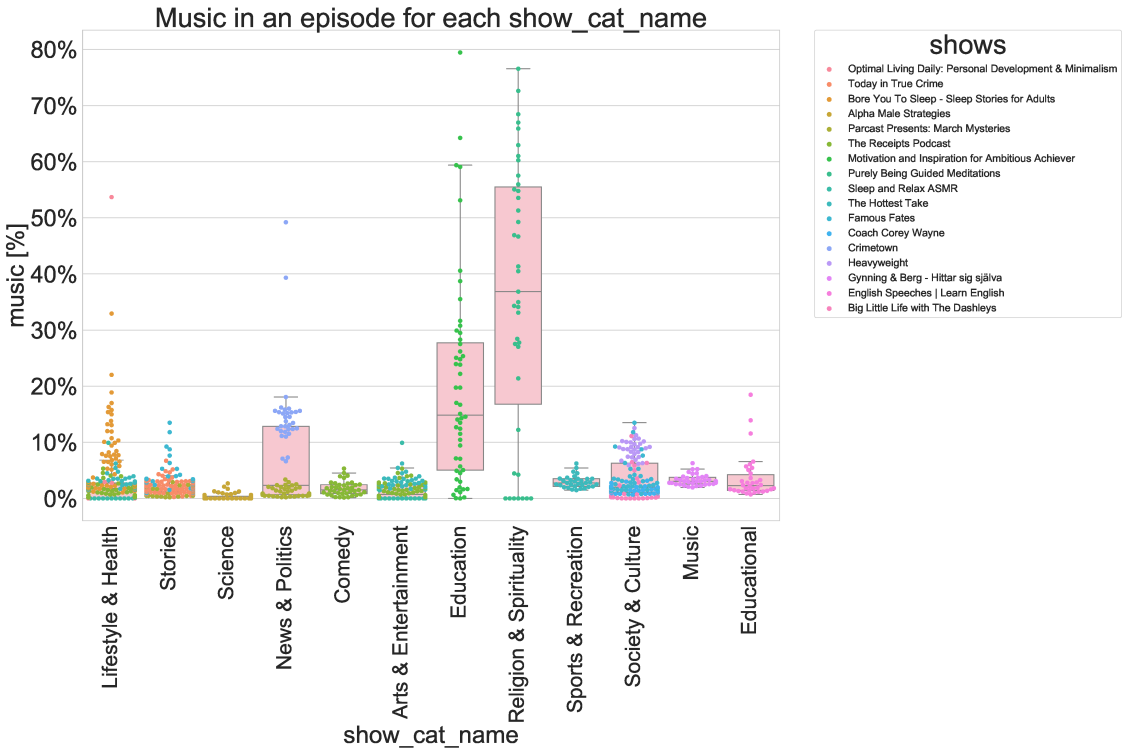

Figure 5.16: Genre level inspection of percentage of music.

In like manner to the genre-wise distributions of monotonousness, speech rate, and music inspection of plots for the rest of the stylistic features (noise, silence, female speech, male speech, pitch average, and speech in general) showed differences in the spread and the location of the distributions. Based on this it can be concluded that these features can also be used to discriminate between podcasts on a genre level. The rest of the plots can be found in appendix E.

5.3.2 Statistical Hypotheses Testing Results

Since the interest was in comparing distributions of different groups to each other (shows or genres), Kruskal-Wallis and Welch's ANOVA tests were used to test the H0 hypothesis that "The means of all the different group's distributions are equal". The alternative H1 hypothesis was then that "At least one of the means is different".

The tests were carried out with the software tool SPSS for only the top six most popular shows and six most popular shows genres in the dataset in order to limit to an appropriate number of groups considering the statistical tests. The top six shows were: "Gymning & Berg - Hitta sig själva", "The Hottest Take", "Today in True Crime", "Famous Fates", "Heavyweight", "Parcast Presents: March Mysteries". The six most popular genres were: "Music", "Sports & Recreation", "Stories", "Arts & Entertainment", "News & Politics", "Society & Culture" respectively.

First, we present the results of the statistical tests for the six most popular podcast shows. Both of the statistical tests, Kruskal-Wallis test and the Welch's ANOVA test, resulted in p-values < 0.01. The

test statistics of Kruskal-Wallis tests for show level can be seen in Figure 5.17. The test statistics of Welch's ANOVA tests for show level can be seen in Figure 5.18. Because the assumption of homogeneity of variance was not met for the inspected shows, the robust test of equality of means was used in the Welch ANOVA test. The results of both of the tests show that the H0 hypothesis can be rejected with a significance level of 0.05. This means that at least one of the tested shows have a different distribution to the rest of the top six shows.

Test Statistics[a,b]

	music	pitch average	speech rate	pitch standard deviation	speech
Kruskal-Wallis H	168.337	214.819	213.476	204.596	121.287
df	5	5	5	5	5
Asymp. Sig.	.000	.000	.000	.000	.000

Test Statistics[a,b]

	male speech	noise	female speech	silence
Kruskal-Wallis H	237.641	154.585	234.583	168.184
df	5	5	5	5
Asymp. Sig.	.000	.000	.000	.000

a. Kruskal Wallis Test

b. Grouping Variable: show name

Figure 5.17: Test statistics of Kruskal-Wallis tests for show level.

Robust Tests of Equality of Means

		Statistic[a]	df1	df2	Sig.
music	Welch	87.621	5	93.287	.000
noise	Welch	59.013	5	105.699	.000
silence	Welch	108.292	5	99.837	.000
speech	Welch	119.690	5	91.590	.000
female speech	Welch	849.468	5	101.684	.000
male speech	Welch	807.896	5	100.919	.000
pitch average	Welch	408.315	5	101.685	.000
pitch standard deviation	Welch	170.857	5	92.559	.000
speech rate	Welch	145.052	5	97.876	.000

a. Asymptotically F distributed.

Figure 5.18: Test statistics of Welch's ANOVA tests for show level.

We carried out a post hoc Tukey test to identify which shows are different from each other. The test showed that for example in terms of the stylistic feature "monotonousness" (pitch standard deviation) the show "Gynning & Berg - Hittar sig sjalva" was different to any other show. The shows "Heavyweight" and "Parcast Presents March Mysteries" were not statistically significantly different from one another in terms of monotonousness. Similarly, "Famous Fates" and "The Hottest Take" were not statistically significantly different from one another. "Today in True Crime" and "The Hottest Take" were similar to each other in terms of monotonousness. The homogeneous subsets of the tested shows in terms of the stylistic feature "monotonousness" can be seen in Figure 5.19.

pitch standard deviation

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05			
		1	2	3	4
Heavyweight	30	73.6600			
Parcast Presents March Mysteries	43	76.0681			
Famous Fates	44		83.2564		
The Hottest Take	32		87.7080	87.7080	
Today in True Crime	194			89.5782	
Gynning & Berg - Hittar sig själva	35				101.2665
Sig.		.634	.051	.836	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.19: Post hoc Tukey for shows - homogeneous subsets for monotonousness.

In terms of the stylistic feature "speech rate" there were also similarities and differences between shows as seen in Figure 5.20. The shows "Parcast Presents March Mysteries" and "Today in True Crime" were similar to each other in terms of speech rate. "Heavyweight" was different to the other shows in terms of speech rate. "Famous Fates" and "Gynning & Berg - Hittar sig själva" were similar to each other in terms of speech rate. "The Hottest Take" was different to the other shows in terms of speech rate.

speech rate

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05			
		1	2	3	4
Parcast Presents March Mysteries	43	3.7395			
Today in True Crime	194	3.7453			
Heavyweight	30		3.9297		
Famous Fates	44			4.1416	
Gynning & Berg - Hittar sig själva	35			4.1691	
The Hottest Take	32				4.4006
Sig.		1.000	1.000	.979	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.20: Post hoc Tukey for shows - homogeneous subsets for speech rate.

In terms of the stylistic feature "percentage of music" the shows "Parcast Presents March Mysteries" and "Today in True Crime" were similar to each other. Similarly, the shows "The Hottest Take", "Gynning & Berg - Hittar sig själva" and "Famous Fates" were similar to each other in terms of the percentage of music in them. The show "Heavyweight" was different to the other shows in terms of the percentage of music. The homogenous subsets can be seen in Figure 5.21. The results of the post hoc Tukey for the rest of the stylistic features can be found from Appendix G.

Percentage of music

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05		
		1	2	3
Parcast Presents March Mysteries	43	.0115		
Today in True Crime	194	.0188		
The Hottest Take	32		.0298	
Gynning & Berg - Hittar sig själva	35		.0331	
Famous Fates	44		.0338	
Heavyweight	30			.0857
Sig.		.166	.787	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.21: Post hoc Tukey for shows - homogeneous subsets for percentage of music.

Next, we present the results of the statistical tests for the six most popular podcast genres. The test statistics of Kruskal-Wallis tests for genres can be seen in Figure 5.22. The test statistics of Welch's ANOVA for genres can be seen in Figure 5.23. Because the assumption of homogeneity of variance was not met for the inspected genres, the robust test of equality of means was used in the Welch ANOVA test. Both of the statistical tests, Kruskal-Wallis test and the Welch's ANOVA test for genre level inspection resulted in p-values < 0.01. These results show that the H0 hypothesis can be rejected with a significance level of 0.05. This means that at least one of the tested genres have a different distribution to the rest of the six most popular genres.

Test Statistics[a,b]

	music	noise	silence	speech	female speech	male speech
Kruskal-Wallis H	99.142	119.185	95.879	111.628	264.984	254.144
df	5	5	5	5	5	5
Asymp. Sig.	.000	.000	.000	.000	.000	.000

Test Statistics[a,b]

	pitch average	pitch standard deviation	speech rate
Kruskal-Wallis H	153.224	217.806	168.247
df	5	5	5
Asymp. Sig.	.000	.000	.000

a. Kruskal Wallis Test

b. Grouping Variable: genre

Figure 5.22: Test statistics of Kruskal-Wallis tests for genre level.

Robust Tests of Equality of Means

		Statistic[a]	df1	df2	Sig.
music	Welch	28.218	5	160.148	.000
noise	Welch	68.096	5	183.213	.000
silence	Welch	34.656	5	166.961	.000
speech	Welch	41.535	5	151.683	.000
female speech	Welch	464.148	5	187.391	.000
male speech	Welch	492.156	5	184.795	.000
pitch average	Welch	272.695	5	173.153	.000
pitch standard deviation	Welch	164.770	5	170.192	.000
speech rate	Welch	93.822	5	158.287	.000

a. Asymptotically F distributed.

Figure 5.23: Test statistics of Welch's ANOVA tests for genre level.

We carried out a post hoc Tukey test to identify which genres are different from each other. The test showed that for example in terms of the stylistic feature "monotonousness" (pitch standard deviation) the genre "Arts & Entertainment" was different to any other genres. Similarly, the genre "Music" was different to any other genre. The genres "News & Politics" and "Society &

Culture" were not statistically significantly different from one another in terms of monotonousness. Similarly, "Stories" and "Sports & Recreation" were not statistically significantly different from one another. The differences and similarities between the different podcast genres for the stylistic feature "monotonousness" can be seen in the table of homogeneous subsets in Figure 5.24.

pitch standard deviation

Tukey HSD[a,b]

		Subset for alpha = 0.05			
genre	N	1	2	3	4
Arts & Entertainment	133	66.9929			
News & Politics	75		76.4400		
Society & Culture	140		79.2007		
Stories	281			86.5209	
Sports & Recreation	32			87.7080	
Music	35				101.2665
Sig.		1.000	.850	.996	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.24: Post hoc Tukey for genres - homogeneous subsets for monotonousness.

In terms of speech rate, there was no statistical significant difference between the genres "News & Politics", "Stories", "Society & Culture" and "Arts & Entertainment". However, "News & Politics", "Stories", "Society & Culture" were statistically different to "Music". The genre "Sports & Recreation" was different to every other genre in terms of speech rate. The differences and similarities between the different podcast genres for the stylistic feature "speech rate" can be seen in the table of homogeneous subsets in Figure 5.25.

speech rate

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05		
		1	2	3
News & Politics	75	3.8023		
Stories	281	3.8065		
Society & Culture	140	3.9759		
Arts & Entertainment	133	3.9826	3.9826	
Music	35		4.1691	
Sports & Recreation	32			4.4006
Sig.		.081	.063	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.25: Post hoc Tukey for genres - homogeneous subsets for speech rate.

As seen in Figure 5.26 there are similarities and differences between podcast genres in terms of percentage of music in a podcast episode. In case of the percentage of music, “Arts & Entertainment”, “Stories” and “Sports & Recreation” are not different from each other. At the same time “Music” and “Society & Culture” are different from “Arts & Entertainment” but similar to “Stories” and “Sports & Recreation”. In terms of the percentage of music, the genre “News & Politics” is different to any other of the six genres. The whole post hoc Tukey analysis for all the stylistic features can be found in Appendix I.

Percentage of music

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05		
		1	2	3
Arts & Entertainment	133	.0136		
Stories	281	.0200	.0200	
Sports & Recreation	32	.0298	.0298	
Music	35		.0331	
Society & Culture	140		.0364	
News & Politics	75			.0713
Sig.		.089	.082	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.26: Post hoc Tukey for genres - homogeneous subsets for percentage of music.

5.4 Conclusions and Discussion of Chapter 5

In this section the conclusions and discussion of chapter 5 are presented. In addition, some limitations of the second phase of the study are discussed.

5.4.1 Conclusions and Discussion

In this chapter we have presented our work on answering the research question RQ2: "How suitable are the listener perceived stylistic characteristics of podcasts for automatic podcast recommendation and search based on the attainable information from podcast audio signal?". The main focus was on the middle level stylistic features and the low level acoustic features described in our framework. We approached answering RQ2 step-wise by first reflecting on what stylistic features were a sensible starting point for modelling stylistic podcast characteristics. The reflection on the stylistic features was done from four perspectives: what listeners care about according to the user study results, what is generally sensible from the perspective of podcast production and consumption, how does the feature align with the research focus, and how easily can the stylistic feature be automatically detected from podcast audio by using available open source tools. From the results of the user study, with the above perspectives in mind, the study was limited to nine stylistic features, namely "pitch", "monotonousness", "speech rate", proportion of "music", "noise", "silence", "female speech", "male speech" and "speech" in general.

We experimented with the selected features by implementing a program which extracts the selected features directly from podcast audio. The program computed the nine stylistic features for

each podcast episode in the dataset. The stylistic features' values were used in order to answer the question: "Can stylistic features be used to discriminate between podcasts and if so, what are the stylistic features which can be used to do so?" We were interested in whether the selected stylistic features could be used to discriminate between podcasts on different levels, from episode and show level to genre levels. Usefulness for discriminating between podcast episodes, shows, and genres would indicate the feature's suitability for search and recommendation applications.

In order to explore if the selected features could be used to discriminate between podcasts, a data exploration was conducted by computing several different distribution plots. In addition, hypothesis testing was conducted to further test if there were any differences in the distributions of different shows or genres, and post hoc analysis was carried out to investigate which of the examined shows and genres were different or similar to each other. The data exploration showed that a lot of difference in the spread and the location of the distributions of all the selected feature values could be observed. The differences could be observed in all the levels from an episode level to a show and to a genre level. Based on this, it can be concluded that "pitch", "monotonousness", "speech rate", proportion of "music", "noise", "silence", "female speech", "male speech" and "speech" all show potential for being used to discriminate between podcasts on an episode, show, and genre level. Additionally, the statistical hypothesis testing showed that there are statistically significant differences in the feature value distributions on a show level and on a genre level for each of the examined stylistic features. The post hoc tests showed homogeneous subsets among the examined shows and genres. For example, "Parcast Presents March Mysteries" and "Today in True Crime", the two crime podcast shows in the set of six shows used for the statistical testing, were similar to each other in terms of speech rate and percentage of music. For example, in terms of podcast genres, the genres "News & Politics" and "Society & Culture" were similar to each other in terms of monotonousness (pitch standard deviation) and speech rate. Furthermore, all the aforementioned features could be detected automatically from podcast audio signal by using available open source tools. Therefore, it can be concluded that these features show potential for being suitable for automatic podcast recommendation and search based on the attainable information from podcast audio signal.

5.4.2 Limitations

In this section we discuss the limitations of the work related to the automatic extraction of the selected stylistic features, related to the exploratory data analysis, and related to the statistical tests.

When it comes to the open source tools which were used in our program to compute feature values for "pitch", "monotonousness", "speech rate", proportion of "music", "noise", "silence", "female speech", "male speech" and "speech", we did not evaluate the performance of these tools on our podcast dataset. Therefore, our results might suffer from inaccuracies. Especially our segmentation to female speech segmenets, male speech segments, silent segmenets, noise and music segments may suffer from noise, since the segmenter which we used had not been trained on podcasts and was optimised for French language as opposed to English, which was most common in the dataset. Additionally, audio segments which contain music or other sounds over speech might have caused inaccuracies in the values of speech rate, pitch standard deviation, and average pitch.

Despite these expected inaccuracies, we trust that considering the quite large size of the podcasts set used for analysis, our approach offered a rough estimate of the aforementioned features which was sufficient for the research purposes. On the subject of segmenting podcast episodes to "female speech" and "male speech" segments, it is worth mentioning that the audio segmenter is only as good as the model and data used for training it. For example, correctly detecting certain types of female or male voices might depend on if similar voices were present in the training data and on if the model managed to capture important qualities of these voices. Additionally, the segmenter does not take into account non-binary people.

Next, we discuss the limitations of our approach to representing the stylistic features of podcasts by only one value across a whole podcast episode. This approach is a rough estimate of the stylistic content of a podcast episode and disregards possible variations within a podcast episode. Variations within an episode could provide interesting insight to for example whether trends within episode variations occur within and across podcast genres. Understanding the variations within an episode could also be used to for example create an automatic preview of potentially interesting parts of the episode for the user.

When it comes to the exploratory analysis and comparing the stylistic feature values between different episodes, shows, and genres to each other, there are some limitations. First, when considering comparing for example podcast shows to each other based on speech rate, monotonousness (pitch standard deviation), and the average pitch, one should compare shows only within the same language. This is because different languages might have different baselines on what is an average speech rate, pitch, or pitch variation. Therefore, if across language comparison is made, we suggest that the stylistic features are first normalised with each language's baselines.

In regards to the exploratory analysis carried out on genre level, it should be noted that the genres in general were represented with few shows. Some genres were represented by only one show. Thus, one needs to be careful with what kind of conclusions are drawn from these genres. Nevertheless, based on the distribution plots it seems very promising that differences between feature values could occur even on a genre level, and thus the examined stylistic features could be sensible for search and recommendation even on a genre level. However, we predict that as more shows are added to each genre, the differences between most of the feature values will become smaller. Further studies are required on the genre level with more shows included in each genre in order to draw stronger conclusions on the genre level distributions.

Chapter 6

Conclusions and Discussion

The main contributions of the research are presented in this chapter. Furthermore, it is discussed, how the work presented in this report contributes to the existing research fields of spoken content retrieval and podcast content modelling. Some application areas and use cases are also discussed. Lastly, ethical considerations are discussed, and areas of future work are identified.

6.1 Conclusions and Main Contributions

Looking into only topic content of podcasts neglects that podcasts are a rich medium for information. The paralinguistic features of speech are neglected. In addition, many podcasts contain background music and other sounds. This master thesis has presented a novel approach to modelling podcast content, not based on podcasts' textual content, but based on stylistic characteristics of podcasts. The research focused on stylistic characteristics which could be automatically extracted directly from the podcast's audio signal. In order to approach modelling of the stylistic characteristics of podcasts we concentrated on answering two research questions: RQ1: "What stylistic characteristics of podcasts do listeners find interesting or important for their podcast listening experience?" and RQ2: "How suitable are the listener perceived stylistic characteristics of podcasts for automatic podcast recommendation and search based on the attainable information from podcast audio signal?"

A framework for stylistic podcast characteristics was developed to guide the approach and method selection. The framework consists of three levels. The top level comprises of *stylistic categories*, which we see as the broadest units for the stylistic characteristics. The stylistic categories are described by middle level *stylistic features* which in turn can be (hopefully) detected by using low level *acoustic features* which are directly measurable from podcast audio. Research question RQ1 focused on listeners' perception on the stylistic features and the stylistic categories from our framework. Research question RQ2 concentrated on stylistic features and their automatic detection, by using their underlying acoustic features, from podcast audio signal. The research question RQ2 approached the topic from a computational perspective.

6.1.1 User's Perceptions of Stylistic Characteristics of Podcasts

It was discovered that listeners observe and care about various different kinds of stylistic features in podcasts. In general, the stylistic features listeners can observe and verbalise in podcasts relate mostly to music or speech, voice, and talking style. Additionally, listeners care about stylistic features related to topic clarity, audio quality, introduction, and presence of ads. Listeners also observe stylistic features related to the dynamics of the people in the podcast and the flow of the podcast, as well as the presence of laughter or swearing. Additionally, podcast listeners care about stylistic features such as whether the podcast contains conversations or interviews, or whether the podcast contains a lot of silence or other sounds.

Most of the stylistic features, which emerged from the user study, referred to either speech, voice qualities, or talking style. Further inspection of these features showed that the participants of the user study cared about stylistic features related to speech rate, tone of voice, clear speaking, pitch related features like monotonousness of speech, extralinguistic factors such as gender, articulation, and other features such as affect in the voice of the speaker. When it comes to participants' preference of the aforementioned stylistic features, all participants had their own individual preferences. In general, almost every kind of identified stylistic feature varied nearly equally in between being preferable or disliked with a few exceptions for example ads being consistently disliked and stylistic features related to dynamics, flow, and people in the podcasts were always viewed positively.

Seven stylistic categories were constructed based on listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about. The seven categories were "Music and Sound", "Ads/Commercials", "Audio Quality", "Speaking Style and Voice Qualities", "Formality Level", "Engagement Level", and "Format". Each one of these categories have a dimensional aspect to them. Most of the dimensions are related to the right balance of these elements in the podcasts and can largely depend on the listener's preference. Hence, many of the dimensions are named "pleasant - unpleasant", "suitable - unsuitable" or "good - bad". Categories in which dimensions differ from the above are "formality level" with the dimension "serious - leisure" and the category "engagement level" which is characterised by the dimension "engaging - not engaging". The seven stylistic categories can be used for reflecting on the stylistic content of podcasts from the user's perspective. The stylistic categories can be used not only for categorising and recommending existing podcasts, but also for designing new podcasts. Podcast creators are encouraged to reflect on these seven categories and their dimensions whenever deciding on the style of their podcast.

Some additional stylistic features were gathered from existing literature and from expert ideas. It was tested how well these features fit to listeners' perceptions on stylistic podcast categories and their perceptions on which stylistic features are important for their podcast listening experience. The features could be seen in Tables 4.1 and 4.2. All of the stylistic features were deemed important by the user study participants. However, some of the features were seen as less relevant for the listening experience than others. Extralinguistic factors, namely age and gender, seemed to be less relevant to the participants.

6.1.2 Audio-based Automatically Extracted Stylistic Differences Between Podcasts

In this research phase, we explored how suitable are the listener perceived stylistic characteristics of podcasts for automatic podcast recommendation and search based on the attainable information from podcast audio signal. From the results of our user study, with four perspectives in mind (what listeners care about according to our user study results, what is generally sensible from the perspective of podcast production and consumption, how does the feature align with our research focus, and how easily can the stylistic feature be automatically detected from podcast audio by using available open source tools), we narrowed down to nine stylistic features, namely "pitch", "monotonousness", "speech rate", proportion of "music", "noise", "silence", "female speech", "male speech" and "speech" in general. We developed a program, which can automatically extract and compute these stylistic feature.

Inspections of the distribution plots of the computed feature values showed that each of the aforementioned stylistic feature shows potential for being used to discriminate between podcasts on each of the levels of interest from podcast episodes to podcast shows and genres. In order to further test this conclusion, we carried out statistical hypothesis tests to test the hypothesis that at least one of the inspected show's distribution was different to the other shows' distributions. Also the podcast genres were tested similarly. All of the statistical tests resulted in that at least one of the tested shows/genres had a different distribution to the others. This confirmed our conclusion that the selected stylistic features ("pitch", "monotonousness", "speech rate", proportion of "music", "noise", "silence", "female speech", "male speech" and "speech") show potential for being used to discriminate between podcast on a show level, and on a genre level. Post hoc analyses were carried out in order to investigate which of the podcast shows and genres were similar or different to each other. The analyses showed different kinds of homogeneous subsets for the different stylistic features. For example, in terms of podcast genres, the genres "News & Politics" and "Society & Culture" were similar to each other in terms of monotonousness (pitch standard deviation) and speech rate.

Furthermore, as our ability to develop a feature extraction program showed, all the aforementioned features could be detected automatically from podcast audio signal by using available open source tools. In the light of the results of this step-wise exploration, it can be concluded that the features "pitch", "monotonousness", "speech rate", proportion of "music", "noise", "silence", "female speech", "male speech" and "speech" show strong potential for being suitable for automatic podcast recommendation and search based on the attainable information from podcast audio signal.

6.2 Discussion

Previously, spoken content retrieval has been applied in many different application areas, but only recently have the applications moved towards more natural speech formats. Examples of these new application areas include search of meetings [47] [48], call center recordings [49], collections of interviews [50][51], historical archives [52], lectures [53], and political speeches [54]. We contribute to the more recent research and application areas of spoken content retrieval by focusing on the

nearly unexplored research area of podcasts, which is a newer and increasingly more popular form of unplanned, unstructured, and natural speech media.

Traditionally, spoken content retrieval has focused on modelling the topical content of the media by utilising various ASR methods and metadata [11] thus ignoring the possible other (stylistic) dimensions of spoken language and audio media. Laver [16] has described spoken word as carrying information on speaker's emotional state, personality, state of mind, or for example the speaker's gender and age. This paralinguistic and extralinguistic information is often conveyed through variations in pitch, loudness, tempo, and rhythm of speech. The traditional spoken content retrieval approaches of content modelling omit these types of information. However, when it comes to spoken content retrieval, user might prefer to attain results of a certain speaking style, certain speaker, a result of certain length, or speech media of certain quality and format as hypothesised in [11].

In our work, we explored to what extent podcast listeners care about these types of paralinguistic and extralinguistic features as well as features related to the quality and format of podcasts. We found that features related to pitch and rhythm of speech are very important to podcast listeners, whereas gender and age are seen as something less important. The results of our study also showed that podcast listeners find the quality and format of podcasts important for their listening experience. Moreover, the participants had very individual preferences regarding these aspects of podcasts. Therefore, our results contribute to confirming the hypothesis from [11] that user might prefer to attain results of a certain speaking style, certain speaker, a result of certain length, or speech media of certain quality and format. In general, our work has contributed to bringing attention to the paralinguistic, extralinguistic, and other stylistic dimensions of spoken content, and has highlighted the importance of considering such dimensions when designing spoken content retrieval applications.

Previously, some research has been done on characterising podcast content. The study "Predicting Podcast Preference - An Analysis Framework and its Application" [72] has proposed an extensive framework (PodCred) of podcast qualities for predicting podcast popularity. Whereas this framework covers everything from the podcast cover photo to an active home page for the podcast, we concentrated on the stylistic dimensions of podcasts with a particular focus in stylistic features which could be automatically detected from the podcast audio. The PodCred framework suggests some features which we considered to be interesting from the stylistic perspective. These features were appearance of (multiple) on-topic guests, participation of multiple hosts, containing discussion/opinions, fluency, speech rate, articulation and accent of the podcaster, use of conversational style, presence of affect, use of humour, presence of advertisements, signature intro/opening jingle, background music, atmospheric sounds, editing effects, and the quality of the recording. We included these features in the user study. Our results showed that all of these features are important for podcast listeners in terms of their listening experience.

The PodCred framework was validated by implementing a basic classification system that makes use of a select set of popularity indicators from the framework. The focus of the validation was on features that could be extracted with a minimum of crawling or processing effort at the time. Therefore, the validation concentrated on feed-level metadata such as the presence of description, the length of that description, and the number of authors listed in the feed, hence excluding the

content of the podcast audio. In the light of the study "Predicting Podcast Preference - An Analysis Framework and its Application" [72], our research has contributed to the knowledge of podcast characteristics from not only the novel perspective of stylistic content but also from the perspective of detecting the characteristics directly from podcast audio signal, as opposed to the metadata level features tested in "Predicting Podcast Preference - An Analysis Framework and its Application" [72].

The study "More than Just Words: Modeling Non-textual Characteristics of Podcasts" [5] explored the non-textual characteristics of podcasts "seriousness" and "energy". These characteristics were also discovered in our user study. The stylistic category "formality level" from our study included participant generated notes such as "serious", "calm", "excited", "light mood", and "hosts seem to know each other and can joke around in an authentic way(indicated by laughter and jokes)". Whereas the authors of "More than Just Words: Modeling Non-textual Characteristics of Podcasts" [5] defined seriousness from the perspective of funniness, we found that listeners cared about the formality level of the podcasts not only from the perspective of humour, but also from the perspective of how people are talking. For example, whether the podcast consists of relaxed chatting or follows a more conventional format, for instance such as in news broadcasts.

In addition to the traditionally employed ASR systems omitting a variety of stylistic information from the way they represent spoken content, they are known to produce erroneous output when applied to unstructured natural speech, such as podcasts [3]. Furthermore, ASR systems scale poorly across languages. We have contributed towards alleviating the aforementioned problems by proposing an additional way of approaching podcast content modelling for retrieval. More specifically, we proposed that the stylistic dimensions of podcasts could be included into the way podcast content is modelled and indexed. We focused on stylistic features of podcasts which could be automatically extracted from podcast audio signal, hoping that this approach would scale better across languages than what usage of ASR does. This also provides means of understanding and modelling podcast content of minor languages, and therefore supports diverse and inclusive podcast applications.

Besser's study [71] researched how podcast retrieval can be better optimized to meet the needs and search goals of users. Besser [71] identified a set of categories classifying the underlying search goals. The categories consisted of searches for person names, searches for podcasts for which the title or a quotation from an episode is known, and searches for information about a general topic or about a current issue or event. We contributed to identifying podcast listeners' retrieval needs from a stylistic perspective and found that listeners care about many different kinds of stylistic features. The stylistic features resulted from our research could, in addition to Besser's findings, be used to optimize podcast retrieval systems to meet the needs and search goals of users. Furthermore, by showing that some of these features are sensible for automatic podcast search and recommendation, our work is a step towards a better user experience of spoken content retrieval applications. Further benefits and applications of our results are discussed below.

The contributions of our study can help enable novel user interactions with large collections of spoken content, such as podcasts. For example, users could be enabled to search or filter content based on stylistic criteria, empowering the users to match their listening experience to their current mood, aspired mood, or personality. This way users could also tie their podcast search better to their

listening context. We envision that users could for example search for an "energetic podcast about [topic]" when they look for entertainment during workout. On the other hand, users could search for a "calm podcast about [topic]" before bed time or for an "engaging podcast about [topic]" when commuting to work.

The above use case is not limited to search but can also extend to context aware recommendation. Similarly, someone who enjoys hearing relatively calm and slow speech could be helped to find pleasant podcasts by either the system recommending podcasts with slow speech rate or by ranking podcasts with low speech rate higher in the returned search results. Another use case is enabling systems to recommend stylistically similar content to recently played podcasts. We hypothesise that this in itself has potential to increase the likelihood that the user finds the recommended podcast pleasant to listen to and hence increases the success rate of the podcast recommendation algorithm.

Finally, if audio media recommendations were made not only on the basis of content, but also on the basis of style, forming of filter bubbles around listeners' consumption could perhaps be mitigated. Filter bubbles are spheres of homogeneous type of content, in which consumers might become stuck. They are the result from personalized searches when a website algorithm selectively guesses what information a user would like to see based on information about the user, such as location or past click-behavior. On the other hand, rigorously recommending content based on style can lead to stylistic bubbles. Therefore, it is important that the creators of recommendation algorithms are aware of the possibility for filter bubbles and think how diverse content can be made available the best way to users by design. All things considered, offering consumers novel ways to discover content based on not only the topical factors but also the style can offer a competitive advantage to other audio media providers and can be a differentiating factor on the market.

6.3 Ethical Considerations

We would like to highlight here that we do not see all features which resulted from the user research as suitable as others to be used as the basis for recommending podcasts. We particularly point to the extralinguistic factors of age, gender, and accent which came up in the user research. These are factors which the speaker cannot easily change about themselves and which might be sensitive to some groups. With such features it is very important to ask oneself what the benefit for using such features is and if using them is necessary. For example, we needed to detect the perceived gender of the speaker in order to compute the pitch variations across an episode as correctly as possible. It might also be useful to automatically detect perceived gender of speakers in order to observe what kind of gender distributions lies in the whole podcast catalogue and if for example any gender disparities can be observed between podcast genres.

It is a different matter if podcast recommendation would benefit from using gender as one of the features or if for example using pitch average across an episode would be enough to approximate the same voice qualities as what we perceive in a speaker's gender. Furthermore, recommending content based on accent should be heavily advised against because using accent has a high risk of leading to demographic or cultural filter bubbles, and can in general be a very sensitive matter.

6.4 Future Work

Whilst reporting our studies we have pointed out some interesting areas for future studies. We provide a summary of interesting future work in this section.

One interesting area for future studies would be to explore the feature value distributions between different users. This kind of study could help gain further insight into users' stylistic taste of podcasts and quantitatively validate the results of our user research.

Another natural next step for this study would be to implement a recommendation algorithm based on the stylistic features proposed in this study, deploy the algorithm into an existing podcast service and conduct A/B testing to see if improvements in recommendation performance can be observed.

A third interesting area for further research lies in completing the framework for stylistic features by feature engineering all the seven stylistic categories to a detectable form. Operationalising of the various stylistic categories and stylistic features could be approached both from the user perspective and from the technical perspective. One could for example research further what characterises an engaging podcast from the users' perspective and then attempt to engineer a working machine learning model which could automatically detect the engagement level of podcast from the podcast's audio.

In general, we believe that our work constitutes a promising step towards a comprehensive set of stylistic podcast categories, stylistic podcast features, and their underlying acoustic features. The results of our research could be used to empower users to discover stylistically more relevant podcasts. A lot of the required technology for detecting stylistic features listeners care about already exists. Only implementation and testing in an application context is needed to take the user experience of podcast search and recommendation to a new (stylistic) level.

Bibliography

[1] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 2009. DOI: 10.1017/CBO9781139644082.

[2] *37 mind blowing youtube facts, figures and statistics – 2019*. [Online]. Available: https://biographon.com/youtube-stats/.

[3] J. Besser, K. Hofmann, and M. Larson, "An exploratory study of user goals and strategies in podcast search.", in *LWA*, 2008, pp. 27–34.

[4] N. Newman and N. Gallo, "News podcasts and the opportunities for publishers", 2019.

[5] L. Yang, Y. Wang, D. Dunne, M. Sobolev, M. Naaman, and D. Estrin, "More than just words: Modeling non-textual characteristics of podcasts", in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, ser. WSDM '19, Melbourne VIC, Australia: Association for Computing Machinery, 2019, pp. 276–284, ISBN: 9781450359405. DOI: 10.1145/3289600.3290993. [Online]. Available: https://doi.org/10.1145/3289600.3290993.

[6] J. Valinsky, *Spotify bet on podcasts. it's working*, Oct. 2019. [Online]. Available: https://edition.cnn.com/2019/10/28/tech/spotify-third-quarter-earnings/index.html.

[7] A. Carman, *Google will start surfacing individual podcast episodes in search results*, Aug. 2019. [Online]. Available: https://www.theverge.com/2019/8/8/20759394/google-podcast-episodes-search-results-transcriber.

[8] P. Nowak, *What is the average reading speed?*, May 2018. [Online]. Available: https://www.irisreading.com/what-is-the-average-reading-speed/.

[9] V. de Boer, R. J. F. Ordelman, and J. Schuurman, "Evaluating unsupervised thesaurus-based labeling of audiovisual content in an archive production environment", *International Journal on Digital Libraries*, vol. 17, no. 3, pp. 189–201, Sep. 2016, ISSN: 1432-1300. DOI: 10.1007/s00799-016-0182-6. [Online]. Available: https://doi.org/10.1007/s00799-016-0182-6.

[10] M.-Y. Chung, "Podcast use motivations and patterns among college students", PhD thesis, Kansas State University, 2008.

[11] M. Larson and G. J. Jones, "Spoken content retrieval: A survey of techniques and technologies", *Foundations and Trends in Information Retrieval*, vol. 5, no. 4–5, pp. 235–422, 2012.

[12] D. W. Oard *et al.*, "Overview of the clef-2006 cross-language speech retrieval track", in *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer, 2006, pp. 744–758.

[13] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, 2008.

[14] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2009.

[15] L. Wilcox and M. A. Bush, "Hmm-based wordspotting for voice editing and indexing", in *EUROSPEECH*, 1991.

[16] J. Laver, "The phonetic description of voice quality", English, *Cambridge Studies in Linguistics London*, 1980.

[17] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods", *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, Oct. 2015, ISSN: 2371-9850. DOI: 10.1109/TAFFC.2015.2432791.

[18] N. Chhaya, K. Jaidka, and L. H. Ungar, "The AAAI-18 workshop on affective content analysis.", in *AAAI Workshops*, 2018, pp. 2–7.

[19] M. Maybury, *Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance and Authoring*. Wiley, 2012, ISBN: 9781118219522. [Online]. Available: https://books.google.se/books?id=sBMXidepUFkC.

[20] P. Ekman, "Basic emotions", in *Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd, 2005, ch. 3, pp. 45–60, ISBN: 9780470013496. DOI: 10.1002/0470013494.ch3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013494.ch3. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch3.

[21] H.-B. Kang, "Affective content detection using hmms", in *Proceedings of the Eleventh ACM International Conference on Multimedia*, ser. MULTIMEDIA '03, Berkeley, CA, USA: ACM, 2003, pp. 259–262, ISBN: 1-58113-722-2. DOI: 10.1145/957013.957066. [Online]. Available: http://doi.acm.org/10.1145/957013.957066.

[22] H.-B. Kang, "Emotional event detection using relevance feedback", in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, IEEE, vol. 1, Jan. 2003, pp. 721–724.

[23] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models", in *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ser. ACII '07, Lisbon, Portugal: Springer-Verlag, 2007, pp. 594–605, ISBN: 978-3-540-74888-5. DOI: 10.1007/978-3-540-74889-2_52. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74889-2_52.

[24] R. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features: A dimensional and categorial experimental approach", *Multimedia Tools and Applications - MTA*, vol. 61, pp. 1–29, Nov. 2011. DOI: 10.1007/s11042-010-0702-0.

[25] M. Xu, X. He, J. S. Jin, Y. Peng, C. Xu, and W. Guo, "Using scripts for affective content retrieval", in *Proceedings of the Advances in Multimedia Information Processing, and 11th Pacific Rim Conference on Multimedia: Part II*, ser. PCM'10, Shanghai, China: Springer-Verlag, 2010, pp. 43–51, ISBN: 3-642-15695-9, 978-3-642-15695-3. [Online]. Available: http://dl.acm.org/citation.cfm?id=1894049.1894055.

[26] X. Y. Chen and Z. Segall, "Xv-pod: An emotion aware, affective mobile video player", in *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 3, Mar. 2009, pp. 277–281. DOI: 10.1109/CSIE.2009.982.

[27] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification", *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010, ISSN: 1941-0077. DOI: 10.1109/TMM.2010.2051871.

[28] H. L. Wang and L.-F. Cheong, "Affective understanding in film", *IEEE Transactions on circuits and systems for video technology*, vol. 16, no. 6, pp. 689–704, 2006.

[29] K.-M. Ong and W. Kameyama, "Classification of video shots based on human affect", *The Journal of The Institute of Image Information and Television Engineers*, vol. 63, pp. 847–856, Jan. 2009. DOI: 10.3169/itej.63.847.

[30] A. Yazdani, J.-S. Lee, and T. Ebrahimi, "Implicit emotional tagging of multimedia using eeg signals and brain computer interface", in *Proceedings of the First SIGMM Workshop on Social Media*, ser. WSM '09, Beijing, China: ACM, 2009, pp. 81–88, ISBN: 978-1-60558-759-2. DOI: 10.1145/1631144.1631160. [Online]. Available: http://doi.acm.org/10.1145/1631144.1631160.

[31] P. Ekman and D. Cordaro, "What is meant by calling emotions basic", *Emotion review*, vol. 3, no. 4, pp. 364–370, 2011.

[32] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[33] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament", *Current Psychology*, vol. 14, no. 4, pp. 261–292, Dec. 1996, ISSN: 1936-4733. DOI: 10.1007/BF02686918. [Online]. Available: https://doi.org/10.1007/BF02686918.

[34] K. R. Scherer, "The dynamic architecture of emotion: Evidence for the component process model", *Cognition and emotion*, vol. 23, no. 7, pp. 1307–1351, 2009.

[35] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position", in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[36] R. E. Thayer, *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[37] J. A. Russell, "A circumplex model of affect.", *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[38] D. Knox, G. Cassidy, S. Beveridge, and R. MacDonald, "Music emotion classification by audio signal analysis: Analysis of self-selected music during game play", in *Proceedings of the 10th International Conference on Music Perception and Cognition*, 2008, pp. 25–29.

[39] C.-H. Yeh, H.-H. Lin, and H. Chang, "An efficient emotion detection scheme for popular music.", May 2009, pp. 1799–1802. DOI: 10.1109/ISCAS.2009.5118126.

[40] U. Glavitsch and P. Schäuble, "A system for retrieving speech documents", in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '92, Copenhagen, Denmark: Association for Computing Machinery, 1992, pp. 168–176, ISBN: 0897915232. DOI: 10.1145/133160.133194. [Online]. Available: https://doi.org/10.1145/133160.133194.

[41] D. James, "The application of classical information retrieval techniques to spoken documents", Aug. 1995.

[42] E. Hauptmann and M. Witbrock, "Informedia: News-on-demand multimedia information acquisition and retrieval", May 1998.

[43] S. Whittaker *et al.*, "SCANmail: A voicemail interface that makes speech browsable, readable and searchable", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2002, pp. 275–282.

[44] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal, "SCAN: Designing and evaluating user interfaces to support retrieval from speech archives.", Jan. 1999, pp. 26–33. DOI: 10.1145/312624.312639.

[45] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval", in *Proceedings of the Fourth ACM International Conference on Multimedia*, ser. MULTIMEDIA '96, Boston, Massachusetts, USA: Association for Computing Machinery, 1997, pp. 307–316, ISBN: 0897918711. DOI: 10.1145/244130.244232. [Online]. Available: https://doi.org/10.1145/244130.244232.

[46] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources", in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '96, Zurich, Switzerland: Association for Computing Machinery, 1996, pp. 30–38, ISBN: 0897917928. DOI: 10.1145/243199.243208. [Online]. Available: https://doi.org/10.1145/243199.243208.

[47] H. Bourlard and S. Renals, "Recognition and understanding of meetings overview of the european ami and amida projects", Jan. 2008.

[48] F. Kubala, S. Colbath, D. Liu, and J. Makhoul, "Rough'n'ready: A meeting recorder and browser", *ACM Comput. Surv.*, vol. 31, no. 2es, 7–es, Jun. 1999, ISSN: 0360-0300. DOI: 10.1145/323216.323354. [Online]. Available: https://doi.org/10.1145/323216.323354.

[49] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations", Jan. 2006, pp. 51–58. DOI: 10.1145/1148170.1148183.

[50] W. Byrne *et al.*, "Automatic recognition of spontaneous speech for access to multilingual oral history archives", *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 420–435, 2004.

[51] F. de Jong, D. Oard, W. Heeren, and R. Ordelman, "Access to recorded interviews: A research agenda", *JOCCH*, vol. 1, Jun. 2008. DOI: 10.1145/1367080.1367083.

[52] J. H. Hansen *et al.*, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word", *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 712–730, Oct. 2005. DOI: 10.1109/TSA.2005.852088.

[53] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the mit spoken lecture processing project", in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[54] C. Alberti *et al.*, "An audio indexing system for election video material", in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2009, pp. 4873–4876.

[55] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content", *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, 2008.

[56] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method", in *Proceedings of the ninth ACM international conference on Multimedia*, 2001, pp. 203–211.

[57] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification", *IEEE Transactions on speech and audio processing*, vol. 9, no. 4, pp. 441–457, 2001.

[58] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal", in *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, 2008, pp. 1–7.

[59] M. Radmard, M. Hadavi, and M. M. Nayebi, "A new method of voiced/unvoiced classification based on clustering", *Journal of Signal and Information Processing*, vol. 2, no. 04, p. 336, 2011.

[60] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states", *Speech evaluation in psychiatry*, pp. 221–240, 1981.

[61] R. W. Picard, *Affective computing*. MIT press, 2000.

[62] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling", *IEEE transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[63] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation", in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[64] B. Wrede and E. Shriberg, "Spotting"hot spots"in meetings: Human judgments and prosodic cues", in *Eighth European Conference on Speech Communication and Technology*, 2003.

[65] K. Laskowski, "Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings", in *2008 IEEE Spoken Language Technology Workshop*, IEEE, 2008, pp. 81–84.

[66] H. Bořil, A. Sangwan, T. Hasan, and J. H. Hansen, "Automatic excitement-level detection for sports highlights generation", in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[67] W. Hürst, M. Welte, and S. Jung, "An evaluation of the mobile usage of e-lecture podcasts", in *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*, 2007, pp. 16–23.

[68] C. Evans, "The effectiveness of m-learning in the form of podcast revision lectures in higher education", *Computers & education*, vol. 50, no. 2, pp. 491–498, 2008.

[69] D. McKinney, J. L. Dyck, and E. S. Luber, "Itunes university and the classroom: Can podcasts replace professors?", *Computers & education*, vol. 52, no. 3, pp. 617–623, 2009.

[70] J. Dewe, J. Karlgren, and I. Bretan, "Assembling a balanced corpus from the internet", in *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, 1998, pp. 100–108.

[71] J. Besser, "Incorporating user search goal analysis in podcast retrieval optimization", *Master's thesis. Saarland University, Germany*, 2008.

[72] M. Tsagkias, M. Larson, and M. d. Rijke, "Predicting podcast preference: An analysis framework and its application", *Journal of the American Society for Information Science and Technology*, vol. 61, no. 2, pp. 374–391, 2010, ISSN: 1532-2890. DOI: 10.1002/asi.21259.

[73] N. Hariri, B. Mobasher, and R. Burke, "Context-aware music recommendation based on latenttopic sequential patterns", in *Proceedings of the sixth ACM conference on Recommender systems*, 2012, pp. 131–138.

[74] M. Schedl, P. Knees, and F. Gouyon, "New paths in music recommender systems research", in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 392–393.

[75] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling.", in *Ismir*, vol. 270, 2000, pp. 1–11.

[76] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge", in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[77] B. Schuller *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism", in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[78] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016. [Online]. Available: http://www.deeplearningbook.org/.

[79] I. Goodfellow *et al.*, "Generative adversarial nets", in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[80] P. M. Asaro, *Participatory design*, May 2020. [Online]. Available: https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/participatory-design (visited on 07/09/2020).

[81] *Definition of monotonous*, 2020. [Online]. Available: https://www.merriam-webster.com/dictionary/monotonous (visited on 07/09/2020).

[82] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking", *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[83] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically", *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[84] K. Lee and D. P. W. Ellis, "Detecting music in ambient audio by long-window autocorrelation", in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 9–12.

[85] I. Otsuka and H. Suginohara, *Method and device for detecting music segment, and method and device for recording data*, US Patent 8,855,796, Oct. 2014.

[86] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis", *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[87] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic music transcription and audio source separation", *Cybernetics and Systems*, vol. 33, no. 6, pp. 603–627, 2002. DOI: 10.1080/01969720290040777. eprint: https://doi.org/10.1080/01969720290040777. [Online]. Available: https://doi.org/10.1080/01969720290040777.

[88] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality", in *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, IEEE, 2018.

[89] E. Keller, "The analysis of voice quality in speech processing", in *International School on Neural Networks, Initiated by IIASS and EMFCSC*, Springer, 2004, pp. 54–73.

[90] K. P. Truong and D. A. v. Leeuwen, "Automatic detection of laughter", in *Ninth European Conference on Speech Communication and Technology*, 2005.

[91] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection", in *International Workshop on Machine Learning for Multimodal Interaction*, Springer, 2008, pp. 137–148.

[92] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks", in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[93] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech", in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 5117–5120.

[94] L. S. Kennedy and D. P. Ellis, "Laughter detection in meetings", 2004.

[95] O. Peleg and M. Wasserblat, *Method and system for laughter detection*, US Patent 8,571,853, Oct. 2013.

[96] A. Clifton *et al.*, "The spotify podcasts dataset", *arXiv preprint arXiv:2004.04270*, 2020.

Appendix A

User Study Detailed Agenda

Understanding what stylic characteristics of podcasts do podcast listeners find important for their podcast listening experience

Workshop Agenda and Discussion Guide | March 2020 | @katariiinam

RESEARCH GOALS & RESEARCH QUESTIONS

The goal of this research is to understand what stylistic characteristics of podcasts do listeners find interesting or important for their podcast listening experience. The goal is approached by answering the following questions.

1. What stylistic features can listeners observe and verbalise in podcasts? Both in terms of what features they like and do not like in podcasts.
2. What are listeners' perceptions on what kind of stylistic categories exist among podcasts based on the stylistic features they care about?
3. How well do stylistic podcast features identified from existing literature and the stylistic features suggested by us fit to listeners' perceptions on stylistic podcast categories and their perceptions on which stylistic features are important for their podcast listening experience?

METHOD

Approach: Participatory design workshops

Location: Online via Google Hangouts
Total: 9 people + 1 stand - in
Time: about 2h
Type: Podcast listeners (listen to p podcasts at least once a month)

Recruiting Criteria

- Podcast listeners: Listens to podcasts minimum 1x a month on Spotify
- Fluent English skills - able to express themselves in English without limitations
- Good IT skills and access to a computer in order to participate in the online workshop

AGENDA SUMMARY

10min

Welcome
Ice breaker

5min

Intro to workshop, context

35min

START NEW REC
Working on the research questions 1
10min writing post its
10min discussion
10min writing post its
5min discussion

10min

Break

25min

START NEW REC
Research question 2
10min grouping
15min discussion & regrouping

25min

Research question 3
10min distribute new notes & make new groups
15min discussion and regrouping

10min

Wrap up & Thank you

30min

Writing down my own reflections about the methods, the group dynamics, the results etc

DISCUSSION GUIDE

Welcome

Intro to workshop, context and tasks
Welcome to this workshop and thank you so much for coming.
The workshop will take about 2h. This workshop will be about podcasts and how they are. I will tell you more soon. But first I need to ask you to sign **a non-disclosure agreement.** There is one copy for yourself and one for me.

Ok so lets start the session. First I want to do a little ice breaker and hear who I have here today.
Ice breaker
Tell us one thing you like about staying at home. Let's do a round. I can start.

Yey this is great! Lets now move to the exercises. So this is intented to be a relaxed, fun and creative workshop so no pressure at all. You don't need to perform in any way and your are not being evaluated in any way.
Please remember that I am truly interested in your personal opinions. There are no right or wrong here and everyone's opinions are equally valuable. If you have questions at any time please feel free to ask them.

I will start the recording now.
Please go to the Mural board I shared with you, also open the Google Drive folder I linked to the Mural board.

Working on the research questions 1
You have different podcasts in the Google drive folder. Please take 10min to listen to them by yourself. You won't have time to listen to them properly. So instead use the browse bar and skim through the podcasts as you please. But try to have time to listen to each podcast at least a little bit.

While you listen to the podcasts, please think what you like or don't like about them **other than the content or the topic talked about** and write them down on the post its on the Mural board. Each participant has been given a number. Go for the circle with your number in it. Make sure to write only one thing to one post it. More different things you can write down the better. Brainstorm, be creative! There is no right or wrong here! After the 10min we will go through the post its.

After 10min - Now lets go around and just read aloud what post its we have.
Here I can point out some examples if the post it has too high level category e.g. funny and ask what makes it funny. Ask to break it down. Why would you think that a podcast is funny?

Now lets do a second round with the same podcast snippets we just heard and try to write down more post its. After the 10min we will go through the post its.

After 10min - Now lets go around and just read aloud what post its we have.

Break

Research question 2

Can you now look at the post-its and try to group them to groups which make sense. Name each of the groups. Also give names to the scales you see. In the end I am looking for something like this: *show the grouping example image in Mural. Explain the line: does not exist in the podcast -- exists a lot in the podcast.* Discuss with each other how the categories could be formed. If you feel like some post it would belong to two or more different groups feel free to write it again on another post it and use that in the other group. Also if you come up with new things to write down feel free to add them. Also, if you feel the need to go and re-listen to the podcasts we went through earlier for this exercise, feel free to.

Now you have groups of these post its. Have a look of them and see if you want to move something. Put a black post it on it. We will then discuss together how we could reorganise the post its with the black post its on them.

Research question 3

Great. Now you have these group of post its with labels on them. I have now a bunch of new post its for you (should be a different colour). Please take them one by one and put them to existing groups. If you feel like they don't fit anywhere put them to the black square on the side.

Ok. Now lets go through the post its which did not fit anywhere. Could you make a new group out of these. Discuss and make it if it works out. If you think that some of these are completely irrelevant please put them to the black square.

Now let's look at everything. Do you agree or would you like to change something?

Wrap up & Thank you
Thank you so much for participating. This was very valuable for me!

Goodbye
Writing down my reflections, notes and thoughts about the workshop

Appendix B

Instructions for the Workshop Exercises

1. “Please go to the Google Drive folder where the podcast episodes are. You will not have enough time to listen to the podcasts properly. The episodes are on average 1h long. Instead skim through the podcasts. You can start at any part of the podcast and skip around the podcast in any way you want. While you listen to the podcasts think what you like or do not like in the podcasts other than the content or topic discussed about? While you listen and notice these things write one such thing per post-it. Write the file name you made the observation from e.g. 22.wav on the post-it as well. Write the things you like on pink post-its and the ones you do not like on blue post-its. Write the post-its around your participant circle on the board so that we can keep track of who wrote what. ”

2. “How would you group the post-its from the 1 exercise to the below lines and assign a label/category for each group? How would you name the scale for the group? You can create as many groups as you see fit. Discuss together how you would make the groupings. ”

3. “How would you distribute these yellow notes to the groups from 2 exercise? Do they fit to existing groups? Do they form new groups? Are some of them redundant? If you think that they are not relevant for your podcast listening experience, please move them to the “trash” square below.”

Appendix C

Feature Matrix for Podcast Sample Selection

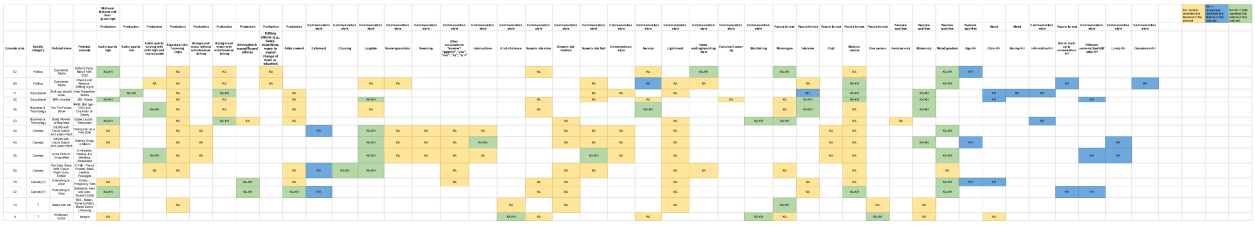

Figure C.1: Feature matrix used to understand feature distribution of the podcasts used in the workshops

Appendix D

Participants' Detailed Background

Prior to the workshop the participants were asked to fill a form in order to answer to some questions related to the participant's background and podcast listening habits. The answers were recorded anonymously. The participants' answers are compiled to the tables D.1 and D.2.

Nationality	Germany (n=3) Finland (n=3) India (n=1) Latvia (n=1) Spain (n=1)
Age & Gender	7 females 2 males 20-29 years old
Profession / Study field	Dental care (n=1) Electrical Engineering (n=1) Interaction Technology (n=2) Law (n=1) Corporate communication and event organization (n=1) Economist (n=1) Human Computer Interaction (n=1) Medicine (n=1)

Table D.1: Participants' background

Frequency of listening to podcasts	Daily Once every week more or less 3-7 times a week It varies depending on how busy is my schedule. At least every month, sometimes multiple times per week Atleast once a week 1/2 weeks 1-6 times in a month 3x/month 1-2 a month
Type of podcasts the participants usually listen to	Science, News, Politics, Economics, Current affairs Documentaries, True Crime Humour and gossip podcasts Football and General Society and culture Advice shows, podcasts about current events/history/culture, entertainment podcasts (e.g. celebrity gossip, funny stories) Lifestyle podcasts related to meditation, philosophy and other ones related to social aspects that make you learn new things
Platforms/channels the participants use for listening to podcasts	Spotify (n=8) YouTube (n=3) Google Podcasts (n=1) Yle Areena (n=1)
Time of listening to podcasts	In the morning on the commute, during chores Morning/evening During commuting hours (morning + before lunch) Afternoon, evening Evenings or my days off work. Sometimes in the middle of a night, if I have trouble sleeping. Daytime - when they are doing household chores No specific time During the day Varies. Rarely mornings
Where the participants listen to podcasts	No specific place At home, on walks, on public transport In the public transport In my kitchen doing dishes/while cycling At home Gym, in the shower, on the bus Home or when traveling. Used to do it at Gym Outside on a walk/run

Table D.2: Participants' podcast listening habits

Appendix E

Plots of Feature Value Distributions

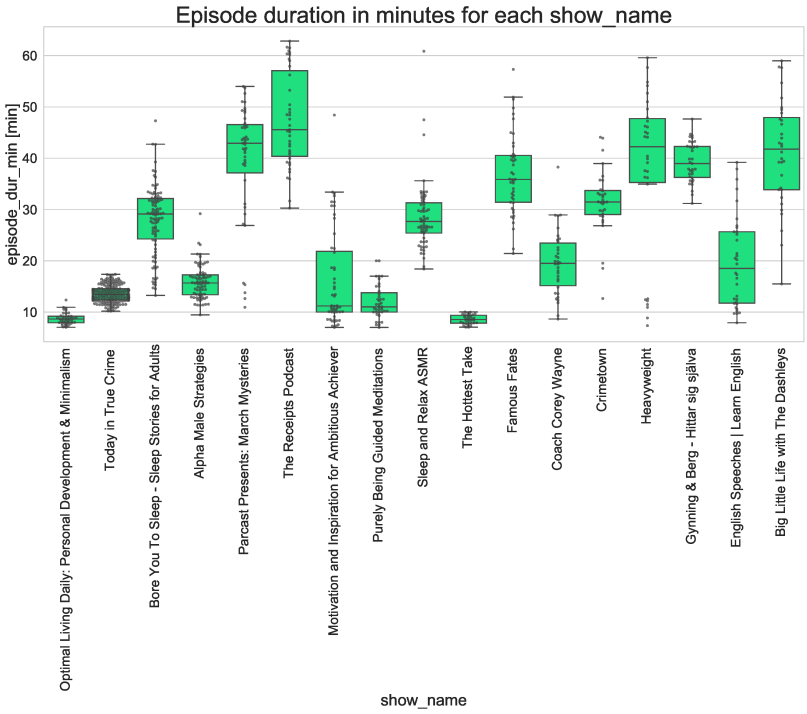

Figure E.1: Episode duration in minutes grouped by shows.

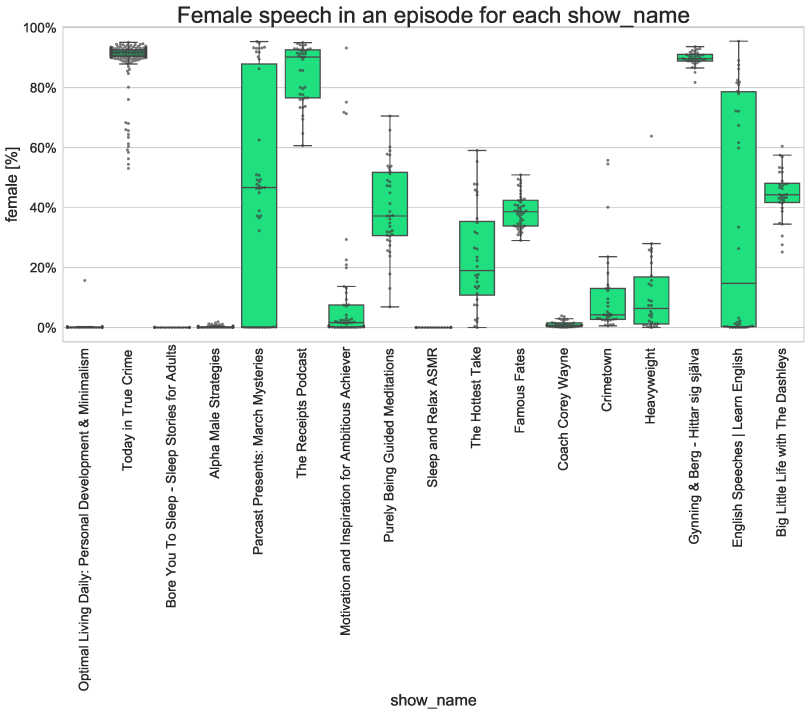

Figure E.2: Percentage of female voices grouped by shows.

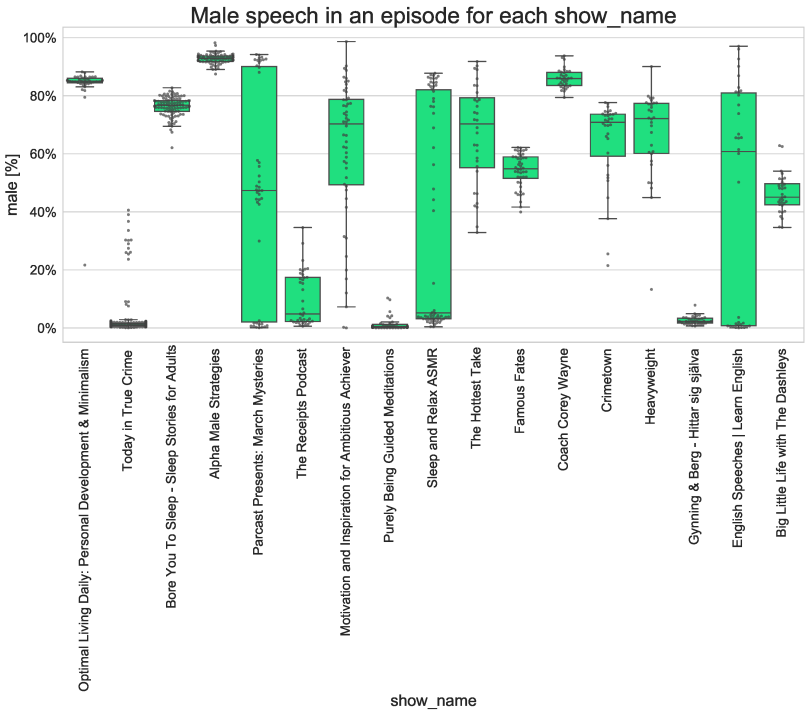

Figure E.3: Percentage of male voices grouped by shows.

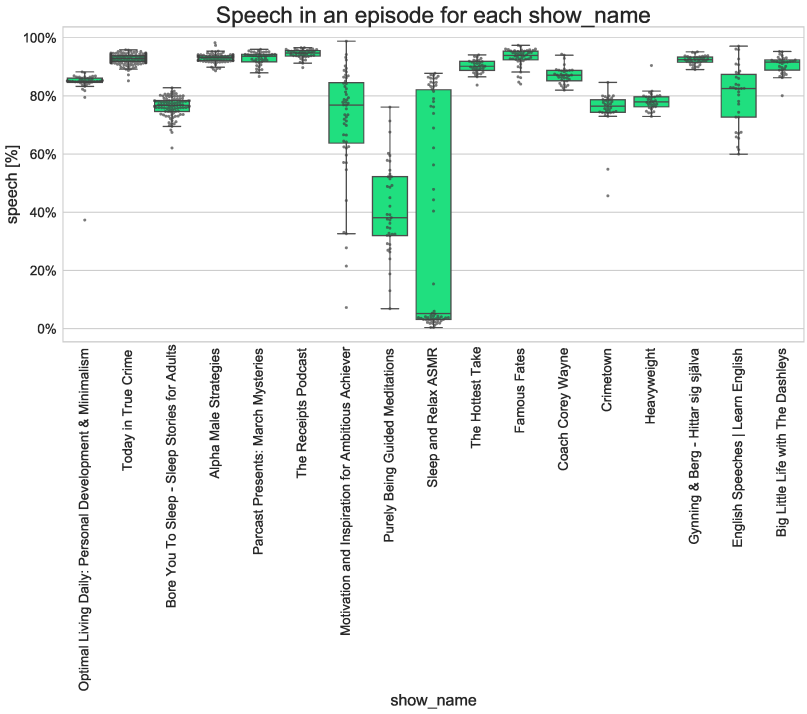

Figure E.4: Percentage of speech grouped by shows.

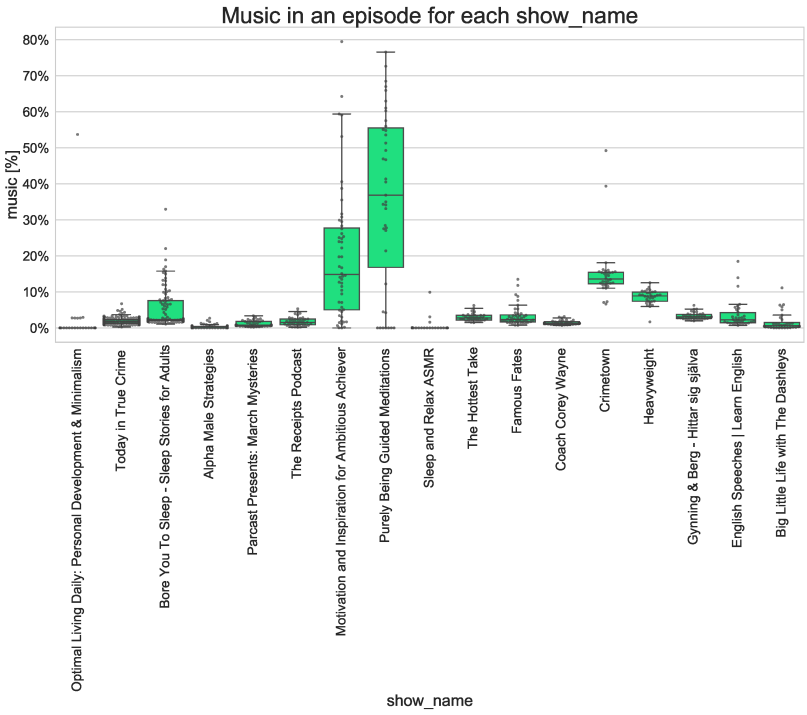

Figure E.5: Percentage of music voices grouped by shows.

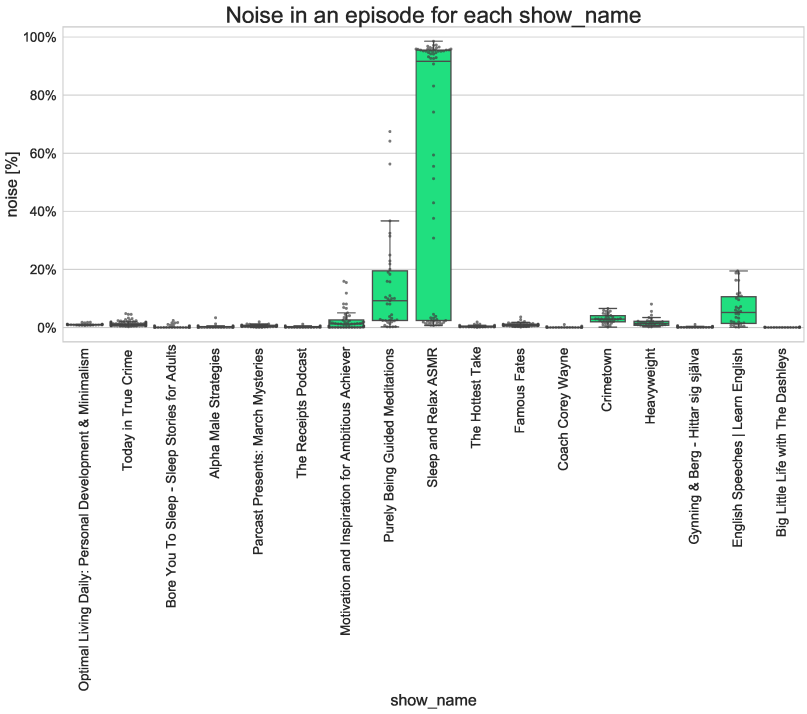

Figure E.6: Percentage of noise voices grouped by shows.

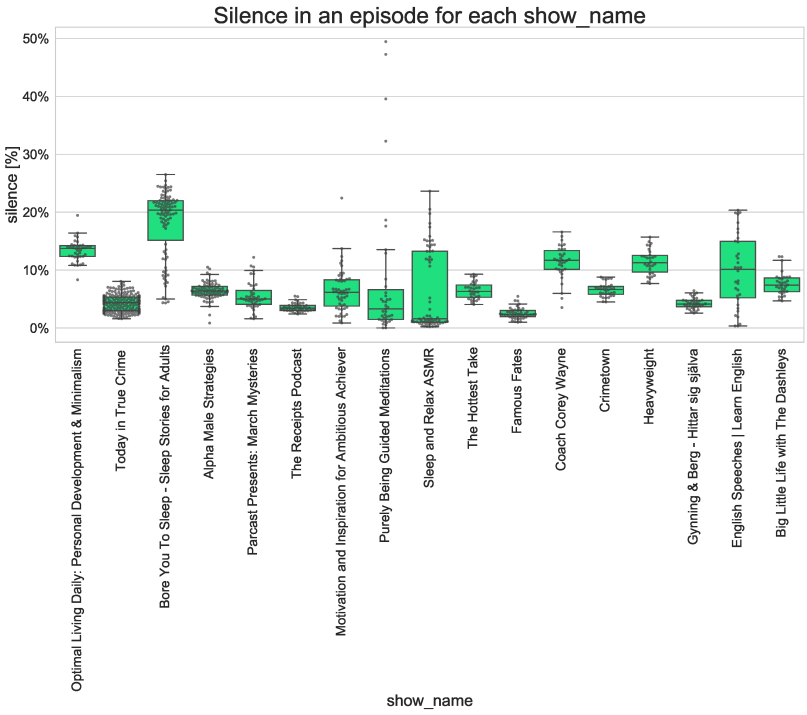

Figure E.7: Percentage of silence voices grouped by shows.

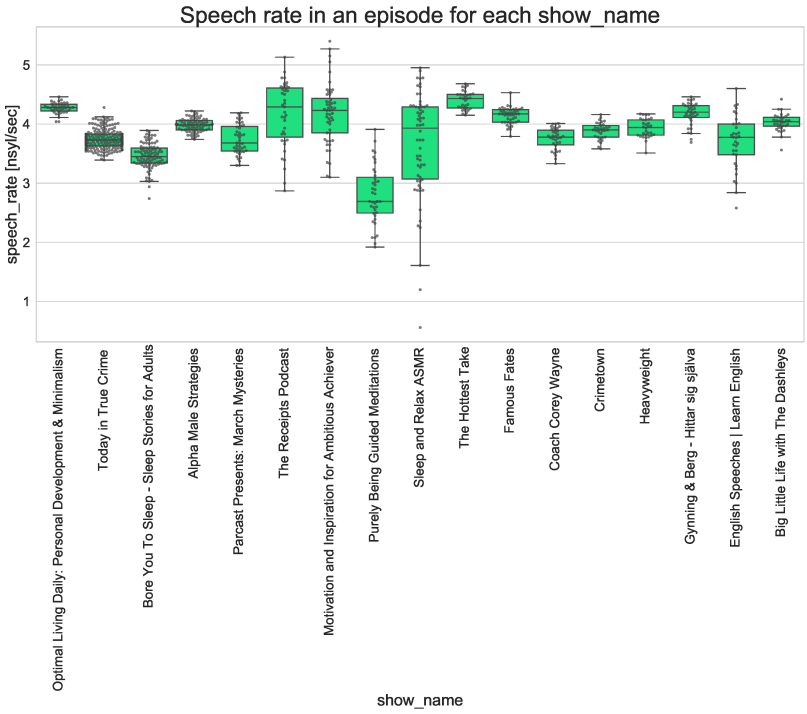

Figure E.8: Speech rate of episodes grouped by shows.

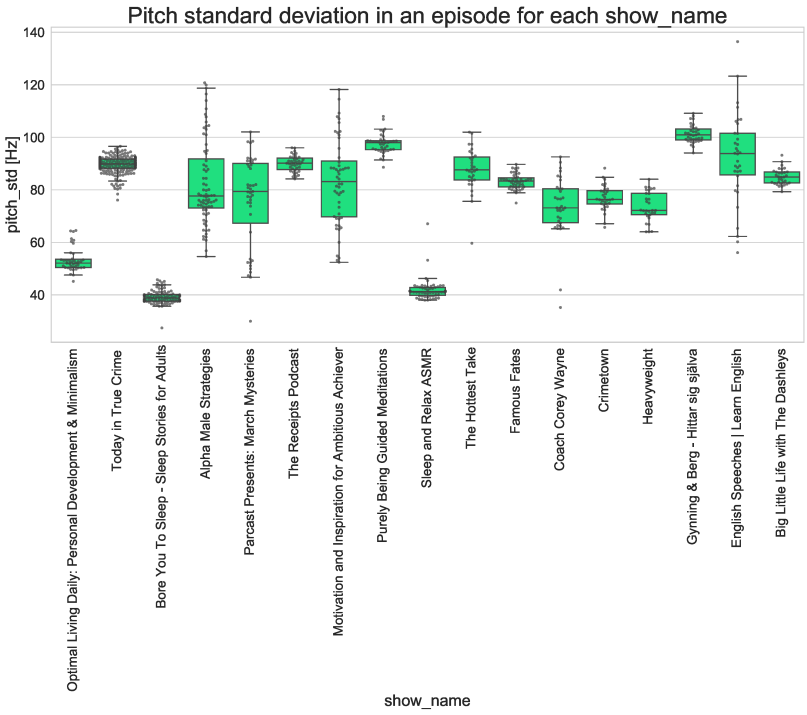

Figure E.9: Monotonousness of episodes grouped by shows.

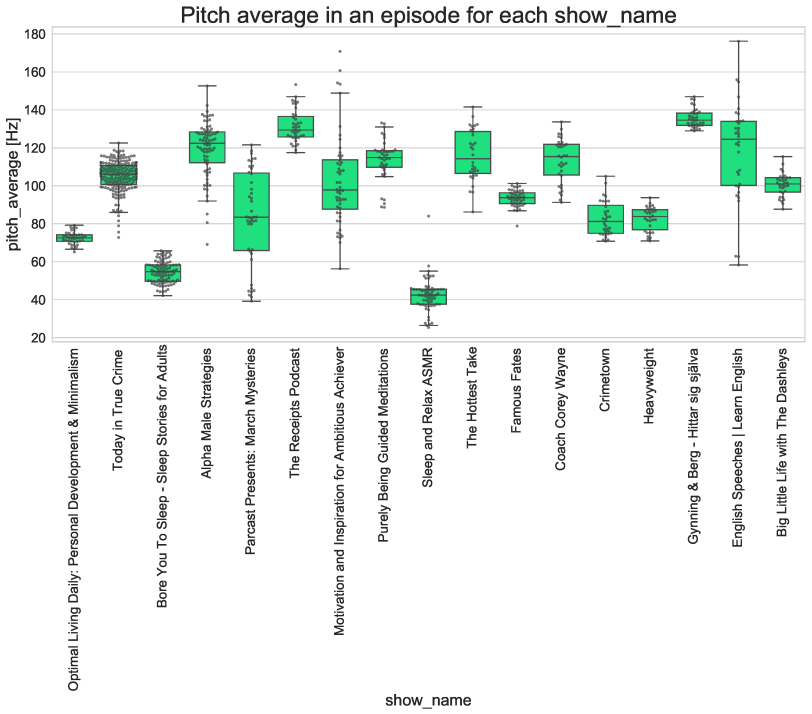

Figure E.10: Pitch average of episodes grouped by shows.

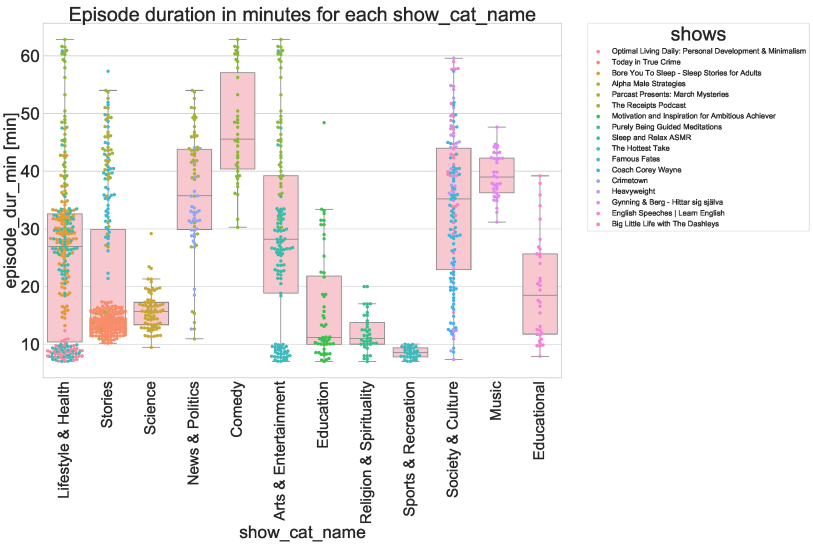

Figure E.11: Episode duration in minutes grouped by genres.

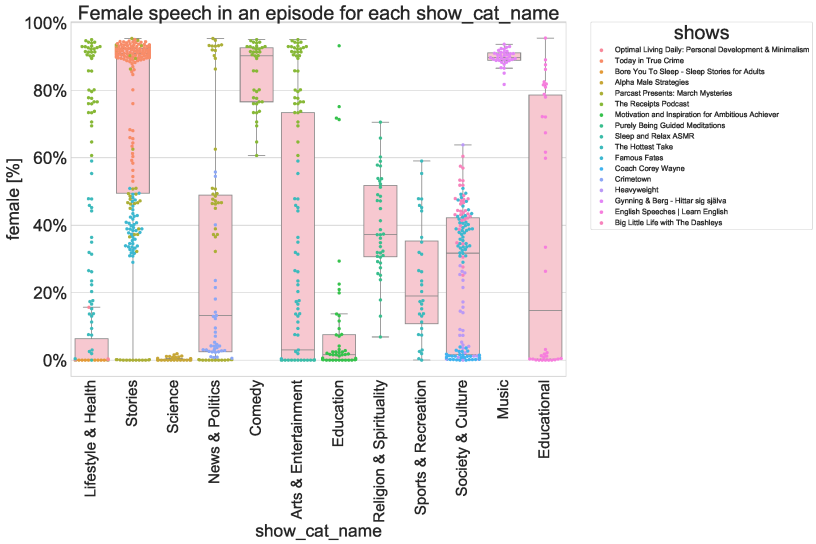

Figure E.12: Percentage of female voices grouped by genres.

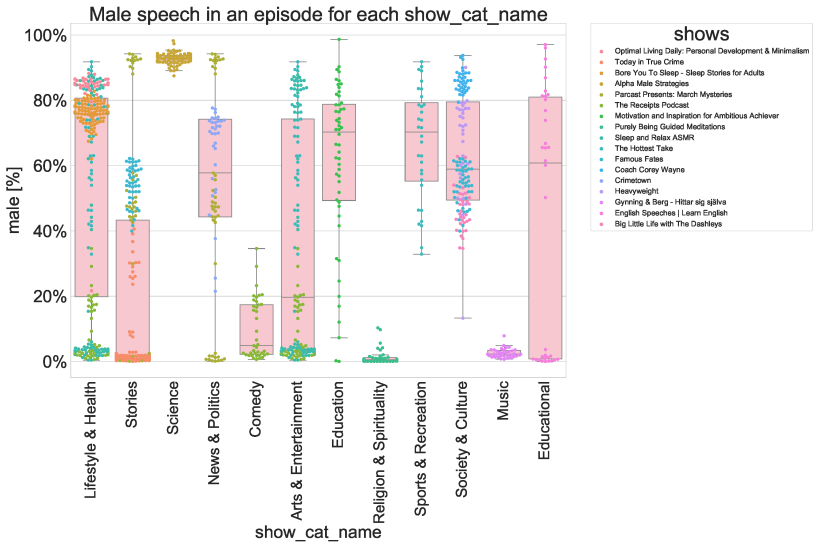

Figure E.13: Percentage of male voices grouped by genres.

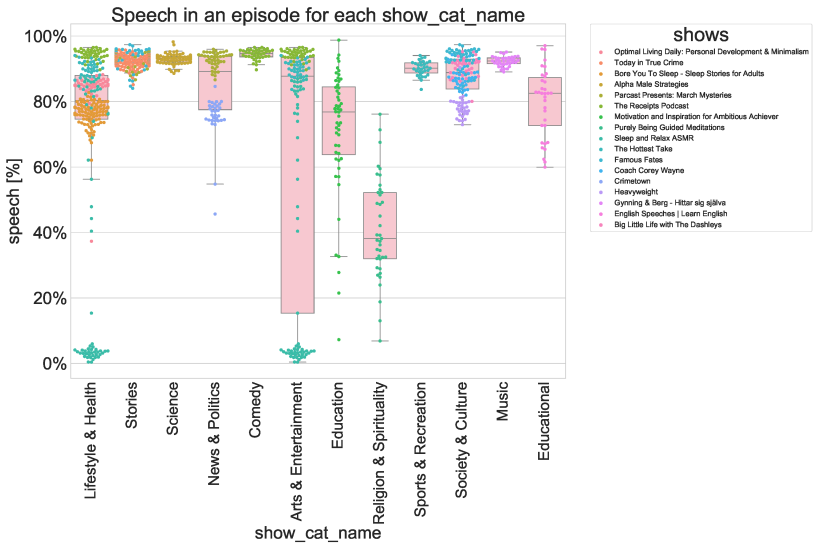

Figure E.14: Percentage of speech grouped by genres.

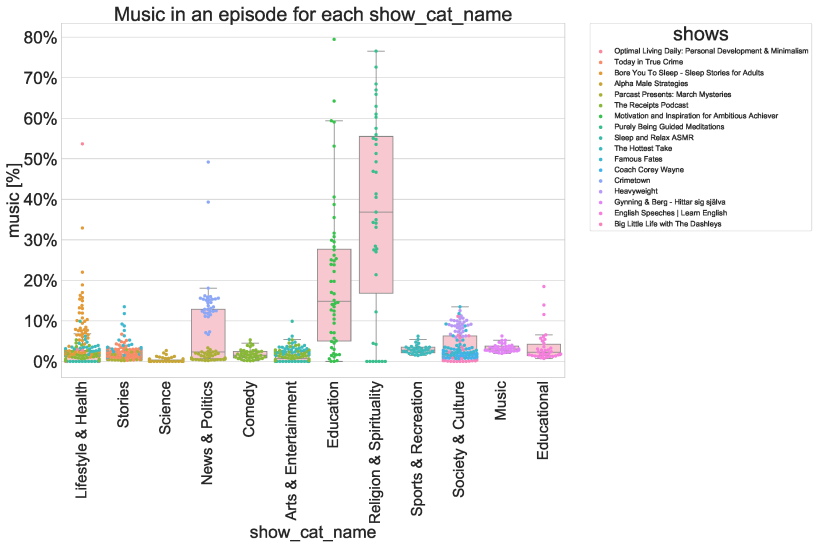

Figure E.15: Percentage of music voices grouped by genres.

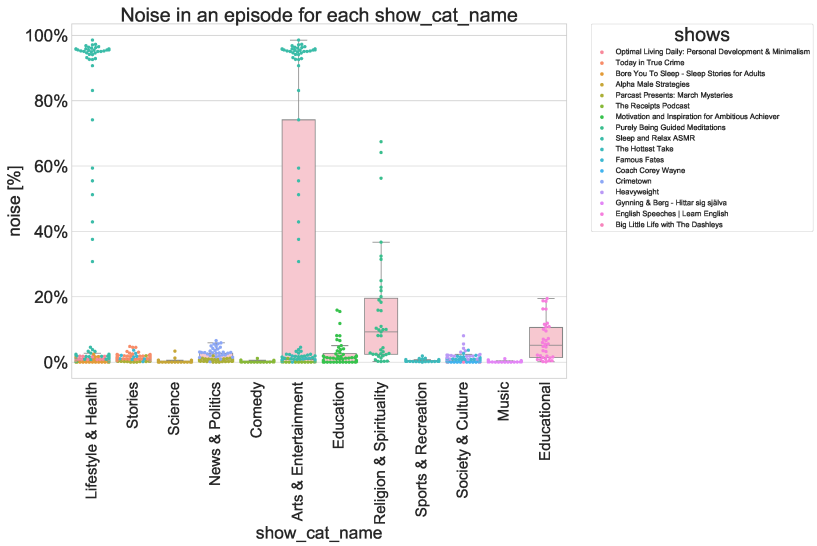

Figure E.16: Percentage of noise voices grouped by genres.

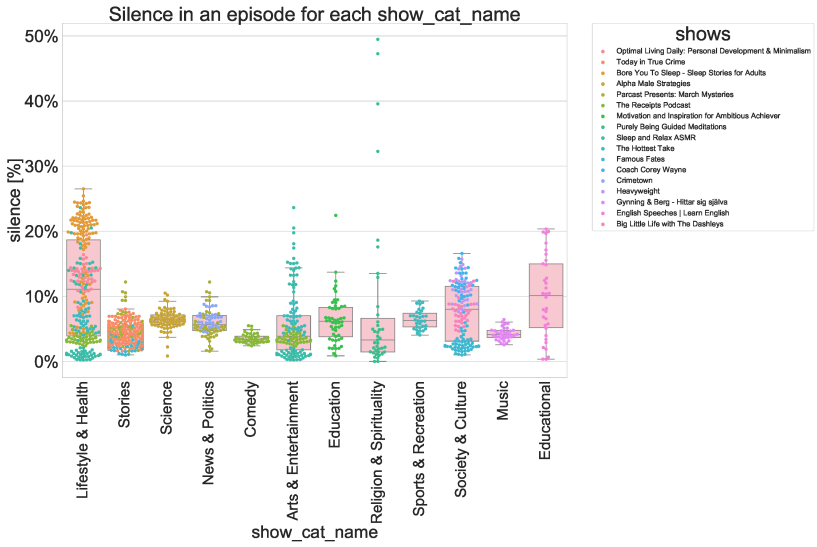

Figure E.17: Percentage of silence voices grouped by genres.

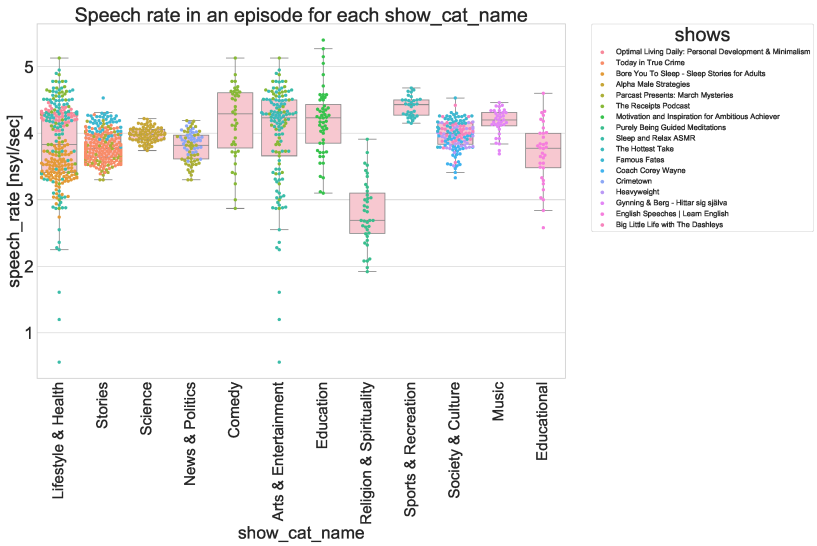

Figure E.18: Speech rate of episodes grouped by genres.

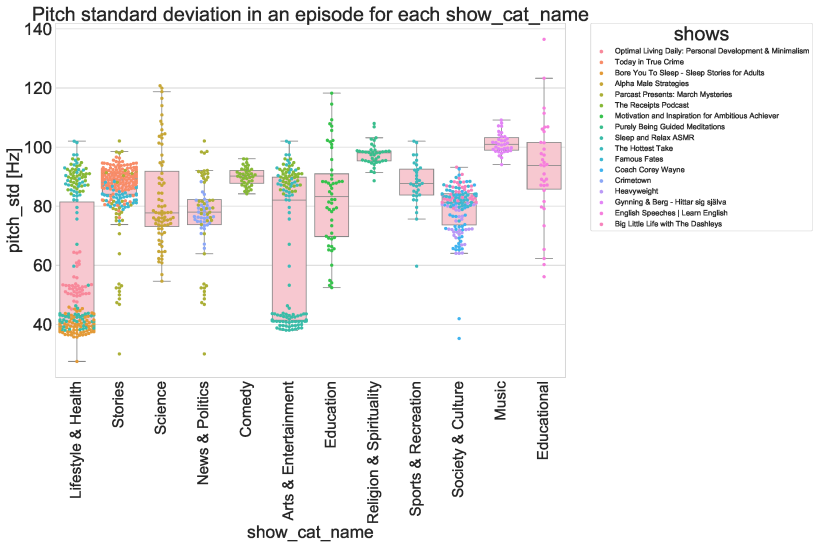

Figure E.19: Monotonousness of episodes grouped by genres.

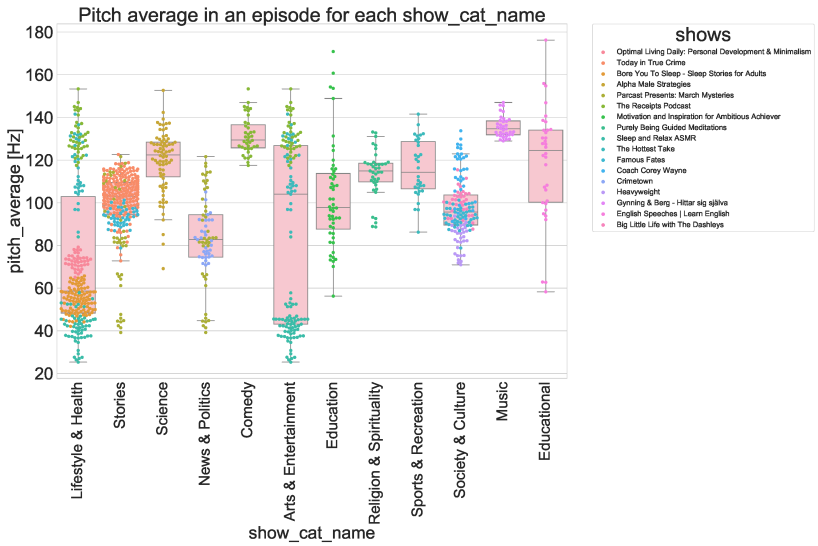

Figure E.20: Pitch average of episodes grouped by genres.

Appendix F

Kruskal Wallis and Welch ANOVA results for Podcast Shows

Test Statistics[a,b]

	music	pitch average	speech rate	pitch standard deviation	speech
Kruskal-Wallis H	168.337	214.819	213.476	204.596	121.287
df	5	5	5	5	5
Asymp. Sig.	.000	.000	.000	.000	.000

Test Statistics[a,b]

	male speech	noise	female speech	silence
Kruskal-Wallis H	237.641	154.585	234.583	168.184
df	5	5	5	5
Asymp. Sig.	.000	.000	.000	.000

a. Kruskal Wallis Test

b. Grouping Variable: show name

Robust Tests of Equality of Means

		Statistic[a]	df1	df2	Sig.
music	Welch	87.621	5	93.287	.000
noise	Welch	59.013	5	105.699	.000
silence	Welch	108.292	5	99.837	.000
speech	Welch	119.690	5	91.590	.000
female speech	Welch	849.468	5	101.684	.000
male speech	Welch	807.896	5	100.919	.000
pitch average	Welch	408.315	5	101.685	.000
pitch standard deviation	Welch	170.857	5	92.559	.000
speech rate	Welch	145.052	5	97.876	.000

a. Asymptotically F distributed.

Appendix G

Post Hoc Tukey results for Podcast Shows

noise

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05			
		1	2	3	4
Gynning & Berg - Hittar sig själva	35	.0017			
The Hottest Take	32	.0044			
Parcast Presents March Mysteries	43	.0057	.0057		
Famous Fates	44		.0099	.0099	
Today in True Crime	194			.0112	
Heavyweight	30				.0186
Sig.		.113	.085	.956	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

silence

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05			
		1	2	3	4
Famous Fates	44	.0254			
Gynning & Berg - Hittar sig själva	35		.0424		
Today in True Crime	194		.0430		
Parcast Presents March Mysteries	43			.0553	
The Hottest Take	32			.0647	
Heavyweight	30				.1144
Sig.		1.000	1.000	.078	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

speech

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05		
		1	2	3
Heavyweight	30	.7813		
The Hottest Take	32		.9011	
Gynning & Berg - Hittar sig själva	35			.9228
Today in True Crime	194			.9270
Parcast Presents March Mysteries	43			.9275
Famous Fates	44			.9310
Sig.		1.000	1.000	.476

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

female speech

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05			
		1	2	3	4
Heavyweight	30	.1119			
The Hottest Take	32		.2343		
Famous Fates	44			.3882	
Parcast Presents March Mysteries	43			.4480	
Today in True Crime	194				.8947
Gynning & Berg - Hittar sig själva	35				.8972
Sig.		1.000	1.000	.444	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

male speech

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05		
		1	2	3
Gynning & Berg - Hittar sig själva	35	.0256		
Today in True Crime	194	.0323		
Parcast Presents March Mysteries	43		.4795	
Famous Fates	44		.5428	
The Hottest Take	32			.6667
Heavyweight	30			.6694
Sig.		1.000	.370	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

pitch average

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05			
		1	2	3	4
Heavyweight	30	82.7410			
Parcast Presents March Mysteries	43	83.5636			
Famous Fates	44		93.4469		
Today in True Crime	194			105.1395	
The Hottest Take	32				116.0309
Gynning & Berg - Hittar sig själva	35				
Sig.		.999	1.000	1.000	1.000

pitch average

Tukey HSD[a,b]

show name	Subset for...
	5
Heavyweight	
Parcast Presents March Mysteries	
Famous Fates	
Today in True Crime	
The Hottest Take	
Gynning & Berg - Hittar sig själva	135.5762
Sig.	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

pitch standard deviation

Tukey HSD[a,b]

		Subset for alpha = 0.05			
show name	N	1	2	3	4
Heavyweight	30	73.6600			
Parcast Presents March Mysteries	43	76.0681			
Famous Fates	44		83.2564		
The Hottest Take	32		87.7080	87.7080	
Today in True Crime	194			89.5782	
Gynning & Berg - Hittar sig själva	35				101.2665
Sig.		.634	.051	.836	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

speech rate

Tukey HSD[a,b]

		Subset for alpha = 0.05			
show name	N	1	2	3	4
Parcast Presents March Mysteries	43	3.7395			
Today in True Crime	194	3.7453			
Heavyweight	30		3.9297		
Famous Fates	44			4.1416	
Gynning & Berg - Hittar sig själva	35			4.1691	
The Hottest Take	32				4.4006
Sig.		1.000	1.000	.979	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Homogeneous Subsets

Percentage of music

Tukey HSD[a,b]

show name	N	Subset for alpha = 0.05		
		1	2	3
Parcast Presents March Mysteries	43	.0115		
Today in True Crime	194	.0188		
The Hottest Take	32		.0298	
Gynning & Berg - Hittar sig själva	35		.0331	
Famous Fates	44		.0338	
Heavyweight	30			.0857
Sig.		.166	.787	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 41.582.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Appendix H

Kruskal Wallis and Welch ANOVA results for Podcast Genres

Test Statistics[a,b]

	music	noise	silence	speech	female speech	male speech
Kruskal-Wallis H	99.142	119.185	95.879	111.628	264.984	254.144
df	5	5	5	5	5	5
Asymp. Sig.	.000	.000	.000	.000	.000	.000

Test Statistics[a,b]

	pitch average	pitch standard deviation	speech rate
Kruskal-Wallis H	153.224	217.806	168.247
df	5	5	5
Asymp. Sig.	.000	.000	.000

a. Kruskal Wallis Test

b. Grouping Variable: genre

Robust Tests of Equality of Means

		Statistic[a]	df1	df2	Sig.
music	Welch	28.218	5	160.148	.000
noise	Welch	68.096	5	183.213	.000
silence	Welch	34.656	5	166.961	.000
speech	Welch	41.535	5	151.683	.000
female speech	Welch	464.148	5	187.391	.000
male speech	Welch	492.156	5	184.795	.000
pitch average	Welch	272.695	5	173.153	.000
pitch standard deviation	Welch	164.770	5	170.192	.000
speech rate	Welch	93.822	5	158.287	.000

a. Asymptotically F distributed.

Appendix I

Post Hoc Tukey results for Podcast Genres

Percentage of music

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05		
		1	2	3
Arts & Entertainment	133	.0136		
Stories	281	.0200	.0200	
Sports & Recreation	32	.0298	.0298	
Music	35		.0331	
Society & Culture	140		.0364	
News & Politics	75			.0713
Sig.		.089	.082	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

noise

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05	
		1	2
Music	35	.0017	
Sports & Recreation	32	.0044	
Society & Culture	140	.0074	
Stories	281	.0102	
News & Politics	75	.0162	
Arts & Entertainment	133		.2668
Sig.		.997	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

silence

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05		
		1	2	3
Stories	281	.0421		
Music	35	.0424		
Arts & Entertainment	133	.0555	.0555	
News & Politics	75		.0597	
Sports & Recreation	32		.0647	.0647
Society & Culture	140			.0775
Sig.		.165	.570	.209

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

speech

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05	
		1	2
Arts & Entertainment	133	.6641	
News & Politics	75		.8528
Society & Culture	140		.8787
Sports & Recreation	32		.9011
Music	35		.9228
Stories	281		.9277
Sig.		1.000	.130

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

female speech

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05		
		1	2	3
Sports & Recreation	32	.2343		
Society & Culture	140	.2503		
News & Politics	75	.3029		
Arts & Entertainment	133	.3043		
Stories	281		.7470	
Music	35			.8972
Sig.		.703	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

male speech

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05			
		1	2	3	4
Music	35	.0256			
Stories	281		.1807		
Arts & Entertainment	133			.3598	
News & Politics	75				.5499
Society & Culture	140				.6283
Sports & Recreation	32				.6667
Sig.		1.000	1.000	1.000	.111

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

pitch average

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05				
		1	2	3	4	5
News & Politics	75	83.1962				
Arts & Entertainment	133	86.2298	86.2298			
Society & Culture	140		97.5773	97.5773		
Stories	281			100.0069		
Sports & Recreation	32				116.0309	
Music	35					135.5762
Sig.		.974	.051	.990	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

pitch standard deviation

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05			
		1	2	3	4
Arts & Entertainment	133	66.9929			
News & Politics	75		76.4400		
Society & Culture	140		79.2007		
Stories	281			86.5209	
Sports & Recreation	32			87.7080	
Music	35				101.2665
Sig.		1.000	.850	.996	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

speech rate

Tukey HSD[a,b]

genre	N	Subset for alpha = 0.05		
		1	2	3
News & Politics	75	3.8023		
Stories	281	3.8065		
Society & Culture	140	3.9759		
Arts & Entertainment	133	3.9826	3.9826	
Music	35		4.1691	
Sports & Recreation	32			4.4006
Sig.		.081	.063	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 65.663.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.