Creating a Speech and Music Emotion Recognition System for Mixed Source Audio

MASTER THESIS

Casper Laugs

Supervised by

Anja Volk

Heysem Kaya

Hendrik Vincent Koops

Universiteit Utrecht

A thesis presented for the degree of Game & Media Technology

Graduate School of Natural Sciences
Utrecht University

August 30, 2020

Abstract

While both speech emotion recognition and music emotion recognition have been studied extensively in different communities, little research went into the recognition of emotion from mixed audio sources, i.e. when both speech and music are present. However, many application scenarios require models that are able to extract emotions from mixed audio sources, such as television content. We coined this recognition problem as MiSME recognition, Mixed Speech Music Emotion recognition. This master thesis studies how mixed audio affects both speech and music emotion recognition using a random forest and deep neural network model, and investigates if blind source separation of the mixed signal beforehand is beneficial, along with a feature importance analysis. We created a mixed audio dataset, with 25% speech-music overlap without contextual relationship between the two.

The speech and music emotion recognition experiments consisted of six experiments each, where the models were trained and tested on different combinations of the three audio types available: single-source audio (speech-only / music-only), mixed audio or blind-source separated audio. Deezer's Spleeter tool was used to create the blind-source separated version of the dataset.

The results showed that both speech and music emotion recognition are possible on mixed audio far above chance-level, meaning that a functional MiSME system can indeed be created. The speech models performed best when blind-source separation was included as a preprocessing step, but there remained a performance gap compared to speech-only audio, suggesting that lower levels of speech emotion recognition performance should be expected on mixed audio. The music models were able to perform better on mixed audio, with and without blind-source separation depending on the model, than on music-only audio. We attributed this to speech-presence forcing the models to favor less ambiguous features during training, resulting in a better generalizing model. The results also showed that both speech and music models trained on single-source audio achieve chance-level performance on mixed audio, rendering them incapable of MiSME recognition.

The feature importance analysis produced many insights regarding which speech and music features are (un)important for mixed audio. It showed that the optimal features were highly dissimilar between audio types for both speech and music emotion recognition, which means that the optimal features for each audio type are different.

This research thus not only shows that both speech and music emotion recognition are possible far above chance-level on mixed audio, but also gives insight into the use of blind-source separation and common speech and music features in a mixed audio scenario. This is important knowledge when estimating emotion from real-world data, where individual speech and music tracks are often not available.

Acknowledgements

Many have supported me through this research, both directly and indirectly. I want to start with expressing my deepest gratitude towards Anja and Heysem for their supervision during these many months. Their guidance was vital and they helped me a lot. They remained supportive through its entire duration, always willing to provide feedback. Heysem was always thinking along, opting for new and interesting additions or solutions when I presented him a problem I was stuck with, and without Anja's greatly detailed feedback this thesis would have been less easy to digest.

I am also greatly indebted to Vincent. As my daily supervisor at RTL, and desk-mate, Vincent helped me greatly. He was always willing to help me with questions and issues, big or small, while concurrently working on so many other projects. His guiding insights helped me at many moments when I was stuck on something, allowing me to keep a steady momentum of progression. I have learned a lot from his guidance. He was a great supervisor, and even a mentor of sorts for me.

I would also like to pay my special regards to RTL's Data Science team, including Daan, Tanja, Rana, Rashid, Maurits, Niels and many others. This was my first time 'working' at a company and in office environment. They made me feel welcome from the moment I arrived. Their hospitality made going to Hilversum every day a joy. We had many laughs, interesting conversations and beertastings during 'BeerClub' at the end of each week. It was something I really missed during the last few months, as the COVID-19 restrictions prevented us from going to the office. I also learned many things there which I could not learn at the university, such as office culture and non-ICT aspects of running a business, such as fiscal numbers and how market change affects their business. I can only hope that the colleagues of my first job after graduating will be half as fun to work with as you guys, I will miss you all!

One of the many 'BeerClub' group photo's

Contents

Chapter 1

Introduction

1.1 Preface

The rise of streaming and video-on-demand has caused disruptive change in the television entertainment industry. Regular television viewership has been steadily declining every year as consumers move to video-on-demand services like Netflix or HBO. This change can also be felt in the Dutch market. Before video-on-demand Dutch broadcasting companies only had each other as competitors, but they now share the market with global giants such as Netflix and Youtube. Disney and Apple have even entered the Dutch market in 2019 with their own services. These media companies do not only compete for users through pricing and content anymore, which has always been part of the industry, but now also through the provided services of the video-on-demand software. Not only is the quality of the software itself important but also many supportive features come into play. For example, strong recommendation systems entice viewers to use the service more, always offering relevant content in such a way that the user never feels like the service has become obsolete. Ensuring that the video-on-demand service meets the expectations of their users, and offers satisfying features, is more important than ever. Any inconvenience or annoyance might result in a loss of subscribers.

RTL The Netherlands is one of the three main players in the Dutch television entertainment market. They own two video-on-demand services, Videoland and RTL XL. Their Data Science department processes usage data to optimize the video-on-demand content while also developing tools to create features that directly and indirectly support the video-on-demand platforms and their users. In relation to this research, the department has been actively exploring if the emotion present in their television content can be processed and used somehow to provide an even better user experience for the users of their platforms. At the time of this research they have successfully found ways to use the visual emotion information. However, audio is a strong complementary carrier of emotion information in television content, specifically music and speech. They wish to use the emotion information contained in the audio as well.

Before this emotion information can be used in any kind of system it must be extracted from the raw audio. Here lies a non-trivial and scientifically relevant challenge for RTL The Netherlands. In RTL The Netherlands's use case the audio is from movies, tv-shows, theater productions and live tv-broadcasts. The audio in these types of content contains both speech and music, often being concurrently present (overlapping) and mixed in with other types of sounds. Both speech emotion recognition (SER) and music emotion recognition (MER) are active fields of research that have produced many models that can computationally classify emotions on their respective types of audio. However, scientific knowledge on how to handle emotion recognition when music and speech occur concurrently is sparse to non-existent. This masters thesis research explores this issue, with the goal of gaining scientific insight into the problem itself, and showcasing that a simple

- **MiSME system** - MIxed source Speech Music Emotion system. This is the system produced in this research that takes a mixed-source audio sample and predicts the speech and music emotion separately.
- **MiSME recognition** - Recognizing both the emotion of speech and the emotion of music in mixed audio.
- **Television content** - Term used for all types of entertainment content which can contain interesting emotion information for RTL The Netherlands. This includes movies, tv-shows and more.
- **Single-source** - Audio with only one source, so either speech-only or music-only audio depending on the context.
- **Mixed audio** - Audio with only both speech and music present.
- **Blind-source separated audio** - The isolated approximation of either the original speech or music audio from the mixed audio, produced by the blind-source separation component.
- **Amplification relationship** - The specific relationship between the music and speech that is sometimes present in television content. Music is regularly used to convey cues about the emotion of speech or other story elements (KUCHINKE1A et al., 2013).
- **Dominant features** - Features that are highly effective at allowing to model to predict emotions.

Figure 1.1: List of terms and abbreviations introduced throughout this thesis

yet functioning *Mixed-source Speech Music Emotion* (MiSME) recognition system can be created. The research was done in cooperation with RTL The Netherlands's Data Science department.

1.2 Terminology

This research tackles a relatively unexplored problem which overlaps with many existing fields of study, meaning that many new problems and solutions are introduced throughout this research. To improve ease of reading many terms have been abbreviated or simplified. List 1.1 gives a quick overview of the terms introduced and used throughout this paper.

1.3 Speech and music emotion recognition

It is important to get a solid understanding of the problem at hand before we delve deeper into the research, especially since this is a relatively new type of emotion recognition problem. The previous section described that RTL The Netherlands wants to extract separate speech and music emotion information from mixed audio. This means that they require a system capable of both speech and music emotion recognition, producing separate emotion meta-data for the speech and music present in the audio. Speech emotion recognition and music emotion recognition are both expansive research fields related to information retrieval. They share many similarities, but also differ on many aspects. The literature review chapter covers their differences more clearly, but for now it is important to note how they are similar and how a system capable of both can be created.

Both speech and music emotion recognition are done using computational models. These models learn a mapping between observable acoustic cues present in the audio and a set of expressible emotions. This is similar to how humans communicate emotion, where the 'sender' adds emotion cues to the sound it produces to communicate emotion information, and the receiver tries to identify these cues and processes these to deduce the communicated emotion information. A simple example would be that the 'sender' would speak much more loudly to communicate that he or she is angry, along with other cues. The 'receiver', either human or a computational model, must identify that the 'sender' is speaking loudly and deduce that it tries to express anger by doing so.

These emotion cues are quantified for the model by extracting descriptive *features* from the audio signal. These features describe various properties of the audio, for example the average loudness or the minimum frequency present in the audio. Most features describe the audio on a low level, often too abstract for non-experts to directly comprehend. These often involve transforming the audio signal into a different representation, binning them and computing complex mathematical derivatives. A large set of features are used, sometimes even more than a thousand, allowing to model to base its emotion prediction on a combination of many individual descriptive features. The computation models learn a mapping between these features and expressible emotion during the training phase, after which they can be put to use.

Creating a MiSME recognition system thus requires at least one model capable of speech emotion recognition and one capable of music emotion recognition. This means that the MiSME system consists of at least two components, a *speech emotion recognition model* and a *music emotion recognition model*. However, these models require descriptive features as input. *Feature extractors* are needed to extract these from any given audio segment. This means that the MiSME system also requires a *speech feature extractor* and a *music feature extractor*. Separate ones are needed because SER and MER use different features, which is covered broadly in the next chapter.

Model training is necessary for the MiSME system to function, as the models need to 'know' a mapping from features to emotions. This requires training and training data, similar to any other machine learning problem. The MiSME recognition problem at hand specifically focuses on mixed audio, this is audio where the speech and music occur concurrently. This means that we need mixed audio samples with annotated speech and music emotion information, which can be used for training and testing. A mixed speech-music dataset was created for this research, which is broadly covered in the Methodology chapter.

By now a functioning MiSME system can be created, as we can train the speech emotion recognition model using the created mixed audio dataset and its speech emotion annotations, and the music emotion recognition model using the same dataset but with the music emotion annotations. The impact of mixed audio on speech and music emotion recognition can then be studied through various performance tests and other experiments.

However, we believe that the inclusion of blind-source separation as a preprocessing step might be beneficial for the performance of both speech and music emotion recognition models in a MiSME context. Blind-source separation is covered in more detail in the next chapter, but it tries to 'unmix' a mixed audio signal, producing isolated estimates of the original sound sources. While these estimates are often not perfect, it is not far fetched to assume that these isolate most of the other sources, reducing the degree to which they affect the feature extraction of the other modality. For example, music affects speech features less when blind-source separation isolates most of the music in the mixed audio before speech feature extraction. Blind-source separation is therefore included in some experiments as a preprocessing step.

We can create a MiSME system with these four or five components: the speech emotion recognition model, the speech feature extractor, the music emotion recognition model, the music feature extractor and the optional blind-source separation component. We are essentially combining a common speech emotion recognition pipeline, so a model and feature extractor, with a common music emotion recognition pipeline into one system. These should be able to produce speech and music emotion information from any type of mixed audio, and they can be trained on our handcrafted mixed-audio dataset. A blind-source separation component is optional. The architecture thus does not differ much from common SER, MER and other machine learning models. A visualization of the MiSME system architecture is depicted in Figure 1.2. This should be enough for now to get an idea of how the system functions and what is required to create a system capable of both speech and music emotion recognition. The Literature review and Methodology chapter cover the MiSME system more in-depth.

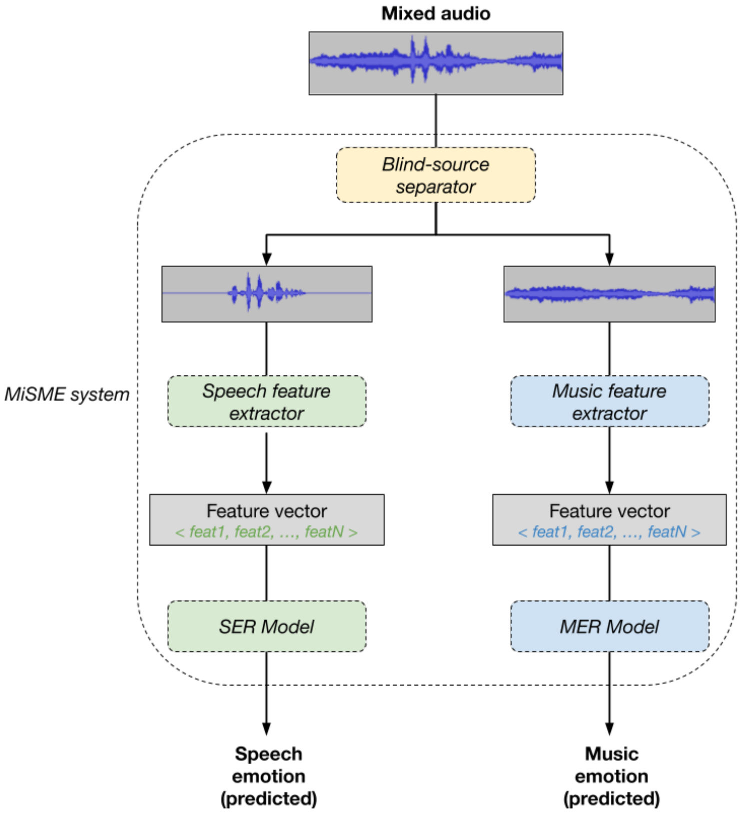


Figure 1.2: The flow of the MiSME system with blind-source separation

1.4 Problem definition

How a MiSME system can be created and how it functions on a basic level should be clear by now. However, as was briefly explained in the preface section, doing speech and music emotion recognition on mixed audio is a non-trivial challenge because it is an fairly unstudied problem. To be more specific, it is unknown how mixed audio affects both speech and music emotion recognition in a general sense, and here lies the main problem. It could be the case that the presence of the other audio source hardly affects the complexity of the emotion recognition tasks, meaning that the model does not have to adapt much compared to single-source audio. However, it is a more likely assumption that they negatively affect each other to at least some degree. The presence of mixed audio 'distorts' feature extraction, as the extracted speech and music features describe the mixed audio signal, rather than just the audio source which must be classified. This makes those features less descriptive compared to features extracted from 'normal' single-source audio. It is not far fetched to assume that less descriptive features result in a more complex emotion recognition task and therefore lower performance, but this requires proper research. If mixed audio indeed results in worse performance compared to non-mixed audio, it also becomes interesting to see how this loss in performance can be minimized.

Exploring this problem by attempting to create a MiSME system, and study its differences compared to non-mixed audio, is therefore of great scientific value for both speech and music emotion recognition research. The knowledge and insights from this research can aid others in the future when tackling similar problems, including RTL The Netherlands. The contributions of this research are covered in the next section.

In most MiSME use cases, possible even every, there is a contextual relationship present between the speech and music. Television content is a great example of this. Music is regularly used to convey cues about the emotion of speech or other story elements (KUCHINKE1A et al., 2013). The exact cues used are often even genre-specific. For example, the same combination of speech and music emotion might mean different things in a scene from a comedy movie than a horror movie. This context, found in the relationship between the speech and music, thus contains valuable emotion information. In an optimal scenario for RTL The Netherlands a MiSME system would be produced that is optimized for their television content, which includes the contextual relationship between speech and music. However, the insights gained from such a model might not transfer to other use cases. We already stated that there is a lack of knowledge regarding how mixed audio affects speech and music emotion recognition in a general sense. Finding a generic solution for the MiSME problem, which is not context specific, would therefore be of greater scientific contribution, as the results would be applicable to all kinds of MiSME recognition problems. Because of this the decision was made to study MiSME recognition this way, purposely excluding the contextual relationship between speech and music from the experiments. This allows us to produce scientifically valuable results, which can still be used by RTL The Netherlands to develop a system of their own, albeit without strong contextual optimization. During the development of their own system they can focus on including this contextual relationship, if desired.

However, at the time of this research it was not known if even a service-ably functioning MiSME recognition system could be created. While we already speculated how mixed audio might affect speech and music emotion recognition, it could be the case that no models can be developed that perform significantly better than chance. While this is a unlikely assumption, there is no proof that it is not the case. It is therefore of high importance to see if a functioning (generic) MiSME recognition system can be created, proving that MiSME recognition is possible to at least some degree regardless of context. We define 'functioning' as significantly above chance level, because that would mean that the models perform better than random guessing. However, we expect to see better performance than just 'significantly above chance-level'. This leads to the following main research question for this research:

> *Can a system be produced which can recognize both the emotion of speech and music in mixed audio, where both are concurrently present, significantly above chance level?*

This research focuses on answering this question, along with producing additional insight into the MiSME recognition problem space.

An advantage of the generic approach taken, where we exclude the contextual relationship between speech and music, is that it allows the MiSME system to be compared to generic but established MER and SER models, and the knowledge surrounding them. These comparisons can be used to gain more insight into the MiSME problem space and are seen as additional challenges. For example, which speech or music features are robust against mixed audio and which are not? What degree of performance decrease can we expect compared to non-mixed audio? Is there even a decrease at all compared to single-source recognition? Exploring these questions alongside the main hypothesis should produce enough knowledge to get a basic understanding of the MiSME problem space.

1.5 Contribution

The main contribution of this research is that it proves that MiSME recognition is possible far above chance-level. However, its contributions are not limited to just that single observation. This thesis describes the development and evaluation of, and research and analysis done surrounding the creation of the first MiSME recognition system. This has produced many insights regarding the MiSME recognition, which are valuable contributions for multiple stakeholders.

In this thesis we describe how models can be trained and evaluated on mixed audio, and which models are most efficient for either speech or music emotion recognition on mixed audio. It also showcases which levels of performance can be expected, and when the inclusion of blind-source separation is beneficial. The feature importance analysis showcases which features are of high importance in a mixed audio context, which can hopefully aid others during feature selection in future work. This research has produced enough knowledge that RTL The Netherlands, and others, should be able to make a MiSME system of their own, as they can copy the strengths of the created system while solving, or preventing, its documented flaws.

The thesis also contributes significantly to the research fields of both speech emotion recognition and music emotion recognition, as it shows how mixed audio affects both recognition tasks on various aspects. This is because all experiments and analyses were done on multiple audio types, including single-source audio. By comparing the difference between the single-source audio experiments and mixed audio experiments we gain a better understanding how mixed-source emotion recognition differs from single-source emotion recognition on multiple levels. This was something which was little to nothing known about until now and is valuable knowledge.

We also hope that this thesis sparks enough interest to motivate others to study MiSME further, allowing it to evolve into a field of research of its own. A MiSME dataset was created in this research, showcasing how such a dataset could be created. Combine this with all other contributions already mentioned and there is more than enough knowledge and data available for future work. This research only scratches the surface of MiSME recognition, as it tackles to problem without any context, and the insightful results raised many new and more complex questions that are interesting to explore further.

1.6 Upcoming chapters

The next chapter covers various relevant literature, providing the necessary background knowledge to understand all aspects of the research. The topics covered are: emotion theory for both speech and music communication, speech emotion recognition, music emotion recognition and blind-source separation.

This is followed up by the Methodology chapter, which covers the design of the MiSME system, the mixed audio dataset, the experiments and more. This chapter provides enough knowledge to understand how to MiSME system was created, and how the results from each experiment and analysis should be interpreted.

The next chapter, the Results, covers the results of all experiments. The Discussion chapter builds upon these results, discussing the implications of the results and combines them with the feature importance analyses to deduce as much knowledge as possible from the experiments. After that the main research question is answered, followed by a reflection on the research itself.

Chapter 2

Literature review

The MiSME recognition task is an interdisciplinary challenge that spans many fields such as machine learning, speech and music information retrieval, signal processing, psychology, vocology and musicology. This means that a large amount of existing research is relevant. In the following section we discuss important basic theories and works of study of all relevant fields, which should be sufficient for understanding the MiSME recognition problem and the choices made in this research. In addition we cover some state-of-the-art work to show where the fields were at at the time of this research. We will start with basic emotion theory (2.1), followed by speech emotion recognition (2.2), music emotion recognition (2.3) and finally blind-source separation (2.4).

2.1 Emotion theory

The system of human emotion is complex and has many variables that define which emotional mental state is experienced. Creating a model that identifies the emotions expressed in audio requires a core understanding of human emotion. In this subsection we delve into the scientific knowledge regarding human emotion. The study of human emotion is called 'Affective science' and there exists no consensus on a single correct way to classify or quantify human emotion. Rather, there exist multiple acknowledged systems and emotion classifications. One of the dominant conflicts within Affective Science is whether emotions should be represented *categorical* or *dimensional*. Both ways of describing emotion have a long list of literature advocating for and against it, so there is no clear 'best option'.

A categorical approach to emotion representation uses 'labels' to identify emotional mental states of humans. The concept is based on the theory that there exists a set of culturally-independent primary emotions from which all secondary or other emotions can be derived (Ekman, 1992; Picard et al., 2001). This thus implies that all human beings share at least those primary emotions. In 1972 Ekman proposed the existence of six basic emotions: anger, fear, disgust, happiness, sadness and surprise (Ekman et al., 2013). While other emotions exist according to Ekman, they are nuanced instances of one of the six basic emotions. The categorical emotions are often called discrete emotions because they should be clearly distinguishable in human facial expression. Ekman's work was groundbreaking at the time and is still today often seen as the go-to theory for categorical emotions. However, his theories have been criticized extensively, suggesting that either an universal set of emotion does not exist or putting the validity of Ekman's proof that this universal set exists into question (Barrett, 2006).

According to Yang and Chen (2012) there exists no consensus on the number of categorical emotions or type of discrete emotion model that should be used with the field of music emotion recognition (MER). There exists no clear 'best' emotion model. They also state that Ekman's primary emotion set is too limited to fully capture the emotions perceived by humans in music. Using

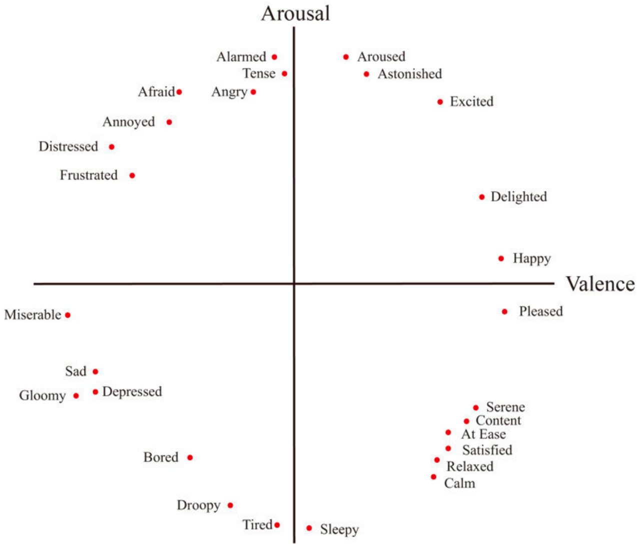


Figure 2.1: Russel's circumplex model, displaying various discrete emotion labels in the valence-arousal space (Seo and Huh, 2019)

a more expansive and complex emotion set is not the solution however as the problem lies in the language used to describe emotions, which is inherently ambiguous and differs between persons (Kim et al., 2010; Yang and Chen, 2012). It thus appears that there is no single 'correct' way to model emotion in a discrete manner, meaning that emotion models should be selected based on the context they are used in.

Complementary to the categorical representation of emotions there exists the dimensional representation, of which also many definitions exist. One of the earliest dimensional emotion model studies by Wundt and Judd (1897) suggested that emotion could be described using three dimensions: Pleasurable vs. unpleasurable, arousing vs. subduing and straining vs relaxing. This dimensional model is based on the theory that one system can represent all emotions similar to the human neuropsychological system that is responsible for all emotions.

Since then many dimensional emotion models have been proposed, but the most well known dimensional model is the circumplex model of affect (Russell, 1980), which represents all emotions using valence and arousal. Each emotion can be expressed as a combination of an arousal and valence intensity, where arousal encapsulates the intensity of the emotion from low to high and valence the degree of negativity-to-positivity of the emotion. An example of this two-dimensional model with categorical emotion labels is visible in Figure 2.1. It was later expanded upon by Russell and Barrett (1999).

Similar to the discussion around discrete emotion models there is also no consensus on which dimensional model is 'correct'. One of the more persuasive works (Fontaine et al., 2007) argues that two dimensions, especially the ones used in Russel's model, are not enough to encapsulate all factors of emotion, arguing that a third or even fourth dimension is necessary.

The question if either the discrete or dimensional representation is superior is open for debate. According to Lazarus (1991), dimensional models, especially the valence-arousal model, blur important psychological distinctions and other aspects of the human emotion processing. Certain emotions might lie very close to each other in the dimensional space but have very different implications for the human that experiences them. For example, 'alarmed' and 'angry' lie very close yet can have very different implications on the person experiencing the emotion. However, using discrete emotions would force binning of more ambiguous or complex emotional states to simpler single emotion labels. It becomes apparent that both types of representation have their own flaws.

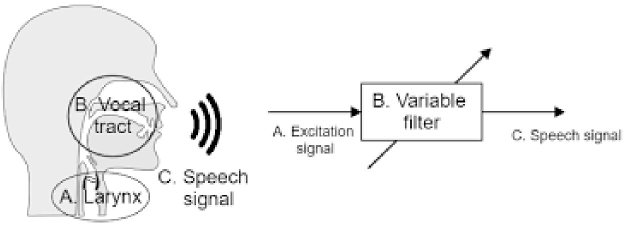


Figure 2.2: Source-filter theory, the vocal tract (B) filters the sound produced by the larynx (A) to create a speech signal containing emotional state cues (C) (Tits et al., 2019)

Modeling human emotion is thus not an easy task. There exist many types of discrete and dimensional models, but Ekman's primary emotion model and Russel's valence-arousal model are generally speaking the most used. Both types of emotion representation have their advantages and disadvantages, and these differences should be taken into account when deciding upon how to represent emotion for a certain problem at hand, which will become relevant to our research later on. While human emotion theory can be discussed in much more detail, this should suffice to be able to understand the human emotion aspect of both the SER and MER field, which are covered in the next sections.

2.2 Speech emotion recognition

The scientific field of speech emotion recognition (SER) focuses on creating computational models that can detect emotion in speech audio. This entire field is based on human speech emotion communication, which is covered first. This is followed by exploring the computational approach taken, from feature usage to existing datasets and created SER models.

2.2.1 Speech emotion communication

Speech itself is one of the primary channels through which humans communicate and express emotions. An emotional state is expressed by a human through speech by changing the sound they produce using a internal muscle-controlled filter, which adds specific acoustic cues to the speech that convey the emotional state (Bachorowski, 1999). To be specific, the source of sound energy produced by the larynx is modulated to convey emotion cues through the shape of the vocal tract and results in changes in vocal pitch, jitter and shimmer. See Figure 2.2 for a visual example. There are also external filters that affect the sound of speech, such as the mouth shape due to facial expression, but these are not considered part of the source-filter theory.

Through studying source-filter behavior various source-filter cue patterns have been found. Banse and Scherer (1996) showed that high arousal emotions (anger, fear, joy etc.) have a positive association with pitch and vocal intensity among other acoustic features, while low arousal emotions (sad, calm etc.) were associated with lower mean pitch and pitch decrease over time. For these emotions more nuanced patterns were also found, for example that the pitch decreases over time for anger but increases for joy. It was also discovered that the experienced intensity of the emotion affects the intensity to which acoustic cues are expressed, suggesting that lower emotional intensity makes the source-filter cue patterns less present or even disappear (Bachorowski and Owren, 1995).

Perceiving and processing these vocal cues correctly is a hard and complex task for the listener. The human error rate in many studies are often high, which can be seen in the later section where we discuss SER datasets. The recognition rate on average hovers between forty to sixty percent depending on the complexity of the dataset, meaning that about half of instances the emotion cues are interpreted wrong by the average listener in a controlled experimental setting. In real life scenarios there are many other factors into play, so the human error rates reported on the datasets might be optimistic. However, Bachorowski (1999) suggests a strong contributing factor for these high error rates is that humans are far less accurate in emotion detection from voices they are not familiar with compared to ones they are familiar with.

2.2.2 A computational approach

Compared to the human brain, computers are able to much more accurately observe and process (digitized) acoustic information, but they lack the complex understanding of human emotion communication and the learning abilities of us humans. The goal of the SER field is to develop computational models that can process these acoustic cues contained in the audio to obtain the expressed emotion for various purposes. These models can then be used for speech emotion recognition tasks unsuitable or uneconomic for humans, such as monitoring the emotion of customers calling a help-desk or the creation of emotion meta-data.

The acoustic cues used for emotion communication are quantified for the model by extracting *features* from various representations of the audio signal that describe various properties of the audio. See Figure 2.3, which depicts multiple representations of the same audio sample from which features can be extracted. During model training the model learns the relationship between the features and emotions. Based on this learned mapping the model can recognize the emotion of new speech audio through the extracted features. An example of simple speech features would be the average loudness in decibel, or the minimum and maximum frequency of the audio.

Features can be extracted either locally or globally from the audio. *Local features* are extracted from each frame of the signal, which are small windowed sections of often consistent length. *Global features* are extracted from the full duration speech signal, assuming it contains only one utterance. Another strategy for feature extraction is phoneme-based segmentation, but it relies much on the quality of the phoneme segmentations which is prone to errors. There exist four types of commonly used speech feature types: continuous, qualitative, spectral and Teager-energy-operator (TEO) features. These will be covered in the following subsections.

There are two excellent works which summarize the state of SER research. El Ayadi et al. (2011) offers great insight into how the SER grew from its earliest works to 2011, while Khalil et al. (2019) covers more recent progress, when deep learning became a popular approach. Below is a simplified combination of these works, along with other findings.

There are a couple of interesting observations made by El Ayadi et al. (2011) that define the field of SER. First off, a universal set of dominant feature for speech emotion recognition does not seem to exist. This is mainly due to the acoustic variability caused by sentence difference, speaking styles and other characteristics between speakers, which directly affect the most common speech features such as pitch and energy contours. Finding speaker-independent features is an on-going challenge. Secondly, there is also no solid agreement within the field if there exist any dominant acoustic features related to valence. Most dominant features when disregarding acoustic variability (speaker-independence) can only distinguish emotions on the arousal plane consistently, while their performance on valence is lacking. Finally, there appeared to be a consensus that global features are superior compared to local features because they were significantly better at distinguishing between emotions that differ in the arousal plane.

As stated earlier, there are four main types of features. Each is discussed below.

Continuous features Continuous features represent acoustic properties strongly related to prosody and the filter-theory, as shown in Figure 2.2. According to El Ayadi et al. (2011) the majority of the researchers believe that continuous features contain many cues to the emotional content of an utterance. The most commonly used global features are pitch, energy, duration and formants-based. Various functionals are used to describe these features such as the mean, deviation, linear regression coefficients and ratio of slope contours.

Voice quality features El Ayadi et al. (2011) summarizes a large number of researches related to voice quality features. Voice quality properties lead to certain impressions of voice, for example a harsh, tense or breathy voice. While there are works advocating the theory that the emotional content of an utterance is related to features of voice quality, the work by Gabrielsson and Juslin (2003) showed that there was no one-to-one mapping between voice quality and affect. Even though no one-to-one mapping could be found, voice quality features were still tested for their SER capabilities, which we will cover later when we discuss produced SER models.

Determining voice quality from a signal and encapsulating that in features is difficult. One approach is the inverse-filtering of the speech signal by removing the filtering effect of the vocal tract to obtain the glottal signal (A. in Figure 2.2). However, this is difficult as the technique is based on an approximation of the vocal tract filter. This filtering can be skipped but requires a model to estimate the voice quality features. Known models can only do this with an accuracy of 68.5% (Hansen and Bou-Ghazale, 1997), which is considered too inconsistent to produce usable features. Therefore, voice quality features are often left out or only used in combination with other, more stable features.

Spectral features Spectral features are popular features to use for SER models (El Ayadi et al., 2011). Many models use Mel-frequency cepstrum coefficients (MFCC) features along with other features for both categorical and continuous emotions recognition. Spectral features are obtained by converting the time-based audio signal into the frequency domain using the Fourier-transform. Spectral features are often used as local speech features because the distribution of spectral energy across the speech range of frequency differs drastically over an utterance. The spectral features operate in two different formats, a linear and a cepstral-spectrum based one. There is no consensus that one performs better than the other (Bou-Ghazale and Hansen, 2000), but in general spectral features are effective at speech emotion recognition.

Teagar-energy-operator features The Teagar-energy-operator features (Teagar and Teagar, 1990) are based on the theory that the muscle tension of a speaker affects the air flow in the vocal system and can be used for stress detection in speech. However, its performance on speech recognition is outperformed by MFCC features, making TEO features only useful for specific stress detection tasks.

Speech emotion datasets

The goal of any SER model is to learn a mapping from the given speech features to the correct emotion as best as possible. This mapping is learned by training the model on a suitable dataset of speech audio samples with known emotions. The quality of the trained model depends on both the suitability of the used model and features, along with the quality of the training dataset used. Many aspects of the dataset being used define how well a model can generalize to actual speech emotion recognition tasks on unseen audio. The quality of a dataset is defined by aspects such as the size, speaker variation, language, emotions present, intensities and sentence length. Training a model with a dataset in another language than on which it will be used, for example, can result in lower performance. In this subsection we explore which datasets have been created for SER training and how they differ.

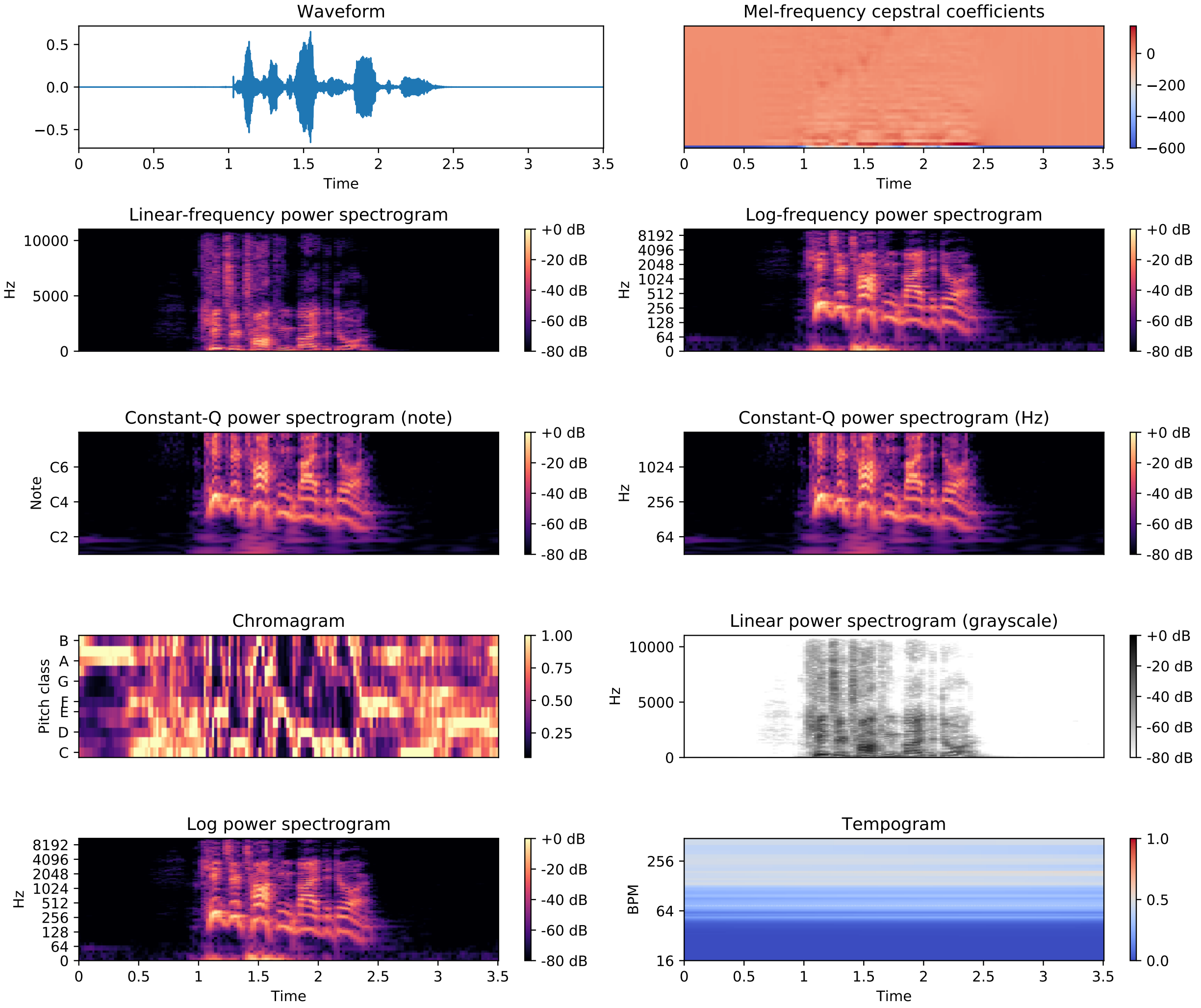


Figure 2.3: Various representations of speech audio from which features can be extracted, created using Librosa and a speech sample from our dataset. See the Librosa documentation for more information about each representation (Brian McFee et al., 2015)

A SER dataset review by Swain et al. (2018) lists a total of 59 datasets, with likely more existing. There is much variation between the datasets, mainly on emotions used, language and size. Some datasets even cover specific scenario's, such as child-adult communication or talking-while-eating.

Availability is a strongly present issue according to El Ayadi et al. (2011), at the time of their research a large majority of the databases were not available for public use. Availability has been improving since then as more recent datasets are publicly available, but many are still inaccessible.

We have selected five publicly available datasets to cover in more detail, selected from the large pool of datasets listed by Swain et al. (2018) with the MiSME problem in mind. These are: Emo-db, the INTERFACE dataset, GEMEP, RAVDESS and CREMA-D. Table 2.1 shows these five datasets in an overview. Please note that all of these datasets use categorical emotions.

Before we cover them it is necessary to discuss human recognition rate. This is the average accuracy to which humans were able to recognize the correct emotion in the samples of that dataset. A lower recognition rate implies that the samples are harder to recognize for humans. This could either be because some samples are hard but realistic, for example low intensity expression that is hard to detect for humans in general. As stated earlier, the less intense the emotion, the more subtle the acoustic cues become. However, it could also be that the captured emotion expres-

Name	Emotions	Actors	Intens.	Utter.	Samples	Avg. annot.	Lang.
Emo-db, 2005	Neutral, anger, fear, joy, sadness, disgust and boredom	10	1	10	300	20	German
INTERFACE, 2002	Neutral-slow, neutral-fast, anger, fear, joy, sadness, disgust and surprise	2	1	175	5520	0.16	Spanish
GEMEP, 2010	Neutral, joy, amusement, pride, pleasure, relief, interest, rage, panic fear, despair, irritation, anxiety, sadness, shame, surprise, admiration, disgust, contempt and tenderness	10	4	2+	7300	10+	Gibberish, French
RAVDESS, 2011	Neutral, angry, fearful, happy, sad, disgust, surprise and calm	24	2	2	4320	22	English
CREMA-D, 2014	Neutral, angry, fearful, happy, sad, disgust	91	4	12	7442	30	English

Table 2.1: An overview of freely available scientific speech emotion datasets, showcasing the differences in included emotions, actors, emotion intensities, unique utterances, samples, language and average emotion annotations per sample.

sion is ambiguous, meaning that it does not contain the correct cues for that emotion. Because most datasets are acted recordings, where the speech captured is expressed by actors following instructions, the ambiguity could be caused by bad acting for example. It is important to keep this distinction in mind, a low human recognition rate does not mean that the dataset is of lower quality, but it could be.

Emo-db Otherwise known as the Berlin emotional database (Burkhardt et al., 2005), Emo-db is a SER database of German actors speaking utterances in 7 different emotions. It was one of the earliest publicly available SER datasets and has been used in many SER research. It has remained popular, albeit that it is used along side more modern datasets just to allow comparison to old SER research. Compared to the other datasets Emo-db is quite small. However, it has a human recognition rate of 80%, the highest of all datasets. This means that it is an easy dataset to train models on, which was beneficial in the early days of SER.

INTERFACE The INTERFACE dataset (Hozjan et al., 2002) was created before Emo-db, but only became partially public available in 2011. It covers the same emotions as Emo-db but splits 'neutral' in a fast and slow variant. INTERFACE is a combination of an English, Spanish, Slovenian and French dataset, with the Spanish dataset being the only one that is publicly available. Albeit that the creators claim that it has a human recognition rate similar to Emo-db, it has an extremely low amount of annotations, making that claim unreliable. However, it is almost twenty time as large as Emo-db.

GEMEP The Geneva Multimodal Emotion Portayal Corpus (Bänziger and Scherer, 2010) has the largest emotion corpus of all five dataset, a grand total of 18 emotions. The emotions are based on a combination of primary and secondary emotions along with the valence and arousal emotion model. Its speech is not in a language but rather gibberish with an western accent. The idea behind this is that models trained on this gibberish would generalize to many western languages because the model is not able to learn language-specific characteristics during training. GEMEP also distinguishes between four emotion intensities, covering a larger spectrum of emotion expression than the other datasets. However, the human recognition rate is only 38%.

RAVDESS The Ryerson Audio-Visual Database of Emotional Speech and Song (Livingstone and Russo, 2018) contains both spoken and sung utterances, of which speech is only relevant for this research. It uses Ekman's primary emotions along with 'calm' and 'neutral' for a total of eight emotion labels. It differentiates between two intensity levels per emotion, compared to the four of GEMEP. It has a human recognition rate of 62.5%.

CREMA-D The CREMA-D dataset (Cao et al., 2014) has a much larger pool of actors compared to the other dataset, a total of 91 actors. It uses only six emotions and differentiates between four intensity levels. It is captured using audio and video. Audio-only experiments had an average human recognition rate of 40.9%.

From just these five SER datasets it is noticeable that the human recognition rate can wildly differ, where more complex datasets using more emotions and intensities score lower. There appears to be a trade-off between a guaranteed quality of the dataset and how expansive it is. It is important to take this in consideration when selecting a SER dataset, which we will come back to in the Methodology chapter.

Produced models and related studies

The field reviews by Khalil et al. (2019) and El Ayadi et al. (2011) offer a broad overview of most SER research until 2019. The most relevant of these works are discussed below, which should be enough to getting a basic understanding on how the field developed and how different types of models and features perform. Readers that are curious and wish to learn more about the fields should study these reviews.

Speech emotion recognition took off in the early 2000s as fields related to speech information retrieval became more interesting due to the internet and media sharing. One of the first SER break-throughs was the work by Nwe et al. (2003). Nwe produced a groundbreaking Hidden Markov Model that was able to recognize the set of primary emotions with an average accuracy of 77% using only Short Term Log Frequency Power Coefficients features. While the used dataset was simple compared to modern SER research, the model made a leap in performance compared to earlier works. These earlier works achieved lower accuracy on more simple datasets (Dellaert et al., 1996; McGilloway et al., 2000).

For a long time Hidden Markov Models were the dominant type of model for speech emotion recognition (El Ayadi et al., 2011). This changed around 2014 as the field of neural networks had seen many major improvements, with new and strong performing neural network variants being discovered such as Recurrent Neural Networks and Long Short-Term Memory (LSTM) Neural Networks. This led to a majority of the more recent SER research using deep learning techniques (Khalil et al., 2019).

Regarding top performing models, Seehapoch and Wongthanavasu (2013) combined common speech features with a SVM-based classifier and reached a recognition rate of 89.8% on Emo-db, a 7-way classification task. This is significantly above the human recognition rate, which means that a model can outperform the average human. This performance was improved upon by Guo et al. (2018), achieving an accuracy of 91.3% using a Convolution Neural Network focusing on amplitude and phase features. This was again improved upon by Bhavan et al. (2019), achieving 92.45% using a bagged SVM model, going against the trend that deep learning was superior.

Wang et al. (2015) produced a model that used a combination of commonly used features (MFCC) with Fourier transform features that correspond to first and second order harmony. These features capture the voice quality of the speaker. The inclusion of these voice quality features improved the recognition rate on their dataset by more than 10%, suggesting that voice quality features can be very effective when combined with other features.

Regarding features, Tian et al. (2015) showed that low level descriptors outperform utterance-level (global) features on acted speech. Even though low level descriptors are preferred over segmental features (i.e. utterance-level) by most researchers (Anagnostopoulos et al., 2015), a couple segmental features such as MFCC and voice quality features are often included. For example, a large number of published SER models include MFCC features (Bhavan et al., 2019; Hasan et al., 2004; Kwon et al., 2003; Neiberg et al., 2006; Sato and Obuchi, 2007; Wang et al., 2015).

Finally, a different approach to developing SER models has been gaining popularity the last few years. Instead of manually selecting a pool of features, the extraction of useful features is included in the learning task of the model. This is called an 'end-to-end' approach. In an end-to-end approach the model must learn to recognize the emotion from the raw or a transformed version of the audio signal instead of a feature vector containing computed features.

Zheng et al. (2015) used principle component analysis (PCE) on a raw log-spectrogram representation of the audio. This produced a list of useful 'raw audio features', which were fed to deep-neural network model. It only reached a recognition rate of 40% on the IEMOCAP database, consisting of 5 different discrete emotions. Trigeorgis et al. (2016) attempted this approach as well, using Long Short-Term Memory (LSTM) neural networks in combination with a Convolutional neural network to create a model that can predict emotions from raw audio signals. Using the RECOLA dataset it outperformed other models that used pre-defined feature-sets, suggesting that there is potential.

To conclude, speech emotion recognition is a fairly developed field that keeps improving. There exist a large number of datasets for SER training differing in language, complexity, emotions and more. Many SER models have been produced, achieving performance above human recognition levels. A deep learning and an end-to-end approach to SER have been gaining traction in recent years and perform well, even though these models are often not top performers on their respective datasets.

2.3 Music emotion recognition

In this section we explore the other emotion recognition task of the MiSME problem, the field of music emotion recognition, or MER in short. This field is noticeably younger, seeing its earliest publications around 2003. The advent of digital music platforms and recommendation systems however led to an increased interest and growth of the field.

MER is based on music emotion communication theory, and focuses on developing computational models that are able to do music emotion recognition, similar to SER. We will first delve into music emotion theory, followed up by exploring the computational approach taken, from feature usage to existing datasets and created MER models.

2.3.1 Music emotion theory

The main motivation for humans to create music is to express emotions (Eerola and Vuoskoski, 2013). This coincides well with the fact that music is a strong communication channel for emotion. How humans express and observe emotions through music has been studied by many, with the earliest studies dating back to ancient Greece (Kramarz, 2017). Generally speaking there exist three main theories on how emotion is expressed through music. The chapter 'Emotional expression in music' of The Handbook of Affective Science (Gabrielsson and Juslin, 2003) provides a broad overview of the many music emotion theories and studies published over time, including these theories. While these theories are barely directly recognizable in modern music emotion models, they served as a foundation for the field and covering them helps with creating an understanding how music emotion can be modeled in different ways.

In the handbook Juslin draws similarities between music emotion communication and Brunswik's lens model (Brunswik, 1956). In music emotion communication a performer encodes a certain emotion in the performance using a set of probabilistic and redundant cues. The listener then tries to decode the performance using similar cues to judge the intended emotion. The success of communication depends on both the ability to encode and decode the signal using cues, along with the degree of correspondence of both party's probabilistic cues. There exist three main theories how these cues take form in music, and why music is able to communicate emotions so well.

The first theory (Cooke, 1959) is that music has three separate expressive elements: An architectural aspect appealing to us because of the beauty of pure form, a pictorial aspect from imitation of natural sounds and a literary aspect as music is a language of emotion and meaning, akin to speech. These three elements can induce emotion (cues) alone or together according to Cooke.

Langer (1953, 2009) proposed a different theory and suggested that the elements of which music exists do not carry fixed lexical meaning like words in language do. Music elements are rather open symbols where the meaning of various elements can be understood "only through the meaning of the whole, through their relations within the total structure". Langer argued that the theory is supported by the idea that there is a mappable relationship between the structure of feelings and structure of elements of the music. This makes it different compared to Cooke's, as Langer suggests that all elements are related and not separately expressive elements.

Finally, Clynes (1977) suggested a theory that humans have biologically preprogrammed spatio-temporal patterns for communicating emotion. These patterns can use any channel (speech, music, facial expression etc.) as long as the pattern is preserved. These patterns can be incorporated in music pieces to communicate the biologically shared patterns (cue) between humans.

Before we delve deeper into the cues used, and which music features can be used to observe these cues, there is one thing that should be made clear. As Juslin and Laukka (2004) noted, there is a difference between the emotions being expressed by the music and the emotions being induced by the music. The expressed emotions are the emotions which the composer or performer tries to express, while the induced emotions are the emotions that the listeners experiences internally. The induced emotion is influenced by many contextual factors such as location and social presence, along with personal factors such as the motivation for listening (Mehrabian and Russell, 1974), and can thus be vastly different or more nuanced than the original expressed emotion by the artist. This causes emotion to be much more ambiguous in music than in speech. The experienced (induced emotion) can differ between individuals for the same music piece, making it a more personal experience, while emotion in speech suffer much less from these personal factors as it often comes down to the ability of deducing the cues to the correct emotion.

2.3.2 A computational approach

Music emotion recognition has also been tackled using computational models, training them on music samples with known emotions with the goal of learning a correct mapping between features and emotions. Many models, datasets, feature studies and more have been published since the early 2000s. Several field reviews (Eerola and Vuoskoski, 2013; Kim et al., 2010; Yang et al., 2018; Yang and Chen, 2012) along with the MER chapters in the works of Gabrielsson and Juslin (2003) and Aljanaki (2016) provide a broad overview of the development of MER since its inception.

- Timing
- Dynamics
- Articulation
- Timbre
- Pitch
- Interval
- Melody
- Harmony
- Tonality
- Rhythm
- Mode
- Loudness
- Musical form
- Vibrato

Figure 2.4: Musical elements that can be used to express emotion (Panda et al., 2015)

Music emotion features

Features can be computed from the music audio to quantify elements of the music. As the music emotion theories posed, the cues used to communicate emotion are often found in specific elements of the music. Quantifying these elements to features allows the model to learn a mapping of those cues to emotions. Panda et al. (2015) created an overview of all musical elements which correlate to emotions. These musical elements are listed in Figure 2.4. According to Yang and Chen (2012) the dominant musical elements for emotion communication are: energy, rhythm, melody and timbre. Energy is strongly related to loudness and dynamics from Panda's list. We will cover these four dominant musical elements and the features used to represent them briefly below.

Energy features They correspond to the loudness of a music piece and are strongly correlated with emotions on the arousal plane. High energy induces strong arousal, and low energy induces lower arousal (Yang and Chen, 2012). Common energy-related features are the root mean-squared loudness, specific loudness sensation coefficients (SONE) and total loudness.

Rhythm features Rhythm is the time-structure of notes and their accompanying strength. It can be described using tempo, meter and phrasing. Hevner (1936, 1937) observed that rhythm strongly affects emotions related to the valence-plane. A 'flowing' rhythm results in higher valence, while 'firm' rhythm results in lower valence. However, the tempo of the rhythm is strongly related to the arousal-plane instead of the valence-plane. A higher tempo correlates with higher arousal. Common rhythm-related features are rhythm strength, regularity, clarity, average onset frequency and average tempo.

Melody features Melody is created by a combination of pitch and rhythm, which causes musical tones to be perceived as single entities. Unlike the other dominant features melody does not have any human music emotion research supporting it. Hevner (1936) even showed that melodic direction had little to no effect on the perceived emotion. However, the inclusion of melodic features has been beneficial for many MER models (Panda et al., 2015). Common melody-related features are vibrato rate and extent, pitch descriptors and contour typology features.

Timbre features Timbre is the perceived sound quality of a musical tone or sound. Timber is related to emotions on both the valence and arousal plane (Yang and Chen, 2012). 'Sharp' timbre positively affects valence and arousal, while 'dark' timbre negatively affects both. Timbre is most often quantified using MFCC features.

Almost all music emotion cue knowledge appears to be in the dimensional emotion space rather than the discrete emotion space. This is because a dimensional approach better fits music emotion communication. As explained earlier, music emotion communication suffers from ambiguity compared to speech emotion communication. Emotion in music is often the induced emotion, which is influenced by a persons bias, their mood and more. Two persons might experience highly similar levels of valence and arousal because that is expressed in the music, but different categorical emotions due to personal interpretation. The example mentioned earlier about anger and fear being closely related in both the arousal and valence plane is a good example of this ambiguity. If forced to use categorical emotions, one person might experience fear while the other experiences anger due to personal interpretation on the same music piece, while their valence and arousal estimates were highly similar.

The benefit of using a dimensional emotion space is that this ambiguity can be expressed. The emotion experienced is not forced into classes, but can be expressed using characteristics in a continuous space. Soleymani et al. (2013) showed that humans are better at judging expressed emotion in music using a continuous space than categorical labels. This is why dimensional emotion models are highly favored within MER.

Music emotion datasets

A MER model must learn a mapping from features to emotions to be able to function. This mapping can be learned through training on music samples with known emotion values. Similar to speech, there exist music emotion datasets to train MER models. Compared to SER, scientific MER datasets are more scarce due to the field being newer and smaller. Licensing is also an issue when creating MER datasets (Kim et al., 2010). Before 2010 there was a lack of freely available MER datasets, but since then the situation has improved significantly.

At the time of this research there was no publication, website or other source of information which listed available MER datasets, meaning that all datasets were found through careful use of search engines. By scouting the Internet a total of seven relevant MER datasets were found, although more likely exist. Table 2.2 shows the seven different datasets. There appears to be an even split between dimensional and categorical datasets, even though MER reviews stated that dimensional emotions were dominant within MER (Panda et al., 2015). Some of these categorical datasets obtained their emotions through mining user-generated descriptive meta-data labels from websites such as Last.fm, rather than conventional annotation gathering.

This brings up an issue that must be mentioned. The ground truth for music emotion datasets are obtained differently than for speech emotion datasets. The ground truth for a speech emotion dataset is the emotion instructed to be expressed by the actor. The quality and validity of those expressions, and thus the dataset itself, are measured using the human recognition rate.

All MER datasets use existing songs of which the intended emotions by the artist are not known. This is solved by obtaining the emotion of music through a consensus of the induced emotion experienced by many listeners. This means that many annotators are required to obtain a reliable ground truth. The validity of the annotations strongly depends on the amount of annotations, and how well the annotators are in-line with the general population regarding personal interpretation bias. A low quality MER dataset can therefore have annotation values deviating strongly from the actual general consensus, while a low quality SER dataset suffers from 'bad' recording containing not enough usable cues for the listener. This is an important distinction between the two fields and should always be taken into account.

1000 song and DEAM These datasets by Soleymani et al. (2013) and Aljanaki et al. (2017) were created for 'MediaEval', a yearly hosted benchmarking initiative for various information retrieval fields including MIR. The music was obtained from The Free Music Archive, picking the top rated songs of various genres. The annotations were done by a strictly selected pool of participants. DEAM is an expansion of the '1000 song' dataset, meaning that DEAM contains all of the '1000 song' samples along with more than a thousand new ones.

Name	Emotions	Samples	Avg anno.	Genres	Source
1000 Songs, 2013	Valence and arousal	1000	10	Various	Free Music Archive
DEAM, 2017	Valence and arousal	1802	10+	Various	Free Music Archive
AMG1608, 2015	Valence and arousal	1608	15 to 32	Contemp. pop	Copyrighted, various music labels
Soundtrack, 2010	Valence and arousal + neutral, happy, angry, sad, surprised and disgust	110	58	Soundtrack	Movie soundtracks, academic use only
ISMIR, 2012	Happy, angry, sad and relax	2904	Unknown	Various	Creative commons licenses
Emotify, 2015	9 induced emotions	400	16 to 42	Pop, electronic, rock & classical	Magnatune recording company
Jamedo, 2019	57 emotions and moods	55000	Unknown	Various	Creative commons licenses

Table 2.2: An overview of the mentioned music emotion datasets, including the average emotion annotations per sample.

AMG1608 Created by Chen et al. (2015b), the AMG dataset is the only dataset to use popular contemporary Pop music instead of license-free music. Due to licensing the audio of the music samples are not available, only the annotation values and meta-data. The songs used were selected using the mood descriptors of All Music Guide[1]

Soundtrack This dataset by Eerola and Vuoskoski (2011) was created to test the correlation between discrete and dimensional emotion annotation. Therefore it contains both valence and arousal annotations, along with six categorical emotions. What also makes it unique is that is uses samples from movie soundtracks, unlike the other datasets which use non-movie music. It also boasts the highest annotation count, but the fewest samples of all other datasets.

ISMIR2012 This dataset was created by Song et al. (2012) and includes four discrete emotions: happy, angry, sad and relax. Samples were obtained through mining Last.fm tags that include terms defined by the authors that relate to one of the four discrete emotions. The samples are available under a Creative Commons license and are mostly Pop songs. The quality of the dataset is hard to judge, as the annotations stem from user generated labels.

Emotify Created by Aljanaki et al. (2016), Emotify focuses specifically on induced emotions, using a video game to collect annotations. It is relatively small compared to the other datasets as it only contains 400 samples spread over nine categorical emotions.

Jamedo Bogdanov et al. (2019) created Jamedo, the largest dataset with a staggering 55.000 samples. It is 18 times larger than the second largest dataset. It also uses 57 categorical emotion and mood labels obtained from Last.fm tags, similar to ISMIR2012. It served as a MER challenge during MediaEval 2019. The best performing model only reached a precision of 0.2 and recall of 0.4, suggesting that this dataset is a challenging learning task for a MER model.

[1] https://www.allmusic.com/

Compared to the speech emotion datasets there appears to be much more variation regarding size, genres of music, method of gathering the ground truth and of course the emotion space used. While there appear to be numerous categorical datasets, which allow for classification rather than regression, the question is if those datasets allow for the creation of a generalizable MER model due to the present ambiguity in music emotion recognition. This is an important aspect to consider when deciding on which MER dataset to use, which we will come back to later.

Produced models and related studies

The field reviews by Kim et al. (2010), Eerola and Vuoskoski (2013) and Yang et al. (2018) offer a broad overview of research done with the field of music emotion recognition up until 2018. The contents of these reviews are summarized in this section, which should be enough to getting a basic understanding on how the field developed and how different types of models and features perform. Readers that are curious and wish to learn more about the field can find more interesting information in the aforementioned reviews.

One of the earliest MER works was done by Li and Ogihara (2003), using a SVM-based model. It was tasked to classify 30-second song excerpts spanning Ambient, Classical, Fusion and Jazz music using 13 different labels related to mood and emotion. It achieved an accuracy of 45%.

While more categorical MER models were developed after Li's work, the vast majority of MER research instead focused on the dimensional emotion space (Yang et al., 2018). This shift started with the works of Schmidt et al. (2010) and Han et al. (2009), which used regression models to predict valence and arousal, and then map those produced values to a categorical emotion space. These models thus still produced categorical labels like earlier works, but they used a dimensional approach with valence-arousal regression instead of direct classification. Han reported an increase from 33% to 95% accuracy on a 11-way classification task with this new approach.

These results motivated other researchers to explore the MER problem completely dimensional. One of the earliest models to do so was by Yang et al. (2008). A SVM was tasked to do valence-arousal regression using the PsySound feature-set, along with spectral contrast and Daubechies wavelets coefficient histogram (DWCH) features. Its root mean squared error performance was deemed impressive considering previous categorical approaches, even though directly comparing them can be ambiguous.

The suitability of dimensional emotions models was also established by Eerola and Vuoskoski (2011). Eerola et al. showed that common categorical emotion labels and a three-dimensional annotation model were strongly correlated, and that the dimensional annotations could be used to predict the categorical emotions with high precision, but not the other way around.

Similar to speech, deep learning became a popular technique in the last few years. The application of a deep learning Gaussian Process (GP) has been explored for both regression (Markov and Matsui, 2014) and regression-classification (Chen et al., 2015a). This technique resulted in a 20% performance increase on valence prediction for regular regression, while arousal performance did not change. For the regression-classification task this technique led to an increase in accuracy of 71.3%, compared to the 63% achieved by non deep learning model.

The effectiveness of various MER features has also been studied by many. According to Yang and Chen (2012) no single type of feature (harmony, spectral, rhythm and dynamics) is able to dominantly recognize emotions in music by itself, multiple types of features are necessary. We also see that many MER works (Chen et al., 2015a; Markov and Matsui, 2014; Panda et al., 2015; Schmidt and Kim, 2011; Song et al., 2012) use a large pool of various types of features and apply feature selection to filter that set, suggesting that various combinations of music features can produce good results.

Panda et al. (2015) showed that melodic audio features are effective, both in combination with more traditional features and by itself. Vibrato and pitch-related features were the most dominant melodic features. In contrast, Schmidt and Kim (2011) showed that feature selection over a large pool of features only slightly outperformed a model using only MFCC-features. While more recent models outperform the more complex feature set used by Schmidt, this research suggests that MFCC features by itself are a good minimal set of feature to use for MER.

To conclude the music emotion recognition section, it is a younger and less developed field than speech emotion recognition, but it has been developing rapidly in the recent years. There are fewer MER datasets available compared to speech, as music licensing is often an issue when creating such datasets. Luckily the situation has been improving and a handful of publicly available categorical and dimensional MER datasets exist.

MER models use the emotion cues contained in the elements of a music piece, such as rhythm or timbre, to recognize the emotion present. However, music emotion communication is ambiguous in nature as the induced emotion can differ between individuals due to interpretation, mood and more. This ambiguity does not lend itself well to the categorical emotion space, making dimensional emotion spaces strongly favored within the MER field.

Many MER models have been successfully developed, with a shift to regression (dimensional emotion space) as it appeared to be superior to classification. Deep learning techniques have also been applied to MER tasks, resulting in better performing models. However, it might be too early to assume that they are superior to all non deep-learning models.

2.4 Blind source separation

In the MiSME recognition task the model is tasked with recognizing the speech and music emotion separately from a mixed audio signal, which is an audio signal consisting of overlapping speech audio and music audio. We have already covered speech and music emotion recognition. However, emotion recognition from a mixed signal means that there is interference of another audio source. The model thus needs to learn how to distinguish which part of the audio signal, or rather the extracted features, is related to the source it is supposed to process. This adds another dimension of complexity to the already complex task of emotion recognition.

Blind-source separation might lessen the complexity of this signal-distinguishing task for the model, which we will delve into in the Methodology chapter. This section covers the theory behind blind-source separation, what it does, along with several examples of successful application and a list of available blind-source separation tools.

2.4.1 Theory

Blind-source separation (BSS) is the act of approximating the original audio signals from a set of observed mixtures of those signals at multiple sensors, without much knowledge about the original source signals.

To phrase it more simple: BSS tries to recreate each original audio signal by learning how the audio is mixed at every sensor, for example a microphone. An inversion of the learned mixture can be applied on the audio captured at a sensor to obtain isolated versions of the audio of each source captured at that sensor. An example: applying the inverse of the mixture on the audio captured at microphone A produces isolated speech and isolated music audio versions of the mixed audio captured at microphone A. Blind-source separation thus tries to 'unmix' the mixed audio at every sensor. After unmixing the audio of all sensors, an original audio source can be recreated by combining the isolated audio of all sensors. For example, the isolated speech captured at microphone A and B after unmixing can be combined to recreate the original speech. A visual example of these steps are depicted in Figure 2.5.

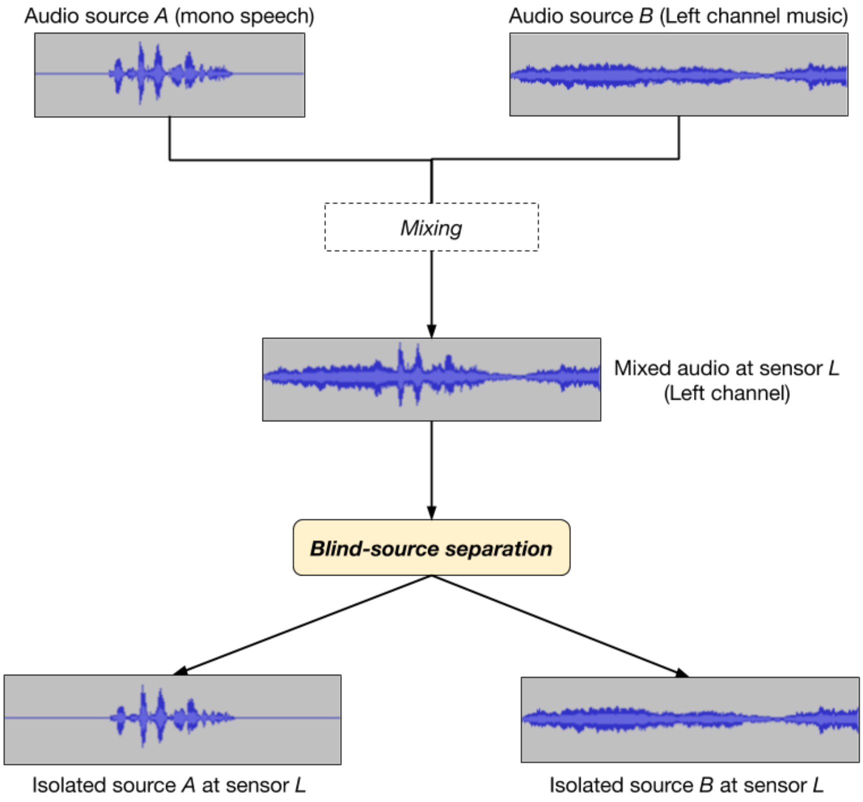

Audio source *A* (mono speech)
Audio source *B* (Left channel music)
Mixing
Mixed audio at sensor *L* (Left channel)
Blind-source separation
Isolated source *A* at sensor *L*
Isolated source *B* at sensor *L*

Figure 2.5: A visual example of how blind-source separation works

The quality of the reproduced audio through BSS is highly dependent on how accurate the calculated mixture is to the actual mixture. Differences between the two results in either audio data from the other source leaking into the isolated audio, or audio data from the correct source being filtered out.

The difficulty of a blind-source separation task is defined by a couple of aspects. The first is that the problem can be *over-determined*, *determined* or *underdetermined* depending on the number of audio sources to sensors. Overdetermined is the easiest, as there are more sensors than audio sources. This means that there are more information sources about the audio (sensors) than actual audio sources. Determined is when they are equal. Underdetermined is when there are less sensors than audio sources, which is the hardest to solve. Speech-music separation is often seen as underdetermined (Grais and Erdogan, 2011).

Another aspect defining the difficulty is if the mixing is *instantaneous* or *convolutive*. Convolutive means that the mixture is affected by time. This could be the case when audio is not captured at the same time at each sensor, for example due to reverberation or strong difference in spatial placement of the sensors. Convolutive BSS is harder to solve than instantaneous.

Finally, the last aspect that defines the difficulty is if it is *time-variant* or *time-invariant*. It is time-variant if the mixture can change over time, for example if speakers move in the audio space. This is much harder to solve than time-invariant, where the mixture is static for the entire duration of the audio.

2.4.2 Research on BSS

Many have studied blind-source separation problems (Demir et al., 2010, 2012; Grais and Erdogan, 2011; Grais et al., 2014; Jansson et al., 2017; Luo et al., 2017). Understanding these works requires significant technical expertise and are generally too complex to cover in this research. However, their results show that blind-source separation can be successfully applied to various scenarios. For example, Xu et al. (2014) showed that BSS of the background music and the singing voice, combined with late-fusion, improved the accuracy of their MER model from 37% to 53% on a 4-emotion categorical classification task.

Blind-source separation tools

There exist a handful of blind-source separation tools, which are generally easy to use. We identified a total of six scientific BSS tools, excluding commercial tools.

Wave-U-Net Created by Stoller et al. (2018), Wave-U-Net is an end-to-end approach trained BSS-tool that uses long temporal context by repeatedly down sampling and applying convolution to the feature maps to combine high and low-level features at different timescales. The model is trained for vocal separation in music tracks.

FASST FASST is a C++ based source separation toolbox created by Salaün et al. (2014). It includes various BSS algorithms limited to the time-frequency domain. User scripts can be written in Matlab and Python. It can be used for various BSS tasks.

Untwist Untwist (Roma et al., 2016) is a Python-based source separation framework. Untwist serves as a framework for the entire source separation pipeline, in which new algorithms can be implemented easily. It supports algorithms that function in the Short Term Fourier Transformation or Quadratic ERB space.

Nussl Created by Manilow et al. (2018), Nussl is another Python-based source separation framework, specifically focusing on underdetermined problems. Sixteen different source separation algorithms are included in Nussl, including ICA, RPCA, NMF with MFCC clustering, Deep clustering and more.

Open-unmix Open-unmix (Stöter et al., 2019) was developed specifically to provide (unexperienced) researchers with an easy to understand framework that included state-of-the-art BSS models, as other frameworks required in-depth knowledge to be used according to the authors. It uses a bi-directional LSTM model based on the work of Uhlich et al. (2015), and its performance on a instrument segmentation task was the highest of all open-source models at the time of its publication, only beaten by one non open-source algorithm called TAK1.

Spleeter Spleeter (Hennequin et al., 2019) is the most recently published BSS toolkit. It was presented at the ISMIR2019 and it is Deezer's source separation library, which has been made publicly available for use. It is a Python-based toolkit that uses Tensorflow for its models. The model is a encoder/decoder CNN with skip connections. It was trained using Deezer's private music dataset, which consists of licensed music. The creators claim that Spleeter slightly outperforms Open-unmix, while being significantly faster to a degree where Spleeter could be run in realtime.

To conclude the blind-source separation section: blind-source separation is the act of 'unmixing' a mixed audio signal by applying an inversion of the calculated mixture at each sensor. The quality strongly depends on how accurate the calculated mixture is to the actual mixture. Blind-source separation has been successfully applied to many scenarios. There also exist a handful of BSS tools which are available for scientific research.

Name	Environment	Algorithm description	Notes
Librosa (Brian McFee et al., 2015)	Python	Spectral features (MFCC, ZTC, Chroma) + Rhythm (Tempogram)	
PsySound3 (Cabrera et al., 2008)	Matlab	Cepstrum, loudness, pitch, roughness, FFT spectrum	
Praat (Boersma and Weenink, 2007)	Standalone (C++)	Spectograms, pitch, formant, intensity, jitter, shimmer, cochleagram, excitation	Speech focused
MIRtoolbox (Lartillot et al., 2008)	Matlab	Various dynamics, rhythm, timbre, pitch and tonality based features.	
openSMILE (Eyben et al., 2013)	Standalone	MFCC, voice quality, Chroma fetures, pitch, loudness, energy, formants, LPC	High feature count
Audio Toolbox	Matlab	Various spectral features, MFCC and pitch	Included in Matlab
Python-speech-features	Python	MFCC, F-bank, log-bank and spectral subband centroid	
Parselmouth (Jadoul et al., 2018)	Python API	None	Python API for Praat
Essentia (Bogdanov et al., 2013)	C++	Many spectral (MFCC, flux, MelBands, Roll-off) and rythm features	High feature count

Table 2.3: An overview of feature extraction tools available for both speech and music feature extraction

2.5 Background on tools, techniques and methods

In this section we will provide some background on the tools, techniques and methods used in this research. The goal is to familiarize readers with these tools and more, while the Methodology chapter explains why they were chosen and how they were put to use.

2.5.1 Feature extraction tools

A large number of feature extraction tools exist that can be used for either speech or music emotion recognition. We identified nine in total, which are depicted in Table 2.3. Of these nine *openSMILE* was used for speech feature extraction, and *Essentia* for music feature extraction. Section 3.4.1 and 3.5.1 cover these decisions in more detail.

OpenSMILE was used with the 'emobase2010' configuration. The 'emobase2010' configuration is a tweaked version of Paralinguistic Challenge feature set (Schuller et al., 2010), which itself is also available as the 'IS10' configuration. 'emobase2010' boasts a total of 1582 common SER features, such as MFCC and Fundamental frequency-features. A list of all feature types included in the configuration can be found in Table 2.4. Only the types of features and their feature count are shown, as this had to be extracted directly from the output files. Each feature type can consist of one or more feature classes, for example the same feature type but on different scales. These feature classes can have multiple derivatives, for example the average, the standard deviation, the minimum value and more. Each derivative counts as a feature. Please see Eyben et al. (2013) for a more detailed overview.

For music feature extraction Essentia was used, using the precompiled 'essentia_streaming_extractor_music' executable. Essentia includes a large amount of common MER features, more than most toolkits. The tool was specifically developed for music information retrieval. The executable that was used produces a total of 2651 common MER features per sample. The feature types, classes and total count per feature type of this feature set are shown in Table 2.5.

Feature type	Features
Mel frequency cepstrum coefficients	630
Log mel frequency band	336
Line spectral pair frequencies	336
Fundamental frequency	82
Jitter	76
Loudness	42
Voicing	42
Shimmer	38

Table 2.4: All feature types included in openSMILE's 'emobase2010' configuration

2.5.2 Models

Four different models were developed for both the speech and music emotion recognition experiments. These are a *Support Vector Machine* (SVM), a *Random Forest* (RF), a *Multilayer Perceptron* (MLP) and a *Deep Neural Network* (D-NN). For the speech emotion recognition experiments all models take the form of a classifier, which produce categorical emotion labels as output. For the music emotion recognition experiments all models take the form of a two-value regressor instead, producing two numerical outputs representing the valence and arousal.

The first three models were built using 'scikit-learn', a popular Python-based machine learning library. The D-NN was built using 'Keras', a popular deep learning library also in Python. Some background on each model is given below, explaining how they function and showcasing their differences to readers unfamiliar with these types of models.

Support Vector Machine

A Support Vector Machine (SVM) uses hyperplanes in the feature vector space differently for classification and regression. For classification it uses hyperplanes to split all data points (feature vectors) into the possible classes. These hyperplanes serve as decision boundaries, defining if an input vector belongs to one class or an other depending on which side of the hyperplane it lies. The linear split quality of a hyperplane for classification is measured by summing the distance of the closest sample of each class to the hyperplane itself, where a larger distance is better. For regression it uses the hyperplanes as an function to estimate the regression value, where the goal is to find a curve (dictated by the hyperplanes) that minimizes the deviation of all data points to it. A lower deviation means that the regression functions lies closer to all data points on average, meaning that the values produced by the function are more accurate.

By default these hyperplanes split the space linearly, but in most cases the data is not linearly separable. This is often solved by applying a kernel that maps all data points (features) non-linearly to a new space. The model then tries to find suitable hyperplanes in this non-linear space. A visualization of mapping these hyperplanes of non-linear space back to linear space is depicted in Figure 2.6.

Random Forest

Random Forest models are based on decision trees. In a decision tree the input is iteratively passed to either the left or right leaf based on if the input (features) meets a certain condition. After a certain number of splits a dead-end is reached, which has a categorical label when the Random Forest is a classifier, or a numerical value when it is a regressor. An example of a simple decision tree is depicted in Figure 2.7.

Type	**Feature class**
Loudness [1]	average_loudness
Complexity [1]	dynamic_complexity
Silence [27]	silence_rate_20dB, silence_rate_30dB, silence_rate_60dB
Spectral [252]	spectral_rms, spectral_flux, spectral_centroid, spectral_kurtosis, spectral_spread, spectral_skewness, spectral_rolloff, spectral_decrease, spectral_strongpeak, spectral_energy, spectral_energyband_low, spectral_energyband_middle_low, spectral_energyband_middle_high, spectral_energyband_high, spectral_entropy, spectral_complexity, spectral_contrast_coeffs, spectral_contrast_valleys
Barkbands[288]	Barkbands, barkbands_crest, barkbands_flatness_db, barkbands_kurtosis, barkbands_skewness, barkbands_spread
Melbands [405]	Melbands, melbands128, melbands_crest, melbands_flatness_db, melbands_kurtosis, melbands_skewness, melbands_spread
Erbbands [405]	erbbands, erbbands_crest, erbbands_flatness_db, erbbands_kurtosis, erbbands_skewness, erbbands_spread
Other [720]	mfcc, gfcc, dissonance, pitch_salience
Rhythm [121]	beats_count, beats_loudness. beats_loudness_band_ratio, bpm_histogram_first_peak_bpm, bpm_histogram_first_peak_spread, bpm_histogram_first_peak_weight, bpm_histogram_second_peak_bpm, bpm_histogram_second_peak_spread, bpm_histogram_second_peak_weight, onset_rate, danceability
Tonal [413]	hpcp, thpcp, hpcp_entropy, hpcp_crest, key_temperley, key_krumhansl, key_edma, chords_strength, chords_histogram, chords_changes_rate, chords_number_rate, chords_key, chords_scale, tuning_frequency, tuning_diatonic_strength, tuning_equal_tempered_deviation, tuning_nontempered_energy_ratio

Table 2.5: A list of all features included in Essentia's 'essentia_streaming_extractor_music.exe' (Based on the output file)

A Random Forest consists of a large number of decision trees that form an *ensemble*. All decision trees in the ensemble make a prediction based on the same feature vector, and the most predicted, or mean output becomes the prediction of the ensemble.

Multilayer perceptron

The multilayer perceptron (MLP) is a type of deep artificial neural network. Similar to all other neural networks, it consists of at least 3 layers: the input layer, one or more hidden layers and the output layer. Each node in the hidden and output layers are perceptrons. The perceptrons use non-linear activiations functions, allowing the model to make non-linear separations because at least two layers of the model always consist of perceptrons. These non-linear separations define to which class the given sample belongs when it is a classifier, and to which numerical value it belongs when it is a regressor. MLPs are always feed-forward, unlike other neural networks. This means that the output of one layer only affects layers deeper in the model, not earlier layers or itself.

Each node (perceptron) in the hidden and output layer are linear classifiers that multiply their input x, in our case the feature vector, by a set of weights w and add a bias b. The result is passed through a nonlinear activation function φ to produce a single output. The function of a single perceptron can be written as follows:

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) \tag{2.1}$$

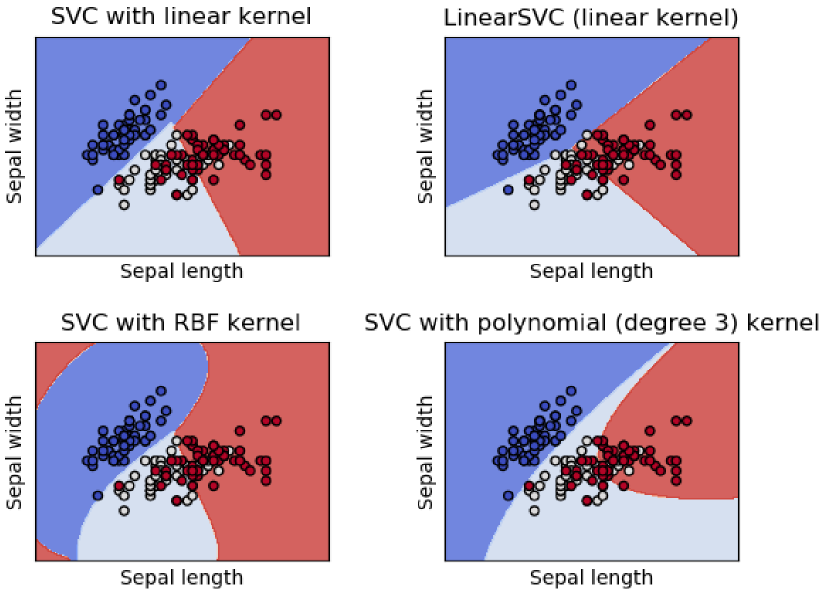


Figure 2.6: State Vector Classifier hyperplanes depicted in four different spaces using the Iris Flower dataset. Each of the four visualizations shows the use of a different kernel, which map the data points and the hyperplanes used for class segmentations into a different space. (Scikit-learn, 2007)

Deep neural network

A network is *deep* when it consists of more than one hidden layer. This means that the MLP can also be considered a D-NN if it has more than one hidden layer. The advantage of a 'deep' model is that each hidden layer transforms the input in a more abstract format usable by the next layer, allowing the model to perform different levels of abstraction. This can be beneficial when learning the correct mapping from input to prediction. Our D-NN uses multiple hidden layers, but no special learning techniques. The MLP uses a single hidden layer, as otherwise it would be too similar to the D-NN model.

Deep neural networks can take many forms. A model can be supervised, semi-supervised or unsupervised and there exist many learning architectures. Most of these aspects are too complex to

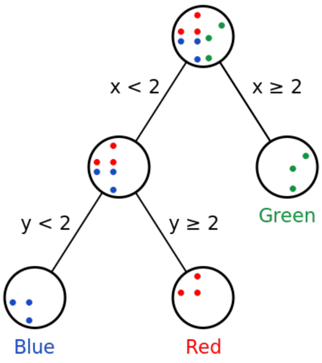


Figure 2.7: A simple decision tree (Victor, 2019)

cover here. We recommend reading the work by LeCun et al. (2015) for more in-depth information.

2.5.3 Blind-source separation

Spleeter was used as the blind-source separation component of the MiSME system. We used the pre-trained 2-stem model, which does vocal-accompaniment separation. The accompaniment is all other musical sources except vocals. This lends itself well to speech-music separation.

This concludes the literature review and background on tools, techniques and methods used. In the next chapter we will go over how the MiSME system, the mixed-audio dataset, the experiments were designed.

Chapter 3

Methodology

We established the goal of this research in Section 1.4, which is to see if a ‘functioning’ MiSME system can be developed, along with additional goals exploring the MiSME problem space. This chapter covers how the MiSME system, the mixed audio dataset and all experiments were created.

The methodology chapter is divided in six sections. First we explain how the MiSME system works conceptually (Section 3.1). This is followed by the datasets used and created (Section 3.2) and the experiments used to test and study the MiSME system (Section 3.3). The final three sections each cover the speech, music and blind-source separation parts of the MiSME system (Section 3.4, 3.5 and 3.6).

3.1 MiSME system design

We adress the MiSME system often as a single entity. However, as explained in Section 1.3, the system has to be capable of both speech and music emotion recognition. This means that it should consist of at least two models, a separate speech emotion recognition model and music emotion recognition model. These two models together allow the MiSME system to produce the required output. Creating one computational model which can do both is another non-trivial challenge better left for future research.

We identified two different approaches of developing the models of the MiSME system. The first option is by training them the traditional way, where both the speech and music model have a feature extraction component. This component extracts suitable features from the audio signal, which are passed to the model to predict the emotion present.

The other approach is to train the models in an ‘end-to-end’ fashion. Instead of relying on extracted features, the models are given the raw audio signal and must learn to do emotion recognition from the audio signal rather than descriptive features. The model is often adapted in such a way that it is capable of transforming the audio in various ways akin to feature extraction, but it must learn how to do this during training. This end-to-end approach adds more complexity to the learning task, because the model must learn how to obtain useful features from the raw audio alongside learning the mapping from the features to emotions.

The decision was made to create the models using the traditional approach with separate feature extraction. While end-to-end models showed potential in the SER field (Trigeorgis et al., 2016), they still underperform compared to the traditional approach. It is not worth it in our opinion considering the fact that the end-to-end approach increases the complexity even further on top of the mixed audio problem, and is therefore better explored in future work.

This brings us to the design of our MiSME system. The system consists at least of four components: *a speech feature extractor*, *a speech emotion recognition model*, *a music feature extractor* and

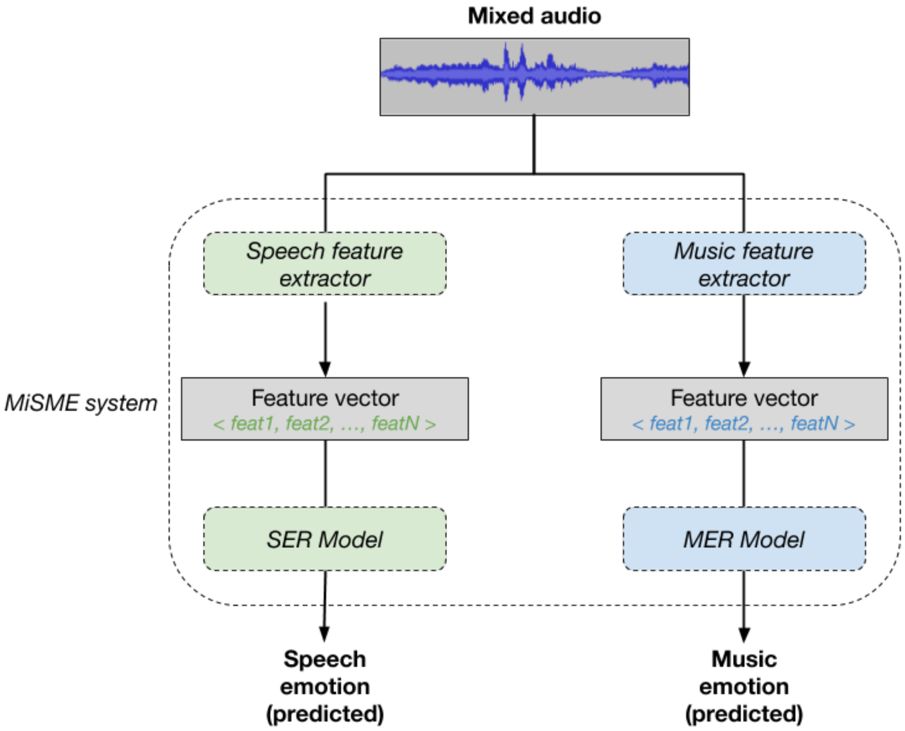


Figure 3.1: The flow of the MiSME system without blind-source separation

a music emotion recognition model. With these four components the MiSME system can output separate speech and music emotions for any given audio signal. However, a fifth component is added when blind-source separation is included in the MiSME system. This fifth component is of course *the blind-source separation component*.

Blind-source separation is included in some version of our MiSME system because we believe that it might be beneficial for the speech and music models. As explained in Section 1.3 and 2.4, blind-source separation tries to 'unmix' a mixed signal, producing isolated approximations of each original audio source seen in the mixed signal. While they are not perfectly accurate, the speech and music models might perform better on these isolated approximations of the speech and music than the mixed audio. Extracted features are (likely) less influenced by the other audio source in these blind-source separated signals, making the features better represent the audio source which the model must process. This decreases the complexity of the recognition task compared to no blind-source separation. Blind-source separation might drastically reduce the degree to which features are affected, and thus it is not far fetched to assume that blind-source separation can be a beneficial preprocessing step.

A visualization of the MiSME system with and without blind-source separation are depicted in Figure 3.1 and Figure 3.2. In the MiSME system without blind-source separation the mixed audio is passed to both the speech and music feature extractor. These each produce a feature vector, one containing all speech features and the other containing all music features extracted from the mixed audio. These vectors are then passed to their respective emotion recognition model as input. From these feature vectors the emotion recognition models produce the final output, an emotion prediction. When blind-source separation is included in the MiSME system, the same happens except that the mixed audio is split into isolated speech and music audio by the blind-source separation component. These isolated audio segment are then fed to their respective feature extractor, instead of the mixed audio.

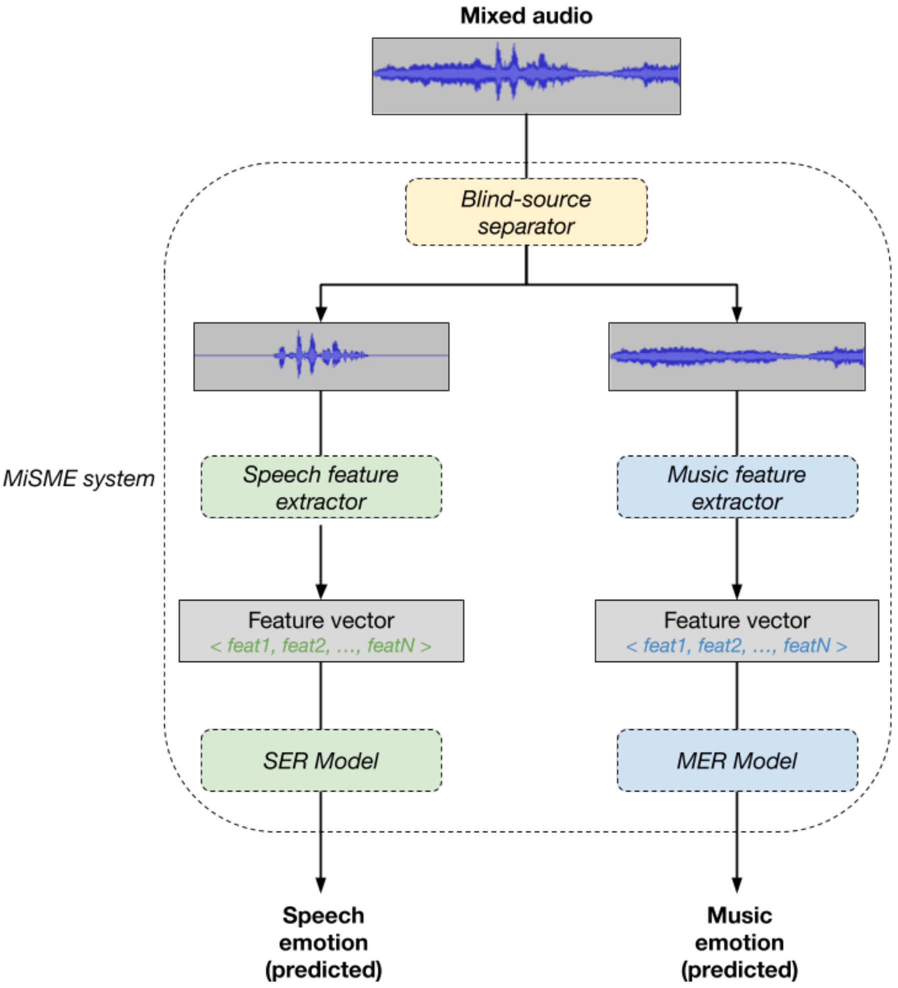


Figure 3.2: The flow of the MiSME system with blind-source separation

3.2 Data

Suitable data is required to properly train the speech and music emotion recognition models. As already mentioned in Section 1.3, no scientific mixed speech-music audio dataset with emotion annotations could be found at the time of this research. This meant that one had to be created.

Recording and annotating a MiSME dataset from scratch was deemed infeasible for the scope of this master thesis research. This meant that the next best scientifically valid option was to create such a dataset by blending samples of existing speech and music emotion recognition datasets. A coinciding advantage of this is that a contextual relationship between the speech and music is avoided because the speech and music are not related, as they stem from separate datasets. A dataset produced this way thus adheres to our decision to exclude the contextual relationship between speech and music from this research, as explained in Section 1.4.

For the speech emotion samples RAVDESS (Livingstone and Russo, 2018) was used. It contains 1440 samples spanning eight different categorical emotions, two different emotion intensities and twenty-four actors. It is a middle ground regarding complexity and difficulty considering all of the SER datasets mentioned in Section 2.2.2. RAVDESS has a human recognition rate of 62.5%. State-of-the-art models achieve accuracy scores between 64% to 75%[1] (Bhavan et al., 2019; Zeng et al., 2019). Overall it is a robust and suitable dataset, offering enough samples and complexity for our experiments without compromising on human recognition rate.

[1] These results were achieved in a speaker-dependent experiment setting

RAVDESS uses a total of eight categorical emotions: *neutral, calm, happy, sad, angry, fearful, disgust and surprise.* These are Ekman's primary emotions in addition to 'neutral' and 'calm'. All emotions except neutral are expressed at two different intensities within the dataset. The fact that RAVDESS uses categorical emotions means that the speech recognition task becomes a 8-way classification task. The speech emotion model has to predict the correct emotion from eight possible emotions.

For the music samples two music emotion recognition datasets were combined. These are Soundtrack and DEAM (Aljanaki et al., 2017; Eerola and Vuoskoski, 2011). They were combined to make it more similarly sized to RAVDESS. This means that our music dataset contains music created for various purposes, as DEAM consists of music from various genres and Soundtrack of movie soundtracks.

Combining them was possible because both datasets are not only annotated in the same dimensional valence-arousal space, but they also use the same value range. This means that the music emotion recognition task becomes a 2-value regression task. The music emotion recognition model must predict the valence and arousal values as accurately as possible, rather than a single categorical emotion label.

It is important to note that a trimmed version of the DEAM dataset was used. Many samples of DEAM contained vocals or human speech, which is fairly normal in music. However, the presence of vocals increases the complexity of both the blind-source separation and speech emotion recognition task. Mixed audio samples will be created by mixing a music sample with a speech sample, which we will cover later. If the music sample contains vocals, there are now multiple speakers present in the mixed audio sample. The blind-source separation component and the SER model now need to distinguish between which vocal-sounds are related to the speech and which are related to the music. We acknowledge that this can occur in certain MiSME use cases, but this increases the complexity significantly. Exploring this non-trivial problem is better left for future research.

To avoid this problem all samples containing vocals in the DEAM dataset were manually filtered out. We define vocals as singing, speech and human vocals being used as beat-samples (hip-hop). After filtering out all samples containing vocals, 948 of the 1802 DEAM samples were deemed suitable for use. Combined with Soundtrack the music dataset has a total of 1058 samples.

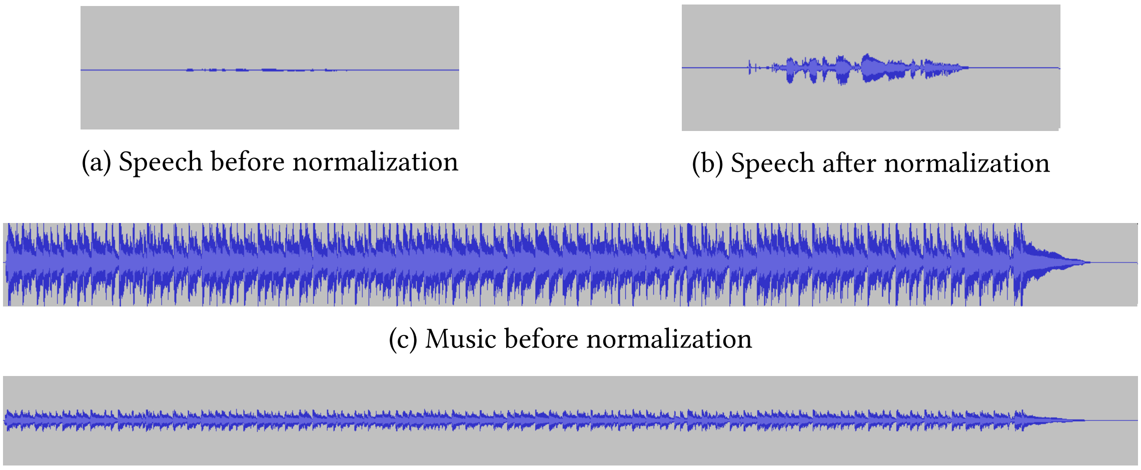
3.2.1 Mixed sample creation

Mixed samples can be created by mixing a speech sample from RAVDESS with a music sample from either DEAM or Soundtrack, producing a new audio sample containing both speech and music. There might exist use cases in which one sound-source is mixed in (perceptually) louder than the other, or where there is a difference in spacial or temporal mixture per sound-source. We simplify it to a scenario where the speech and music are mixed in equally as loud over both channels without any spacial or temporal difference. This means that both the speech and music are perceptually hear-able and centered.

Loudness normalization

The original samples from the three datasets can not be used for mixed sample creation as is. Generally speaking the speech recordings are much lower in volume than the average music sample, but there is also strong variation in perceptual loudness difference between the speech and music samples themselves. Mixing speech and music samples equally will not suffice in this situation as this would lead to samples with extreme deviation in speech-to-music loudness difference.

To allow for equal mixing, all speech and music samples are normalized before mixing using the EBU R128 standard (EBU-Recommendation, 2011). EBU R128 is the normalization standard used



(a) Speech before normalization

(b) Speech after normalization

(c) Music before normalization

(d) Music after normalization

Figure 3.3: A visual example of R128 normalization

by many EU broadcasting companies and it uses perceived loudness rather than actual loudness (decibel). Applying the normalization makes each speech and music sample about as equally loud for the average human ear. This means that R128 normalization makes equal mixing possible, because each speech and music sample are now almost equally as loud and the loudness difference between the original samples is largely gone.

However, perceived loudness is seen as a strong predictor for arousal according to Olsen et al. (2015). A small experiment was done beforehand to test if R128 normalization indeed negatively affected emotion recognition due to loss of perceived loudness difference. The use of R128 audio was justified based on the results, which can be found in Section 4.2.1 and 4.3.1. The normalization did not lead to significantly worse performance on both speech and music emotion recognition. To be precise, we saw a negligible decrease in performance on the SER task and an increase in performance on the MER task.

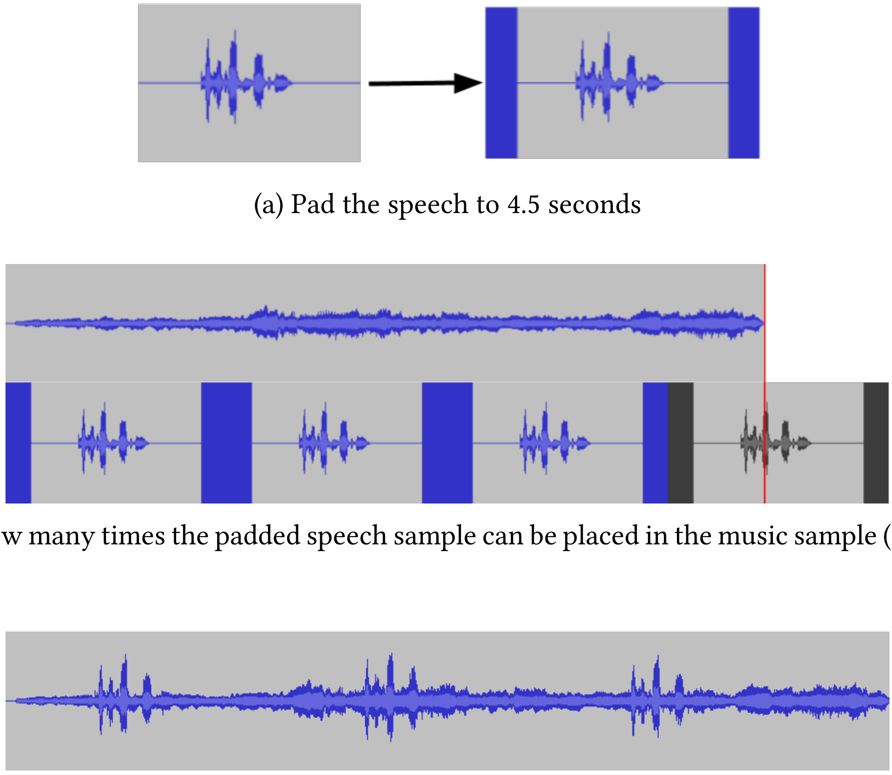
R128 normalization was done using FFMPEG-normalize[2]. It includes a 'dual-mono' mode which compensates for the increase of 3 LUFS (Loudness Unit Full Scale) when a single track mono file is played stereo. All RAVDESS samples are mono, while DEAM and Soundtrack samples are stereo. This 'dual-mono' mode was thus necessary when normalizing the RAVDESS samples, because they become dual-channel mono after mixing. Also, the samples from RAVDESS were resampled from 48khz to 44.1khz using SoX[3] to make them consistent with the sample rate of the music samples.

Mixing speech and music samples

While the R128 normalization solves the 'equal-mixing' problem, there is another issue that needs to be adressed, namely the difference in duration of speech and music samples. The speech recordings of RAVDESS are generally between three and four seconds in duration. However, each recording starts with a short moment of silence, followed by the utterance and another short moment of silence. The actual speech is only between one to one-and-a-half second long. The music samples from Soundtrack range between 11 and 27 seconds in duration, and the samples from DEAM between 44 seconds and almost 9 minutes. The average duration over all music samples is 45.9 seconds. There is thus a large difference between the average length of actual speech in a speech sample and the average length of a music sample.

[2] https://github.com/slhck/ffmpeg-normalize

[3] http://sox.sourceforge.net/Docs/Documentation



(a) Pad the speech to 4.5 seconds

(b) Calculate how many times the padded speech sample can be placed in the music sample (three successful placements)

(c) Blend the speech placements in the music sample

Figure 3.4: A visual example of how mixed samples are created

Creating a mixed sample by just mixing the speech sample into the music sample leads to mixed samples where only a small portion of the audio is affected by actual mixed audio presence. Based on the average duration of a speech and music sample it would mean that less than 2.5% of the audio is actually affected by mixed audio presence. This is such a small portion that it likely does not stress the MiSME recognition ability of the system enough. The music model might be able to work around that small portion of affected audio, treating most of the audio as just plain music audio instead of mixed audio.

A higher portion of mixed audio presence thus required to sufficiently study the MiSME recognition task. To increase this portion the speech samples are mixed in the music samples at centered 4.5 second intervals, a slightly bigger window than the longest speech sample. A visual example can be found in Figure 3.4. We see that a speech sample is first padded to 4.5 seconds in length, after which the number of fits within the music sample are calculated. In the example three of the four placements fully fit, meaning that the speech is mixed in the music sample three times. The number of speech placements can easily be calculated using the following formula that uses just one modulo operation:

$$P = D_m \bmod 4.5 \tag{3.1}$$

where:

$P \quad = \textit{Number of speech placements}$
$D_m = \textit{The duration of the music sample in seconds}$

This increases the portion of audio affected by mixed-source audio significantly. On average the portion of audio affected by speech-music overlap in a mixed sample now becomes around 25%, instead of 2.5%.

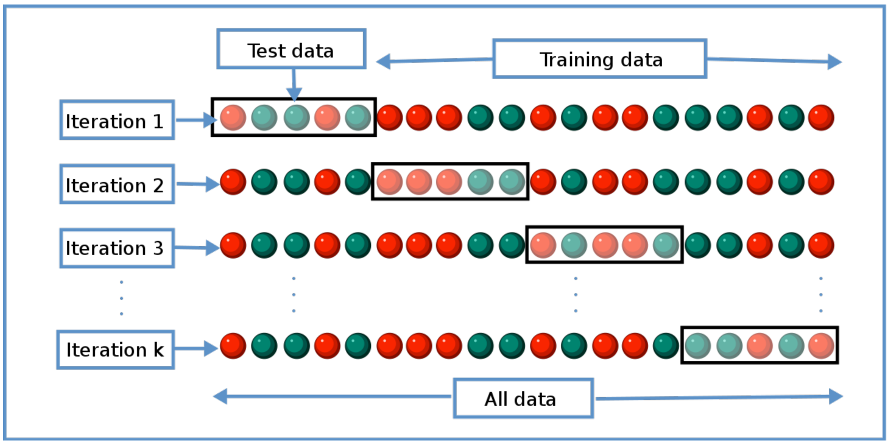


Figure 3.5: A visual example of how k-fold cross validation works (Wikipedia, the free encyclopedia, 2019)

This solution affects both the speech and music emotion recognition tasks. For music emotion recognition the entire mixed sample can be used, extracting features from the entire mixed sample with an average speech-music overlap of 25%. For the speech emotion recognition task one of the speech placement windows is used. Which one of the possible N placements is decided randomly to avoid bias towards certain sections of songs. This means that speech features are extracted from one 4.5 second snippet of the mixed sample. The speech emotion model then uses these extracted features to predict the speech emotion.

All mixed samples were created using a python package called 'pydub'. Both the speech and music sample were mixed unaltered, so no change to their volume or mixture of the left and right audio channel. The mixed sample were saved in a .wav format at 44.1khz.

Dataset creation

We are now able to produce usable mixed sample thanks to R128 normalization and the repeated placement of the speech sample. In an optimal scenario the mixed-audio dataset would contain a mixed sample of all possible speech-music combinations. This would result in a total of 1,523,520 samples. Producing, storing and using more than one-and-a-half million mixed samples in the experiments is unfeasible for the scope and available resources of this research. A mixed audio dataset of a smaller size is thus required.

Mixing every speech sample once with a random music sample is not a suitable solution, it is prone to randomness and could result in an unbalanced dataset. Therefore a solution was applied that falls between the two extremes. Each speech sample is mixed with five random music samples, producing five mixed samples per speech sample for a total of 7200 mixed samples.

But there is one issue which must be adressed to prevent data-leakage. Cross validation is common practice when developing emotion recognition models. The dataset is split up into several folds, where the model is trained on all folds except one, which is used for testing the performance. This is repeated until every fold has been used for testing once, meaning that every possible train-test combination has been covered. The performance metrics over all of these iterations are averaged to obtain a better estimate of the models performance compared to no cross validation.

The decision was made to tackle the speech emotion recognition task in the MiSME problem as speaker-independent. This means that the model is not allowed to see samples of an actor during training and testing, all samples of an actor may only appear in one of the two sets. Otherwise the SER model becomes speaker-dependent, because it can adapt to the speech pattern of that actor

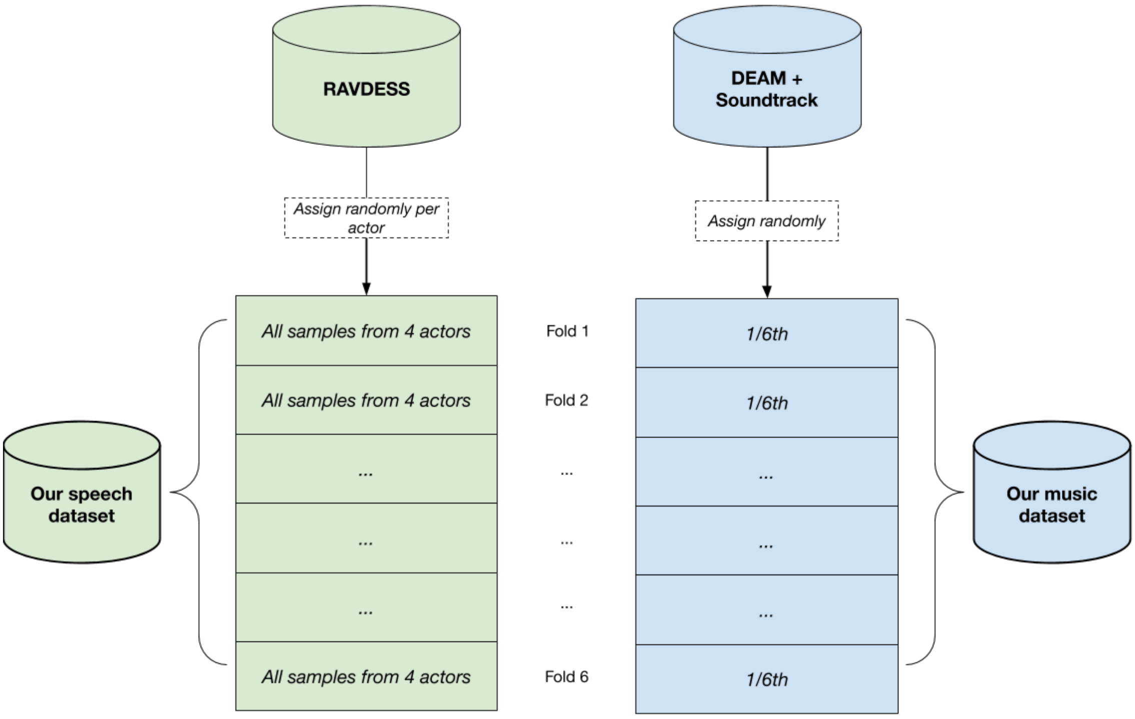


Figure 3.6: A visual example of how the datasets are divided in six folds

and achieve possibly better performance due to that advantage.

Speaker-independent cross validation can easily be achieved by placing all samples of an actor in the same fold. We chose for 6-fold cross validation, meaning that each fold contains all samples from four unique actors. However, mixed samples are created by mixing each speech sample five times with a randomly chosen music sample. Speaker-independence is assured by keeping the mixed samples in the same fold as the speech sample fold, so all mixed samples of an actor appear in the same fold. To also avoid the same dependence problem for music, the music samples are also divided into six folds. By creating mixed samples from speech and music of the same fold, and placing them in the same fold in the mixed dataset, both speaker-independence and music-independence is ensured because each actor and music sample only appears in the same fold over all datasets.

A visual example of these solutions are depicted in Figure 3.6, along with Figure 3.7 showing that only mixed samples are created from the same folds. In Section 4.1 the resulting dataset and its distribution in the valence-arousal space are shown, showcasing its validity.

The mixed audio dataset can now be created. However, we not only test on mixed audio in this research, but also blind-source separated audio as stated earlier. The blind-source separated audio dataset is an exact copy of the mixed audio dataset, only with the correct isolated audio segment produced by the BSS algorithm, which is covered in Section 3.6, instead of the mixed sample. So each full-length mixed sample used for music emotion recognition becomes the isolated music version of the full-length mixed sample, and the speech sample becomes the isolated speech version of the predetermined 4.5 second segment of the mixed audio sample. This 4.5 second segment is one of the speech placements, as explained earlier.

To summarize how all of the data was created for this research, mixed audio samples are created by mixing a speech and a music sample equally regarding loudness and mixture. To ensure that both sources are perceptually hear-able R128 loudness normalization is applied beforehand. To increase the speech-music overlap from 2.5% to about 25%, the speech is mixed in the music using 4.5 second intervals. The music emotion recognition task is done on the entire mixed audio sample,

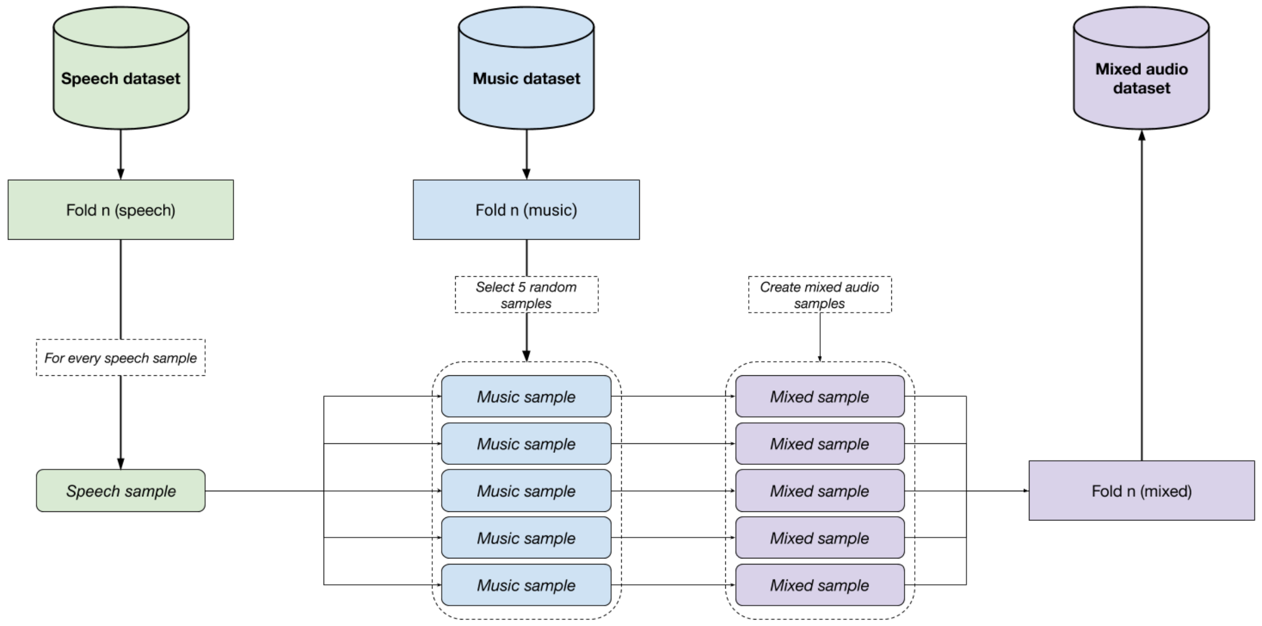


Figure 3.7: A visualization of how the folds are sampled to create the mixed audio dataset - Note: only speech and music samples from the same fold are used

while speech emotion recognition is done using one of the speech placements, picked randomly. To ensure proper speaker and music-independence, the speech and music samples are divided into six folds before mixing. Mixed samples can only be created from speech and music samples of the same fold. As a middle ground, each speech sample is mixed with five different music samples, resulting in a total of 7200 unique mixed audio samples. A blind-source separated copy is made of this mixed audio dataset, where the samples are replaced by the isolated speech or music approximations produced by the blind-source separation algorithm.

3.3 Experiment setup

As identified in the previous section, the MiSME system consists of two emotion recognition systems, a speech emotion recognition system and music emotion recognition system. They each have a feature extractor and emotion recognition model. These must be tested and evaluated separately. Common practices from both fields of study can be used to do this, albeit with some adaptation to the MiSME problem space.

The common practice we use is k-fold cross validation, where one part of the dataset is used for testing and the rest for training. This is repeated until every part has been the testing set once. The performance of all iterations is then averaged to obtain the actual performance of the model. See Figure 3.5 for a visual example of cross validation.

As explained in the previous section, we have three version of the speech and music datasets: *a clean single-source version*, *a mixed-audio version* and *a blind-source separated version*. They are all divided in the same six folds, containing the same samples. Various scenarios can be created by using different audio copies of each fold for training and testing. For example, a cross-validation experiment where the training is on speech-only audio, but the testing is on mixed audio. With this method the MiSME recognition capabilities of the MiSME system are tested.

There are a total of six experiments that test the capabilities of the MiSME system, allowing us to answer the main research question. The speech and music variations of these six experiments are strongly similar but there exist some minor differences, which are covered in their separate subsections. Table 3.1 shows all six experiments for both the speech and music models of the

Experiment	Trained on	Tested on
A.	Speech-only	Speech-only
B.	Speech-only	Mixed
C.	Speech-only	BSS-speech
D.	Mixed	Mixed
E.	Mixed	BSS-speech
F.	BSS-speech	BSS-speech

(a) Speech experiments

Experiment	Trained on	Tested on
A.	Music-only	Music-only
B.	Music-only	Mixed
C.	Music-only	BSS-music
D.	Mixed	Mixed
E.	Mixed	BSS-music
F.	BSS-music	BSS-music

(b) Music experiments

Table 3.1: The audio types on which the models are trained and tested in the six main experiments

MiSME system. They are all possible combinations of training and testing on either single-source, mixed or blind-source separated audio. We use the terms ‘BSS-speech’ and ‘BSS-music’ to reference which of the two outputs of the blind-source separation algorithm are used. ‘BSS-music’ is the output labeled as ‘accompaniment’ by Spleeter, and ‘BSS-speech’ is the output labeled as ‘vocals’.

Experiment A serves as a baseline scenario by training and testing the models on single-source audio. The performance in this scenario could be seen ‘optimal’ because it is speech or music emotion recognition on a dataset that is not affected by mixed audio.

Experiment B and **experiment C** test the single-source trained models on either mixed or BSS-audio. This gives an indication how mixed-audio affects both speech and music emotion recognition. How well does a speech-only/music-only trained model translate to MiSME recognition? Does mixed-audio cause a degradation in performance and if so, how much?

In **experiment D** and **experiment F** the models are trained and tested on either mixed audio or blind-source separated audio. These experiments test how well MiSME-specialized models perform. The performance difference between these experiments and the three earlier experiments allow us to put the performance of a MiSME system in perspective to a non-specialized system. A logical assumption would be that these specialized models (D and F) outperform the non-specialized models (B and C) because they can adapt to the MiSME problem during learning. The performance difference between experiment D and F also shows if blind-source separation is beneficial for MiSME recognition.

Experiment E was included to cover all possible train-test combination. Training on mixed audio but testing on BSS-audio seems counterintuitive, but it might produce interesting results.

Speech experiments

There are some differences between the speech and music experiments. As mentioned in the Section 3.2, the speech emotion recognition task is an 8-way classification. This is because RAVDESS uses a total of 8 categorical emotions.

Measuring performance is fairly simple for classification. A model can either produce the correct label, or not. This means that accuracy, the percentage of correctly produced labels, becomes the main performance metric for the speech experiments. A higher accuracy means a better model generally speaking. However, the accuracy per emotion might be very different. Therefore the precision and recall per emotion are also reported.

Music experiments

The music emotion recognition task however is a 2-value regression task. Both DEAM and Soundtrack annotate the emotion using a valence and arousal value. Measuring performance of a regression task is harder as it is not a binary problem. Two models might produce different arousal values than the actual arousal value, but one might be closer to the correct value than the other. Therefore the difference between the models output and the actual value are often used, this is called the error.

Common practice is to use the root mean squared error (RMSE), which we also use. The RMSE is the main performance metric for the music emotion recognition task, as it directly measures how far off the model is on average compared to the actual values. A lower RMSE generally means a better model.

In addition to the RMSE metric, the R2 score and 2-tailed Pearson correlation (PCC) test are also included. This is because the RMSE is a more complex metric to understand than accuracy. These extra metrics provide additional insight. The RMSE, R2 and PCC are reported separately for valence and arousal.

Additional experiments

While not a separate experiment by itself, feature importance is also calculated during all experiments listed in Table 3.1 using permutation importance. Permutation importance outputs importance values for all speech or music features used. These values describe the impact of the features on the predictions. The importance thus describes how much the model relies on each feature. This is separate from the model's actual performance. Permutation importance on a poorly performing model describes how the bad performing model relies on its features to achieve that performance, nothing else.

However, these feature importances can be used to study how two models differ, possibly explaining their performance difference. To be more specific to our use case, they give insight into how the models differ between experiment scenarios. Which features are important and unimportant for certain audio types compared to others? These kinds of insights can be obtained through the feature importance analysis.

To keep the scope limited without compromising too much, the feature importance analysis is limited to experiment A, D and F. These are the three experiment where the models are trained and tested on the same audio type. This analysis will be covered in the Discussion chapter rather than the Results chapter, as it is strongly speculative in nature.

An extra experiment is also included. Using the obtained feature importances, a feature importance ranking can be made by ordering all features based on their importance value. Some experiments are rerun using only limited amounts of the most important features, instead of the entire feature set. These experiments give us insight into how much of the feature set is essential for decent performance, and what levels of performance could be achieved with minimal feature sets in a MiSME recognition task. This could be useful in cases where computing power is limited and only a few features can be used.

So to summarize, the speech and music models of the MiSME system are tested using six nearly identical experiments where various audio types are used for training and testing. These allow us to establish to what degree mixed audio affects speech and music emotion recognition, and how well a MiSME specialized system can perform compared to generic speech and music emotion recognition models. We also test the difference between non-BSS and BSS-audio. Speech emotion recognition performance is measured using accuracy, while music emotion recognition is measured using the root mean squared error. Finally, feature importance is calculated on all main experiments using permutation importance. This allows us to compare various models, showing which features might be (un)important for certain audio types. This can be used to explain possible performance differences. The feature importances are also used to rerun the models with only limited sets of the most important features, showcasing how well they can perform with highly limited feature sets.

3.4 Speech emotion recognition pipeline

As explained earlier, the speech emotion recognition part of the MiSME system consists of a feature extractor and a speech emotion recognition model. This section describes both components.

3.4.1 Speech feature extractor

It is common practice to use a set of handpicked speech features for any SER task, often supported by either acknowledged studies or good reasoning. However, it is not known how well existing SER feature knowledge translates to mixed audio. This makes using a large feature set that contains many common SER features advantageous over handpicking a smaller feature set. A larger feature set more likely contains suitable (effective) MiSME features, and it also produces more valuable feature importance results, which we use to study how features perform differently on mixed audio than on speech-only audio.

A handful of audio feature extraction toolkits exist (see Table 2.3). We used *openSMILE* (Eyben et al., 2013). OpenSMILE is a feature extraction toolkit focused on speech. It can be run using various 'configurations', resulting in different feature sets. We used openSMILE's 'emobase2010' configuration, which contains a total of 1582 speech features. See Section 2.5.1 for a more detailed description of the features included in this configuration. All features are global (duration-independent), and are suitable for any kind of SER problem according to the manual.

3.4.2 Speech emotion recognition model

The speech emotion recognition task in our experiments is a 8-way classification task as dictated by RAVDESS's annotations. The output is always one of the eight possible emotion labels: neutral, calm, happy, sad, angry, fearful, disgust and surprise. For this kind of emotion recognition task a classifier is needed. To keep the scope manageable and the results generalizable the decision was made to implement four simple but different classification models. The four models are: a *Multilayer Perceptron Classifier* (MLPC), *Random Forest Classifier* (RFC), *Support Vector Classifier* (SVC) and a *Deep Neural Network* (D-NN).

The first three models were built using 'scikit-learn', a popular python-based machine learning library. The D-NN was built using 'Keras', a popular deep learning library also in Python. A background on each model can be found in Section 2.5.2.

All four models will be tested on the speech-only audio emotion recognition task to see which models are fit for use (Experiment A in Table 3.1). One or two models will be selected based on their performance, which will then be used for all of the other experiments listed in Table 3.1.

While already briefly mentioned, we want to stress that we treat the speech emotion recognition task in this research as *speaker-independent*. This means that the model is not allowed to see samples of an actor during training and testing. This would allow the model to obtain better performance because it is familiar with that actor's voice. Published models that reported performance on RAVDESS treated their experiments as *speaker-dependent* (Bhavan et al., 2019; Zeng et al., 2019) . The model-selection experiment will also include separately reported speaker-dependent performance on RAVDESS, making it possible to compare our four simple models to these publications.

3.5 Music emotion recognition pipeline

The music emotion recognition pipeline consists of the music feature extractor and the music emotion recognition model. This section describes both.

3.5.1 Music feature extractor

A large music feature set was preferred for the same reason as mentioned in Section 3.4.1, a large feature set increases the chances of finding suitable MiSME features and produces more valuable feature importance results. For this reason *Essentia* (Bogdanov et al., 2013) was favored over openSMILE and the other toolkits (see Table 2.3), using the precompiled 'essentia_streaming_extractor_music' executable. We left out all non-global (duration-dependent) features, resulting in a feature set consisting of 2651 features in total. Many common MER features are contained in this large feature set, see Table 2.5.

3.5.2 Music emotion recognition model

The music emotion recognition task in our experiments is a two-value regression task. The model must predict the level of valence and arousal within a range of 1 to 9, as dictated by the emotion annotations of DEAM and Soundtrack. The same four types of models are used for the music emotion recognition experiments, but now in regressor form. These are: a *Multilayer Perceptron Regressor* (MLPR), *Random Forest Regressor* (RFR), *State Vector Regressor* (SVR) and *Deep Neural Network* (D-NN). For a background on these models see Section 2.5.2, but keep in mind that the regressors output numerical values instead of categorical labels.

The music emotion recognition capabilities of all four models are tested on music-only audio (see experiment A in Table 3.1), similar to the speech models. One or two models will be selected based on their performance, these models will then be used for all of the other experiments listed in Table 3.1.

Unfortunately we can not compare the performance of our models to any published models. The issue is not that we use a custom dataset (DEAM + Soundtrack), but rather that neither of the two datasets has seen published results to which our models can be compared to. While there exists various published models on DEAM, see Aljanaki et al. (2017), they all use the continuous annotations instead of the static annotations used in this experiment. Continuous valence-arousal predictions is a different problem and thus those reported metrics are not comparable to our static prediction metrics.

There is one more thing which needs to be adressed. There are two values which need to be predicted per sample, the valence and the arousal. Not all models are able to output multiple dependent values (valence and arousal) from the same set of independent variables (feature vector). Only the MLPR and D-NN are able to do this. To compensate for this both the RFR and SVR will have a separate valence and arousal model. They are trained identically, using the same parameters and inputs. Only the ground truth values are different.

3.6 Blind-source separation component

The blind-source separation component is used to produce a blind-source separated copy of the mixed-audio dataset. From all available tools mentioned in Section 2.4 Deezer's *Spleeter* tool was used (Hennequin et al., 2019). The authors claim that it was the best performing BSS tool at the time of publication, as it performed better than Open-Unmix, the top performing BSS tool before Spleeter.

Spleeter is originally designed for music stem separation. The tool offers various pre-trained models for various stem combinations (piano, drums etc.). One of these pre-trained models is the 2-stem separator. It is trained for vocal-accompaniment separation. The accompaniment is all other musical sources except vocals.

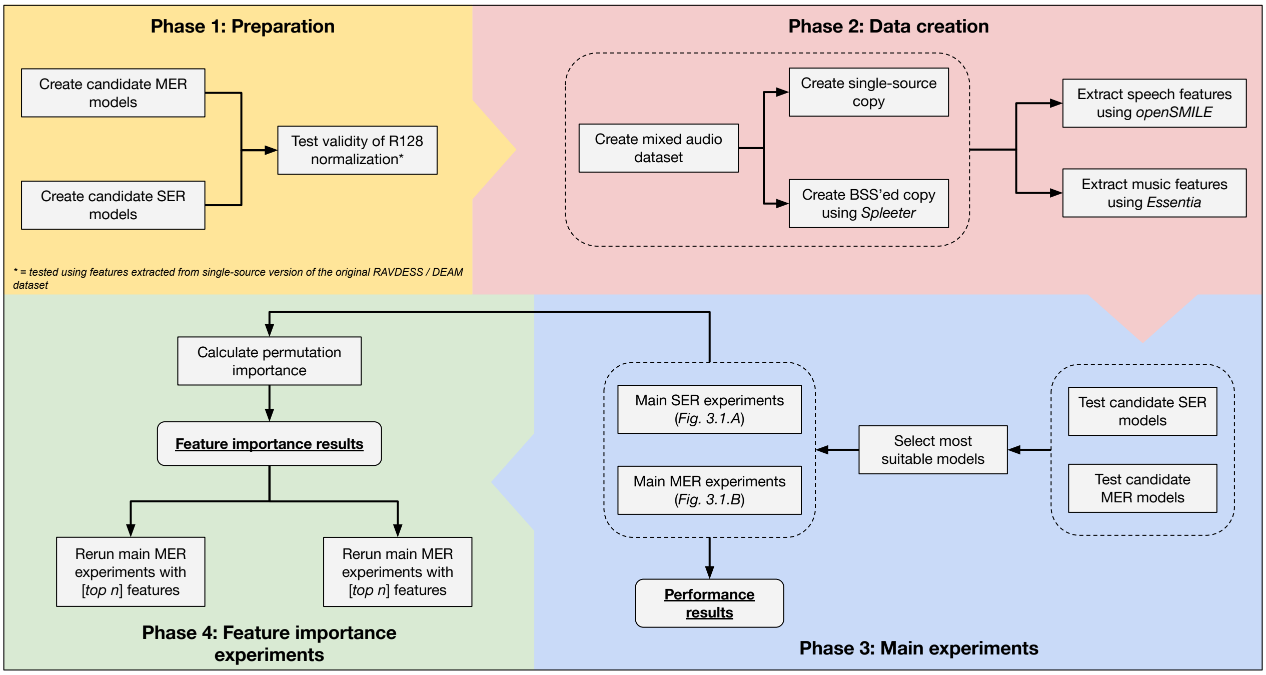

Phase 1: Preparation
Create candidate MER models
Create candidate SER models
Test validity of R128 normalization*
* = tested using features extracted from single-source version of the original RAVDESS / DEAM dataset
Phase 2: Data creation
Create mixed audio dataset
Create single-source copy
Create BSS'ed copy using *Spleeter*
Extract speech features using *openSMILE*
Extract music features using *Essentia*
Phase 3: Main experiments
Test candidate SER models
Test candidate MER models
Select most suitable models
Main SER experiments (*Fig. 3.1.A*)
Main MER experiments (*Fig. 3.1.B*)
Performance results
Phase 4: Feature importance experiments
Calculate permutation importance
Feature importance results
Rerun main MER experiments with [*top n*] features
Rerun main MER experiments with [*top n*] features

Figure 3.8: An overview of all experiments and other actions that are part of this research, all discussed in this chapter. It is divided into four phases, which read from top-left to top-right, bottom-right and finally bottom-left.

While speech is not exactly the same as singing, we expected that the model would be able to do speech-accompaniment separation because speech is strongly similar to singing. Also, because our music dataset is purely instrumental the music sample can be seen as accompaniment as they do not contain vocals.

The quality of the two-stem model was evaluated using various handpicked samples from our own dataset. For most samples the isolated audio appeared to be nearly identical to the original, based on perceptual evaluation. However, the isolated audio sounded like it was of lower quality, and in some samples instruments leaked into the isolated speech. These instruments always sounded similar to speech, for example a violin. We still deemed Spleeter to be fit for use regardless.

3.7 Summary

We have now covered all aspects of the research design and creation of the MiSME system. An overview of how all experiments and other elements are related is depicted in Figure 3.8. This diagram helps with understanding the relationship between and 'flow' of all experiments, of which the results will be discussed in the next chapter.

To summarize this chapter: The MiSME system consists of a separate SER and MER pipeline, with both a feature extraction component and an emotion recognition model. Using predefined feature extraction was favored over the 'end-to-end' learning approach due to uncertainties. Optionally, a blind-source separation component is included in some versions of the MiSME system.

RAVDESS was selected as our speech dataset, and DEAM and Soundtrack are combined to create our music dataset. All music samples with vocals were filtered out, as they will likely increase the complexity of the blind-source separation and speech emotion recognition tasks.

To properly create mixed samples all samples (RAVDESS, DEAM and Soundtrack) were normalized using the R128 normalization standard. A mixed sample is created by mixing the speech sample repeatedly in the music sample at 4.5 second intervals. This increases the speech-music overlap from 2.5% to 25% on average. The speech and music datasets are divided into six folds. Mixed samples are created from speech and music samples of the same fold, ensuring that an actor or music piece is only seen during training or testing. Each speech sample is mixed with five randomly chosen music samples, producing five mixed samples per speech sample. This results in a mixed audio dataset of 7200 unique samples.

Both the speech and music emotion recognition models are tested using six different experiments. These are all possible combinations of training and testing on one of the three available audio types: single-source, mixed or blind-source separated audio. Comparing the performance between these experiments should provide enough information to answer the main research question. In addition a feature analysis is performed on the three experiments where the same audio type is used for training and testing. This is done using permutation importance and should provide valuable insight into which features are fit for MiSME recognition and how the models differ between experiments.

Four types of models are created for both the speech and music emotion recognition task. These are a State Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP) and Deep Neural-Network (D-NN). Each model takes the form of a classifier in the speech emotion recognition task, and the form of a two-value regressor in the music emotion recognition task. The performance of these models are tested early on on the single-source emotion recognition task to test their suitability. One or two models will be selected to be used for all other experiments to avoid a bloated results section and an increasing scope.

Speech feature extraction is done using openSMILE with the ‘emobase2010’ configuration. It produces a total of 1582 speech features and contains a large number of common SER features. For music feature extraction Essentia was used, using their precompiled feature extractor. It produces a total of 2651 music features per sample and also contains a large amount of common MER features.

For blind-source separation Deezer’s Spleeter was used. At the time of the research it was the best performing BSS tool. The pretrained 2-stem model, which is originally trained for vocal-accompaniment separation, lends itself well to speech-music separation.

This concludes the Methodology chapter. The next chapter covers the results from all experiments, along with descriptive statistics of the datasets created in this research.

Chapter 4

Results

This chapter is split into three sections: Section 4.1 covers the creation of the mixed audio dataset, Section 4.2 covers the results of the speech emotion recognition experiments and Section 4.3 covers the results of music emotion recognition experiments. The analysis of the feature importances obtained through permutation importance are not included in this chapter, they can be found in Section 5.1.1 and 5.2.1.

4.1 Created datasets

RAVDESS, DEAM and Soundtrack were used to create the mixed audio datasets. Remember that the music emotion recognition task is on the entire mixed audio sample, while the speech emotion recognition task is done on one of the 4.5 second speech placements. In total three versions of the speech and music datasets were created: a single-source, mixed and blind-source separated version. While the single-source version is the original audio from RAVDESS/DEAM/Soundtrack, the other two were created using mixing and blind-source separation as explained in Section 3.2. This section describes the created dataset, to show that they were fit for use, as many actions were taken to create these datasets, such as dataset merging, random mixing between folds and R128 normalization.

Let us start with the combining of DEAM (Aljanaki et al., 2017) and Soundtrack (Eerola and Vuoskoski, 2011). While they use the same range for their valence and arousal annotations, the distribution of samples within that space is noticeably different between datasets, as can be seen in Figure 4.1. This can be attributed to how the samples were gathered. Soundtrack consists of a set of music samples selected by experts, where each sample should represent an extrema of

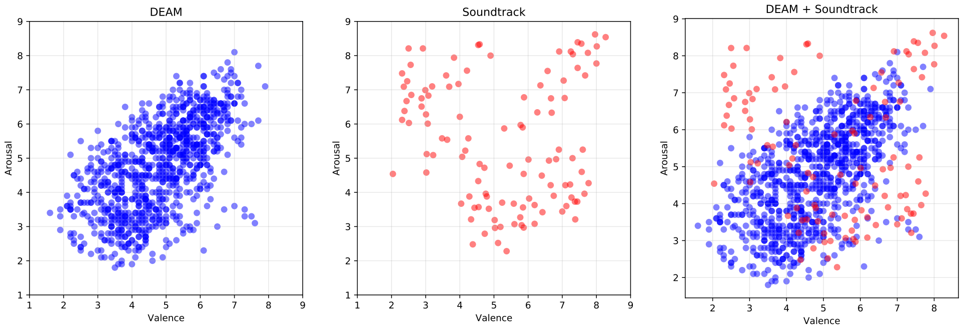


Figure 4.1: Distribution of the music samples contained in DEAM and Soundtrack

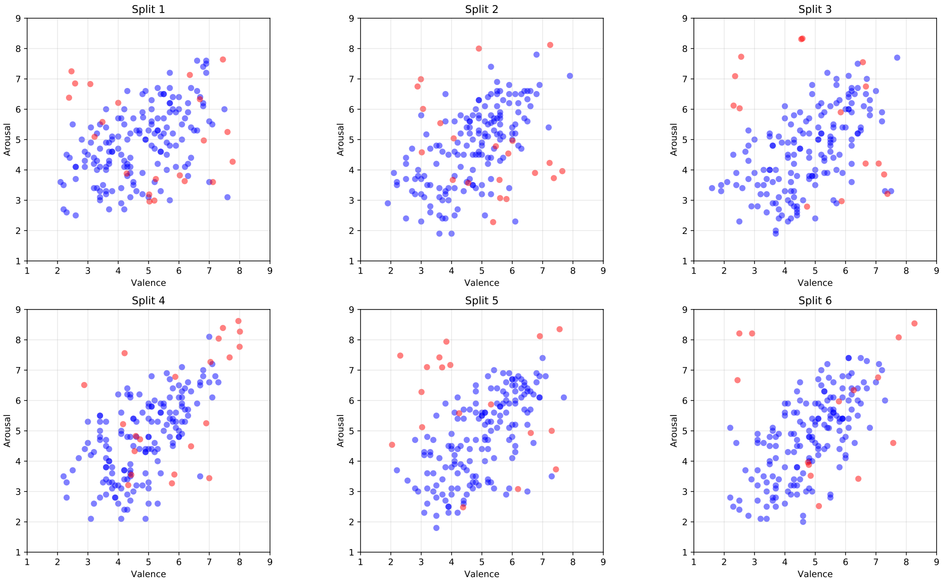


Figure 4.2: Distribution of music samples per split (blue = DEAM, red = Soundtrack)

certain categorical or dimensional emotions. For example high arousal or 'sadness'. This causes Soundtrack samples to occupy specific areas in the valence-arousal space, as those represent those extremes. DEAM on the other hand consist of music from various genres, not picked by experts on their possible emotion representation but based on their popularity (number of plays) on Free Music Archive. This means that the distribution of samples in DEAM better represent an 'average' set of music pieces than Soundtrack.

Figure 4.1 shows that DEAM does not cover the 'high-valence low-arousal' and 'low-valence high-arousal' corners well, if at all, while Soundtrack does. We speculate that this is because music pieces with extreme opposite valence and arousal might clash with average music taste, making them less popular, or that creating such pieces is hard to do. Whatever the cause might be, we can see that they are not common. While the samples in Soundtrack might be less generalizable due to the selection bias, it is only one-ninth of the size of DEAM. Combining both allows for better coverage of the total valence-arousal space, while not skewing the distribution too far from what the average music piece looks like due to dataset size difference.

Another thing that should be discussed is the creation of the folds. All datasets, including their audio variations, are split into six folds that preserve speaker-independence and music-independence during cross validation. During the initial split four actors were randomly assigned to each fold, this randomization was seeded. This means that the split can always be reproduced using the same seed. Because all actors have the same number of samples, emotions, utterances etc. a visualization was deemed unnecessary. All folds had at least one female and male actor, so we deemed the splits suitable as there was no severe gender-imbalance among the folds.

Regarding the splitting of music samples, they were assigned to each fold by randomly shuffling the order of samples and splitting them in six equal sized partitions. Because there are a total of 1057 music samples and six folds, fold 1 has 177 samples and the others 176 samples. This shuffling was seeded. Figure 4.2 depicts the valence-arousal space coverage of each set. We spotted no distressing difference between folds and deemed them fit for use.

Finally, the selection of mixing combinations of speech and music samples during the creation of the mixed dataset was also done randomly, but seeded. When selecting the five music samples, already selected music samples were taken out of the pool of candidate samples to prevent duplicate mixed samples from occurring. A histogram depicting the frequency distribution of music

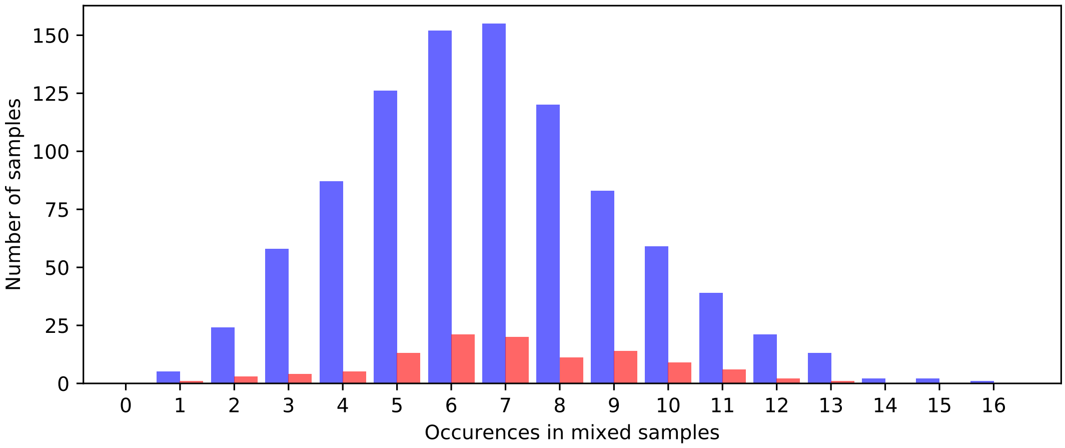


Figure 4.3: Frequency distribution of music samples appearing in the mixed and blind-source separated datasets (blue = DEAM, red = Soundtrack)

samples occurring in the mixed samples is depicted in Figure 4.3. On average a music sample occurs in 6.8 mixed samples. It is interesting to note that all music samples occurred at least once in the mixed dataset, so no samples are lost due to random selection.

Overall we see no (severe) flaws in the created datasets that would make them unfit for use. The combining of DEAM and Soundtrack allows for better valence-arousal space coverage, while keeping the distribution generalizable. The creation of the six folds showed no imbalance and during mixing all music samples occurred in at least one mixed sample, so none are lost. We therefore deemed the datasets fit for use.

4.2 Speech emotion recognition experiments

All of the speech emotion recognition experiments can be divided into four groups: the R128 normalization test, candidate model selection, the main research question experiments (see Table 3.1) and the most important features experiments. The results of each group of experiments is reported in separate subsections.

For all experiments feature values were standardized to have zero-mean and unit variance before training. This was done using the 'StandardScaler' function of Sklearn. Ground-truth labels are a single integer, ranging between 1 and 8 representing an emotion label in the following order: neutral, calm, happy, sad, angry, fearful, disgust and surprise. The models were trained in a speaker-independent setting using 6-fold cross validation during all experiments unless explicitly stated otherwise.

4.2.1 Validity of R128 normalization

At an early stage a small scale experiment was performed to see if R128 normalization was valid to use. As a refresher, R128 normalization would allow for mixed sample creation with equal loudness. Without it there would be strong variation in perceivable loudness of the music and speech sample, which we deemed problematic.

A state vector classifier (SVC) using the default Sklearn settings was trained and tested on both a normalized and a non-normalized version of RAVDESS, this is speech-only audio. The accuracy of the model dropped from 61% to 60.3% when applying R128 normalization on the 8-way classification task. While this is a decrease, we deem it not severe enough to make R128 normal-

Model	**Accuracy$_{\text{SD}}$**	**Accuracy$_{\text{SI}}$**
D-NN	77.4%	61%
MLPC	76.5%	60.3%
Bagged ensemble SVM (Bhavan et al., 2019)	75.6%	-
SVC	74.8%	57.4%
RFC	67.2%	55.1%
D-NN (Zeng et al., 2019)	65.5%	-
Dummy$_{strat}$	13.5%	13%
Dummy$_{mf}$	9.2%	13.3%

Table 4.1: The speaker-dependent and -independent performance of all our candidate models and other published models on RAVDESS

ization unsuitable for use within our research. The statements by Olsen et al. (2015) suggested a more severe loss of performance, as they state that perceptual loudness is a 'strong predictor' of arousal.

4.2.2 Candidate model selection

As mentioned in Section 3.4.2, four candidate SER models were created: a State Vector Classifier (SVC), Multilayer Perceptron Classifier (MLPC), Random Forest Classifier (RFC) and Deep Neural Network (D-NN). Their implementations are discussed below.

SVC The SVC used all default settings defined in Sklearn, so a radial basis function (RBF) kernel, a regularization parameter of 1 and a kernel coefficient of $1/(n\_features * X.var)$.

RFC The number of estimators (decision trees) of the RFC was set to 1000. We limited the maximum number of features to be considered per split to $\sqrt{N_{features}}$. This is considered superior to other options when training for classification, but since it is the most aggressively trimming option the number of estimators was increased from the default 100 to 1000 to compensate this aggressive trimming.

MLPC The MLPC consisted of one hidden layer with 512 perceptrons with a 'logistic' activation function. This is generally speaking the best activation function for classification tasks. Only one hidden layer was used because it would otherwise be too similar to the D-NN model. Early stopping was turned on for the MLPC, with a maximum number of 5000 iterations (epochs), a batch size of 256 and an adaptive learning rate.

D-NN The D-NN consists of three hidden layers of 512 neurons each, with a dropout-rate of 0.5. The hidden layer neurons were set to use 'ReLu' activation, the output layer to 'softmax'. Hidden layer neuron bias was initialized as 0.01. The number of epochs was set to 25 to avoid overfitting on the test set, using the categorical cross-entropy loss and the 'adam' optimizer. An attempt was made to implement dynamic early-stopping but it did not function consistently and was therefore dropped in favor of a static epoch parameter.

Each model was tested in both speaker-dependent and independent setting on the RAVDESS dataset to get an idea of their speech emotion recognition abilities. This is speech-only audio. The results are depicted in Table 4.1. In addition to the four candidate classifiers two dummy models were added. The performance of these dummy models are similar to chance-level and serve as a baseline performance which we can compare our candidate models to. $Dummy_{mf}$ always outputs the most frequently occurring class, while $Dummy_{strat}$ picks a class randomly based on the probability distribution over the entire training set.

Let us start with speaker-dependent performance, which is not the preferred setting, where we can compare the candidate models to published models. All of our candidate models outperform the model created by Bhavan et al. (2019) (65.5%), as seen in Table 4.1. The MLPC (76.5%) and D-NN (77.4%) also outperform the model by Zeng et al. (2019) (75.6%), while the SVC model comes close (74.8%). The RFC performs noticeably worse than the other three candidate models (67.2%). Nevertheless, this shows that all four models are capable of speech emotion recognition close to, or better than, 'state-of-the-art' models on RAVDESS.

When moving to the preferred speaker-independent settings we see a drop in accuracy for all models, ranging between 12% to 17%. This was expected as the models can not benefit anymore from samples of actors appearing in both the training and testing set. This shows that speaker-independent emotion recognition is a harder task than speaker-dependent emotion recognition on RAVDESS. However, all models still perform far above chance level. Accuracy scores of 55% to 61% on a 8-way classification task is still impressive, especially considering that the human recognition rate on RAVDESS is 62.5%(Livingstone and Russo, 2018). All candidate models thus seem suitable for speaker-independent speech emotion recognition on RAVDESS.

The decision was made to continue with the D-NN and RFC, using these two models for all other speech experiments. The D-NN was picked because it is the top performer on both the speaker-dependent and independent experiment. Using the top performer could be advantageous when MiSME recognition is not that different from normal speech emotion recognition. The RFC was chosen because it is the only candidate model that can compute feature importance through permutation importance, which is highly favored over regular feature importance inspection methods[1]. While it is the worst performer of all four, its architecture and learning-approach is drastically different from the D-NN, which could be advantageous.

4.2.3 Main experiments

Both the D-NN and RFC model were tested in six different experiments (see Table 3.1). As a refresher, the six experiments are all possible combinations of training and testing on either speech-only, mixed or blind-source separated audio. These cover a wide range of scenarios. The results from these experiments are depicted in Table 4.2, Table 4.3a and Table 4.3b. The first table shows the accuracy of both models on all six experiments, while the other two depict the precision and recall for the four most important experiments. Also, when we speak of 'blind-source separated audio' in these experiments, we mean the produced speech output, not the accompaniment.

Experiment A In experiment A the models were trained and tested on speech-only audio, the optimal speech emotion recognition scenario. The performance achieved serve as the baseline performance, showing what the models can achieve when there is no mixed-audio interference, which was an accuracy of 55.1% (RFR) and 61% (D-NN).

Experiment B and C In experiment B the models were tested on mixed audio instead. This discrepancy between training the models on speech-only audio but tasking them to classify mixed audio resulted in such a strong decrease in accuracy that both models perform only slightly better than the dummy model (13%). The accuracy of the RFC dropped from 55.1% to 14.8%, and the D-NN from 61% to 14%. The same happens in experiment C, where the models were tested on

[1] https://scikit-learn.org/stable/modules/permutation_importance.html

	Training	Testing	RFC	D-NN	Dummy-Strat
A	*Original*	*Original*	55.1%	61%	13%
B	*Original*	*Mixed*	14.8%	14%	13%
C	*Original*	*BSS-speech*	15.6%	17.1%	12.8%
D	*Mixed*	*Mixed*	30.1%	29%	12.8%
E	*Mixed*	*BSS-speech*	27.3%	21.4%	13.4
F	*BSS-speech*	*BSS-speech*	41.4%	43.1%	13.1%

Table 4.2: Accuracy scores of the RFC, D-NN and Dummy models on all six speech experiments

blind-source separated audio (15.6% and 17.1% respectively). The performance thus appears to drop to near chance-level when tasking a speech-only trained (non-specialized) model to classify audio with speech-music overlap.

Experiment D In experiment D the models were trained and tested on mixed audio, so without blind-source separation. Both models perform significantly better compared to experiment B and C, with the accuracy of both models more than doubling (30.1% and 29% respectively). However, there is still a large accuracy difference compared to Experiment A (55.1% and 61%).

Experiment E Experiment E is included to cover all possible train-test combinations. This only shows that a mixed-audio trained model performs worse on blind-source separated audio than non blind-source separated audio.

Experiment F In the final experiment, experiment F, the models were trained and tested on blind-source separated audio. We see another jump in performance (41.4% and 43.1% respectively). They now significantly outperform their counterparts of experiment D (30.1% and 29%), with the only difference being that blind-source separation was included. Not only is the best performance on any form of mixed audio (experiment B to F) achieved using blind-source separation, the jump in performance is very significant as it increases the accuracy by 13.3% and 14.1% respectively.

However, there are a few interesting things to note regarding precision and recall (See Tables 4.3a and 4.3b). Recognizing the 'neutral' emotion appears to be challenging for the RFC on any form of mixed audio, as the recall remains close to zero. While lower precision and recall make sense due to accuracy differences between audio types, the precision and recall for the 'Sad' emotion also stand out. The precision and recall for 'Sad' see a stronger decrease compared to the others. It thus appears that 'Neutral' and 'Sad' are harder to recognize for the RFC on mixed audio. We also see that 'Neutral', 'Happy' and 'Surprised' score a zero on both precision and recall on Experiment B. This means a speech-only trained RFC is completely unable to identify those emotions when classifying mixed audio.

Regarding the D-NN we see that it does not struggle as much with 'Neutral' and 'Sad' samples as the RFR, as the precision and recall decrease more in-line with the differences in accuracy on all experiments. Overall its performance is much more evenly distributed over all emotions, suggesting that it is better capable at identifying all emotions.

To conclude, the best performance on both models was achieved with blind-source separation. It resulted in significantly higher performance than without blind-source separation, bringing it closer to single-source classification levels of performance than chance-level. However, there remains a performance gap of around 14% to 18% compared to single-source classification.

	Emotion	S - S	S - M	M - M	B - B
RFC					
	Neutral	.376	.0	.25	.286
	Calm	.542	.153	.334	.444
	Happy	.493	.0	.257	.404
	Sad	.36	.131	.147	.249
	Angry	.644	.286	.373	.502
	Fearful	.656	.167	.233	.443
	Disgust	.52	.185	.301	.361
	Surprised	.66	.0	.289	.44
D-NN					
	Neutral	.575	.2	.201	.237
	Calm	.632	.139	.399	.521
	Happy	.656	.154	.247	.365
	Sad	.436	.128	.194	.296
	Angry	.646	.163	.455	.571
	Fearful	.614	.131	.25	.354
	Disgust	.652	.176	.29	.447
	Surprised	.686	.15	.243	.554

(a) Precision

	Emotion	S - S	S - M	M - M	B - B
RFC					
	Neutral	.333	.0	.002	.05
	Calm	.812	.644	.654	.705
	Happy	.385	.0	.21	.257
	Sad	.26	.358	.09	.18
	Angry	.698	.008	.656	.549
	Fearful	.516	.039	.174	.354
	Disgust	.552	.057	.201	.43
	Surprised	.74	.0	.269	.606
D-NN					
	Neutral	.521	.002	.069	.225
	Calm	.688	.892	.417	.579
	Happy	.438	.01	.239	.357
	Sad	.5	.049	.299	.277
	Angry	.771	.008	.453	.546
	Fearful	.604	.034	.149	.384
	Disgust	.625	.051	.325	.441
	Surprised	.693	.006	.258	.538

(b) Recall

Table 4.3: Precision and recall scores on the four most important train-test combinations. S = Speech-only, M = Mixed, B = Blind-source separated speech

4.2.4 Most important features experiments

Feature importance was calculated for all of the above discussed RFC experiments using Sklearn's permutation importance. The number of repeats was set to five. The importance of a feature is calculated by taking the mean over all six folds. There are in total 1582 features. The resulting feature importance values are covered in Section 5.1.1.

The obtained feature importances were used to create a feature ranking, ordering the features based on their importance from high to low. As explained in Section 3.3, an additional experiment was done to see how well the models can perform with only limited sets of the most important features. We limited this to the setting of experiment F, where the model is trained and tested on blind-source separated audio as this lead to the best performance on any form of mixed audio recognition.

The number of most important features to be used were set at 20, 50, 100, 250 and 500 respectively. The accuracy scores of both models can be found in Table 4.4.

The performance achieved using only the 20 most important features is impressive (38.8% and 37% respectively). The accuracy difference between using all features and the twenty most im-

	RFC	D-NN
All features	41.4%	43.1%
Top 20	38.8%	37%
Top 50	42.6%	38.7%
Top 100	43.1%	41.8%
Top 250	43.6%	43.7%
Top 500	43.3%	43.4%

Table 4.4: Accuracy scores for various sets of top ranking features on blind-source separated audio (Speech)

portant is only 2.5% to 6%. What is even more interesting, is that the RFC outperforms its 'all features' counterpart when using only the 50 most important features. The D-NN is able to do the same with the 250 most important features, at which point both models also achieve optimal performance (43.6% and 43.7% respectively). The performance for the D-NN in this case is only slightly higher than using all features, but the RFC seems to benefit more from the limited set of features as it increases the accuracy by more than 2%.

It thus appears that both models can do MiSME recognition quite well using only a small number of features and blind-source separation, but we need to keep in mind that the gain in performance could be caused by the feature set over-fitting on the dataset. However, the performance obtained using such a limited set of features suggest that the model can be effective with only a few features, which is valuable when computing power is limited.

4.3 Music emotion recognition experiments

The music emotion recognition experiments can be divided in the same four groups: the R128 normalization test, candidate model selection, main music experiments and the top features experiments. Each group of experiments is reported in separate subsections.

Similar to the speech experiments all feature values were standardized to have zero-mean and unit variance before training. The valence and arousal target values were not standardized and thus kept at their original range of 1 to 9. The models were trained using 6-fold cross validation unless explicitly stated otherwise.

4.3.1 Validity of R128 normalization

A similar small scale experiment was performed to see if R128 normalization negatively affected music emotion recognition, the possibility of which was suggested by Olsen et al. (2015). A state vector regressor (SVR) model using the default Sklearn settings was trained on both a normalized and a non-normalized version of the DEAM dataset using 10-fold cross validation.

This normalization caused a drop in root means squared error (RMSE) performance from .863 to .847 for valence and from .866 to .848 for arousal. This is an increase in performance, as the error decreased. Contrary to the statements by Olsen et al. (2015), the loss of perceived loudness difference results in better music emotion recognition performance for our dataset. This should suffice as proof that R128 normalization can be used.

4.3.2 Candidate model selection

Again, four candidate models were created and tested to see how suitable they are for music emotion recognition. These four candidate models are: a State Vector Regressor (SVR), Multilayer Perceptron Regressor (MLPR), Random Forest Regressor (RFR) and Deep Neural Network (D-NN).

SVR The SVR is identical to the speech SVC regarding parameters. See Section 4.2.2 for more information.

MLPR The MLPR created consists of one hidden layer with 1024 perceptrons, twice the amount of neurons compared to the MLPC. This was done to compensate for the increased number of features compared to speech. It uses the 'logistic' activation function. Again, we used only one hidden layer to prevent the MLPR from being too similar to the D-NN. Early stopping was turned on, limited to 1000 iteration using an adaptive learning rate and a batch size of 256.

	Valence			Arousal		
Model	**RMSE**	**R^2 (stdev.)**	**PCC**	**RMSE**	**R^2 (stdev.)**	**PCC**
RFR	.921	.565 (.106)	.709	.934	.595 (.078)	.742
SVR	.989	.421 (.149)	.646	1.007	.452 (.088)	.685
D-NN	.978	.448 (.144)	.667	1.024	.396 (.201)	.682
MLPR	1.137	-.016 (.274)	.58	1.163	.001 (.289)	.619
Dummy	1.296	-.688 (.131)	-.066	1.384	-.936 (.303)	-.09

Table 4.5: Performance of all candidate models on music-only audio - RMSE = Root Mean Squared Error, R^2 = Coefficient of Determination, PCC = Pearson correlation coefficient

RFR For the RFR the number of estimators was set to 100 estimators, rather than the 1000 of the RFC. This is because the maximum number of features to be considered at each split was set to 0.333. This is less aggressive compared to using $\sqrt{N_{features}}$, resulting in larger trees and longer training. This is generally considered better for regression. All other settings were kept at default.

D-NN The D-NN consists of three hidden layers with 512 neurons each and a dropout-rate of 0.5, similar to the speech D-NN. The hidden layer neurons were set to use the 'sigmoid' activation function and the output layer the 'linear' activation function. The bias of each neuron in all hidden layers was set to zero. The bias of the output neurons was set to 5, the median value of the range used for valence and arousal. This speeds up training because all neurons start with the median value of the total range. The training parameters were set to 25 epochs using the 'MSE' (mean squared error) loss function and the 'adam' optimizer. Early stopping could not be implemented successfully, it behaved inconsistently.

In contrast to the speech models we are unable to compare the candidate models performance to any existing models, as there are none on DEAM or Soundtrack (see Section 3.5.2). For this reason the performance of all candidate models were evaluated using our own music dataset. The results are depicted in Table 4.5. A dummy regressor is included to simulate 'chance-level' performance. It uses the mean-strategy, always reporting the mean valence or arousal value of the training set.

As explained earlier, we deem the root mean squared error (RMSE) the most important metric. We judge the quality of the models using this metrics, but the Coefficient of Determination (R^2) and Pearson correlation coefficient (PCC) score are reported as well. All reported PCC scores except for the dummy model had a 2-tailed p-value of less than 0.001.

The RFR achieves the best performance on both the valence and arousal dimension (*.921, .934*), closely followed by the SVR (*.989, 1.007*) and D-NN (*.978, 1.024*). These two models perform noticeably worse on arousal than valence compared to the RFR. The MLPR falls behind the other three regarding performance on both dimensions (*1.137, 1.163*), but all four models outperform the dummy model significantly (*1.296, 1.384*).

The decision was made to continue with both the RFR and D-NN and use them for all experiments listed in Table 3.1. The RFR was chosen because of its performance and permutation importance capabilities. The D-NN was included because it was also selected for speech, making the models used for the speech and music experiment identical. As also explained earlier, the RFR and D-NN differ strongly regarding architecture and learning approach, which is another benefit of using these two models.

It is hard to judge if the achieved RMSE performance is actually 'good' regarding our dataset and the valence-arousal space it covers, as there are no published models to compare them to. To give an idea how close the predictions are to the actual values, hex-binned heat maps of the D-NN and RFR are included in Appendix A.

	Experiment		Valence			Arousal		
	Training	**Testing**	**RMSE**	**R^2 (stdev.)**	**PCC**	**RMSE**	**R^2 (stdev.)**	**PCC**
A	*Original*	*Original*	.921	.565 (.106)	.709	.934	.595 (.078)	.742
B	*Original*	*Mixed*	1.048	.274 (.17)	.632	1.087	.26 (.103)	.7
C	*Original*	*BSS-music*	.975	.455 (.116)	.673	1.008	0.453 (.08)	.7
D	*Mixed*	*Mixed*	.942	.524 (.129)	.69	.931	.601 (.078)	.742
E	*Mixed*	*BSS-music*	.965	.476 (.142)	.674	1.043	.369 (.164)	.674
F	*BSS-music*	*BSS-music*	.909	.589 (.091)	.717	.928	.602 (.086)	.742

(a) Random Forest Regressor

	Experiment		Valence			Arousal		
	Training	**Testing**	**RMSE**	**R^2 (stdev.)**	**PCC**	**RMSE**	**R^2 (stdev.)**	**PCC**
A	*Original*	*Original*	.978	.448 (.144)	.667	1.024	.396 (.201)	.682
B	*Original*	*Mixed*	1.387	-1.229 (.304)	.291	1.538	-2.092 (1.139)	.315
C	*Original*	*BSS-music*	1.097	.126 (.19)	.565	1.132	.091 (.328)	.614
D	*Mixed*	*Mixed*	.947	.517 (.094)	.688	0.96	.546 (.056)	.723
E	*Mixed*	*BSS-music*	1.037	.304 (.106)	.613	1.131	.101 (.318)	.65
F	*BSS-music*	*BSS-music*	.943	.523 (.111)	.694	.979	.499 (.144)	.713

(b) Deep Neural Network

Table 4.6: Performance metrics of the RFR and D-NN on all six music experiments

The heat maps show that most predictions lie close to their actual values, with a large majority of the prediction being within 1.0 difference of the actual valence and arousal for both the D-NN and RFR. This is fairly decent performance considering the total range of 1 to 9. The difference in heat maps also shows that the RFR outperforms the D-NN, as the sample count is higher in the bins on and next to the diagonal line. However, they also show that the RFR struggles with high-arousal recognition.

Based on the RMSE performance and these heat maps we deem the RFR and D-NN models capable of music emotion recognition on our dataset, and thus suitable for further use.

4.3.3 Main experiments

The RFR and D-NN were tested on the six experiments listed in Table 3.1. These cover all possible combinations of training and testing on one of the three audio types: music-only, mixed and blind-source separated audio. The results of these experiments are reported in Table 4.6, showing their RMSE, R^2 and PCC on all six experiments. Heat maps of experiment A, D and F are also depicted in Appendix A, which provide a more detailed overview on the performance differences between experiments. Also, when we speak of 'blind-source separated audio' in this section, we mean the accompaniment output, not the speech.

Experiment A In experiment A (see Appendix A.2) the models were trained and tested on music-only audio, the optimal music emotion recognition scenario. The RMSE performance on this experiment serves as the baseline performance, showing what levels of performance can be achieved when there is no speech-interference. The RFR achieved a RMSE of *.921* for valence and *.934* for arousal, and the D-NN *.978* and *1.024* respectively.

Experiment B In Experiment B the models were tested on mixed audio instead. This discrepancy between training the models on music-only audio but tasking them to predict valence and arousal on mixed audio resulted in a strong increase in RMSE. The RFR increase to *1.048* and *1.087* respectively, and the D-NN (*1.387, 1.538*) now performs significantly worse than the dummy model (*1.296, 1.384*).

Experiment C However, testing on blind-source separated audio, as seen in experiment C, results in a noticeable improvement over experiment B. The performance of the D-NN now falls in-between the dummy model and its experiment A counterpart, as it achieves a RMSE of *1.097* and *1.132* respectively. The RFR (*.975, 1.008*) now performs more similar to itself in Experiment A, and even better than the D-NN in experiment A. The level of performance of the RFR is impressive considering it never saw speech-interference during training.

Experiment D In Experiment D (see Appendix A.3) the models were trained and tested on mixed audio. Something odd happened here. The RFR (*.942, .931*) achieves a lower RMSE on arousal than its experiment A counterpart (*.921, .934*), which is unexpected considering experiment A was seen as 'optimal' as there was no speech-interference. Even more interesting is that the D-NN significantly outperforms its experiment A counterpart on both valence and arousal, as it achieves a RMSE of *.947* and *.96* respectively compared to *.978* and *1.024* in experiment A. How this could have happened is broadly covered in the next chapter.

Experiment E Experiment E is not really relevant, but it shows that mixed-audio trained models perform worse on blind-source separated audio.

Experiment F In final experiment, Experiment F (see Appendix A.4), the models are trained and tested on blind-source separated audio. The RFR sees an improvement in performance, as it achieves a RMSE of *.909* and *.928* respectively. It now performs noticeably better on the valence dimension and slightly better on the arousal dimension compared experiment D (*.942, .931*). However, the RFR now also outperforms its experiment A counterpart (*.921, .934*) on both valence and arousal, the same thing that happened with the D-NN in experiment D. This is another unexpected occurrence, which is discussed in the next chapter.

The D-NN sees a slight improvement regarding valence compared to experiment D, but a significant decrease in performance on arousal as it achieves a RMSE of *.943* and *.979* respectively compared to *.947* and *.96*. This means that the D-NN performs worse on blind-source separated audio than raw mixed audio, and the other way around for the RFR.

The results of the six main music emotion recognition experiments are thus quite interesting. While the RFR appears to be consistently better than the D-NN model, the expected pattern of improvement between experiments, as seen in the speech experiments (see Section 4.2.3), is not present here. The D-NN can outperform its single-source performance using raw mixed audio, and the RFR can do the same with blind-source separated audio. It is unexpected that higher performance is achieved when speech-interference is present. These events will be discussed in the next chapter.

4.3.4 Most important features experiments

Again similar to the speech experiments, feature importance was calculated for all RFR experiments using Sklearn's permutation importance. The number of repeats was set to five. The importance of a feature is calculated by taking the mean over all six cross validation folds. There are in total 2651 music features. The resulting feature importance analysis is covered in Section 5.2.1.

Separate valence and arousal feature rankings were created by ordering all features from high to low based on their importance value (feature importance was separately calculated for valence

Number of features	Valence	Arousal
All features	.909	.928
Top 20	.878	.876
Top 50	.84	.866
Top 100	.848	.857
Top 250	.858	.876
Top 500	.881	.884

Table 4.7: RMSE performance of the Random Forest model using only certain amounts of top ranking features

and arousal). Experiment F was rerun using the RFR with the model only receiving only limited portions of the most important valence and arousal features. These were set to the 20, 50, 100, 250 and 500 most important features respectively. The valence and arousal RMSE for each run is reported in Table 4.7.

With only the 20 most important features the RFR model can already outperform its experiment F counterpart, which used all 2561 features, as it achieves a RMSE of *.878* and *.876* respectively compared to experiment F (*.909, .928*). The performance difference is quite significant, which is even more impressive considering it is done with less than 1% of the entire feature set. Optimal performance is reached when using the 50 most important valence features and 100 most important arousal features (*.84, .857*). The RMSE improvement is large compared to experiment F, which used all features. To put the performance difference into perspective, an heat map of the RFR using only these features is depicted in Appendix A.1.

The degree to which using these limited sets leads to an actual better model is hard to establish. As said before, the increase might be due to the feature set being perfect for the dataset, meaning that we are over-fitting the model. Then again, most of the importance assigned to features is likely due to their MiSME recognition ability, with possibly only the top few being particularly effective on our dataset and therefore being considered extra important.

Much better performance can already be achieved with using only the top 250 or 500, likely because better decision trees can be made when only effective features have to been considered. We think that using a limited set of the most important features will likely result in a better model, making it beneficial to do when creating the model. However, a slightly larger set than the optimal amount should be favored to ensure better generalization.

4.4 Summary

The results of each important set of experiments is summarized shortly below, to clearly state what the results show.

R128 normalization The preparatory experiments showed that R128 normalization did not negatively affect speech and music emotion recognition. It even led to better performance for the MER task.

Candidate model selection The Random Forest and D-NN models were picked from the four candidate models to be used in the main experiments, based on their single-source audio performance and differences in architecture.

Main speech experiments The models achieved an accuracy of 55.1% (RFR) and 61% (D-NN) respectively on speech-only audio. When the same model was tasked to classify mixed or blind-source separated audio, its performance dropped to near chance-level (around 14%). Training the models on mixed audio increased the performance to 29% and 30.1% respectively, which is significantly above chance-level. The inclusion of blind-source separation led to another significant increase in performance, increasing the accuracy to 41.4% and 43.1%. This means that optimal speech performance on mixed-audio was achieved using blind-source separation, but there remains a noticeable performance difference compared to speech-only audio, suggesting that lower performance should be expected in MiSME scenarios.

Main music experiments The root mean squared error (RMSE) metric showed that models trained on music-only audio performed noticeably worse on mixed and blind-source separated audio, similar to the speech experiments. However, something unexpected happened when the models were trained on mixed and blind-source separated audio. The D-NN achieved a lower RMSE when trained on mixed audio (*.947, .96*) compared to its music-only performance (*.978, 1.024*), meaning that was able to predict valence and arousal more accurately on mixed audio than music-only audio. The same occurred with the RFR when blind-source separation was included, as it achieved a RMSE of *.909* and *.928* respectively compared to music-only audio (*.921, .934*). This was unexpected, as the performance on music-only audio was considered the 'upper ceiling' as there is no speech-interference. This is something that will be discussed in detail in the next chapter. Still, this means that both models are able to do MiSME recognition equal to, or better than, music-only audio.

Using the most important features The final experiments, where the models used certain amounts of the most important features obtained from the permutation importance calculations, showed that both speech and music emotion recognition can be done with high levels of performance using only 1% of the entire feature set. Using only limited amounts of the most important features actually led to significant performance improvements on the MER task, as the optimal set led to a RMSE of *.84* and *.857* respectively compared to using all 2561 features (*.909, .928*).

Chapter 5

Discussion

In this chapter we will discuss the results presented in the previous chapter, along with a feature importance analysis. Based on the results and this discussion we will answer the main research question by the end of this chapter. To reiterate, the main research question is:

> *Can a system be produced which can recognize both the emotion of speech and music in mixed audio, where both are concurrently present, significantly above chance level?*

The speech and music experiments will be covered separately first, after which the main research question will be answered in Section 5.3. To finish the chapter Section 5.4 reflects on how this research went, discussing both positive and negative elements and possible future work.

5.1 Speech experiments

The speech emotion recognition task in our experiments is a 8-way classification task, as defined by the number of categorical emotions included in the RAVDESS dataset. Chance level accuracy is 13.5% considering the distribution of samples over the eight emotions, as there are only four 'neutral' samples per actor compared to eight samples for each other emotion. This is due to neutral only being expressed at one intensity.

We tested the speech emotion recognition capabilities of four different models using the speech-only audio version of our dataset. All four models outperformed the model by Bhavan et al. (2019). Our deep neural network (D-NN) and multilayer perceptron classifier (MLPC) advanced state-of-the-art on RAVDESS, outperforming the model by Zeng et al. (2019). It is noteworthy that we were able to achieve 'state-of-the-art' performance with such simple models. This can likely be attributed to the large feature set used compared to the published models. Still, this showed that the models used in this research were highly capable of speech emotion recognition on RAVDESS and therefore suitable models to use for this research.

Two of the four models were used for all speech experiments, the D-NN and Random Forest Classifier (RFC). A couple of interesting observations can be made from the results of the six speech emotion recognition experiments, which were all possible combinations of training and testing on different audio types.

The first thing we observed is that the speech-only trained models performed only slightly better than chance-level on both mixed and blind-source separated audio (Experiment B and C). This shows that non-specialized models can not handle MiSME recognition and specialized models, preferably with blind-source separation, must be developed to successfully do speech emotion recognition on mixed audio.

In addition, the fact that the speech-only trained models achieved ‘chance-level’ performance on blind-source separated audio proves that the speech audio produced by the blind-source separation component is not similar (enough) to the original speech to allow the model to function properly. This is interesting because we perceptually evaluated many blind-source separated samples to judge the quality of Spleeter, and we deemed the isolated speech perceptually very similar to their original samples. This shows that while blind-source separation produces perceptually similar audio for us humans, it is still affected too much by noise and inaccuracies to be processable by a speech-only trained model.

The third and final thing we observed is that the best accuracy for both models on any experiment with music-interference (Experiment B to F) was achieved when trained and tested on blind-source separated audio (Experiment F). The performance difference between using BSS and not using it was quite significant, as the accuracy jumped from 30.1% to 41.4% for the RFR and 29% to 43.1% for the D-NN. We believe that this is solid proof that blind-source separating mixed audio and training a model specifically for that type of audio is strongly superior to a raw mixed audio specialized model. We thus deem blind-source separation strongly beneficial for speech emotion recognition in a MiSME recognition task, and it should always be included if possible.

The results from the six main speech emotion recognition experiments thus showed us that specialized models need to be created for speech emotion recognition in MiSME scenarios, where the application of blind-source separation is strongly beneficial.

5.1.1 Feature importance analysis

Feature importance was computed for all 1582 speech features using permutation importance on all of the six experiments of the random forest model. These feature importances provide us with insight how the model achieved its performance on each experiment, showing which features had the biggest impact (importance) on the predictions made. Our goal is to compare the feature importances of various models, exploring possible causes of the performance difference between models while also showcasing which features are important for certain audio types. This should result in valuable knowledge for future MiSME research, as the feature importance analysis should show which features are effective on mixed audio.

We limited the scope of this analysis to experiment A, D and F. These are the experiments where the models were trained and tested on the same audio type. We deem these the three most important experiments as they showcase the performance of specialized models on the audio type which they were trained for. The results from experiment B and C are enough to show that speech-only trained models do not translate well to mixed or blind-source separated audio, a feature analysis is not necessary to further explain this.

A total of two sets of figures and one set of tables were created to visualize and analyze the computed feature importances. The figures can be found in the Appendix and are discussed in the following sections. These two sets of figures describe the models on a ‘feature-level’. The first set of figures, see Appendix B.1, shows the most important and unimportant features for each model on all three experiments. This was limited to the 20 most important features and 5 most unimportant features. These figures should show if there are any strongly dominant or problematic features for certain audio types.

The second set of figures, see Appendix B.2, shows the ten features with the strongest positive and negative chance in importance between any of the three experiments. To allow for fair comparison the feature importances were normalized to sum to one before calculating the absolute difference. These figures show if there are any individual features which massively change in importance between audio types, indicating that they become highly (un)suitable for that audio type.

Finally, Table 5.2 shows various feature importance statistics per feature type. These feature

	Top 20	Top 50	Top 100	Top 250
Speech only - Mixed	0	1	7	38
Mixed - BSS-speech	3	3	10	36
Speech only - BSS-speech	2	8	14	49
All three	0	0	0	8

Table 5.1: Shared features among the feature ranking of all audio types

types are defined by the openSMILE documentation. This table is deemed most descriptive. Most statistics are limited to the 250 most important features of the entire feature set. This was done because it provides a less skewed view of which features are important for the model, as the experiments showed that the top 250 contained most features beneficial to the model (see Section 4.2.4). This means that a large amount of the 1582 speech features are either suboptimal or negatively affect the model. Calculating statistics over the entire feature set would skew the metrics due to the dominance of suboptimal and negative features in the total features set.

The feature importance analysis is split into multiple subsections, each covering one aspect or observation. Let us start with the similarity of the most important features between audio types.

Feature similarity

Before we go into feature specific observations it is interesting to see how similar the most important features are between experiments. Table 5.1 depicts the number of features appearing in the N most important features for any two, or all three, experiments.

We see that there are almost no shared features between the speech-only model and mixed audio model (experiment A and D) in the top 20, 50 or 100. When looking at the 250 most important features of both models, we see that they share 38 features, so less than one-fifth. This suggests that vastly different features are used for mixed audio recognition. We suspect that many optimal speech emotion recognition features are not robust against mixed audio and are getting replaced by suboptimal ones which are robust, as the accuracy difference between the mixed and speech-only audio models ranges between 25% and 30%.

We also see the same pattern between the mixed audio model and blind-source separated audio model (experiment D and F), and the speech-only model and blind-source separated audio model (experiment A and F). While they share more features in the 20 to 100 range, a large majority is different, meaning that most features still get replaced by other features between audio types. We attribute this again to differences in robustness of features on certain audio types.

However, the fact that there are so few shared between the mixed and blind-source separated audio models likely means that blind-source separation allows the model to use more optimal features, as significantly better levels of performance were achieved when BSS was included in the model. But we also see that the speech-only model and blind-source separated model (experiment A and F) share only a few features. This likely means that blind-source separation is not able to reproduce the speech audio to such a degree that most optimal speech features remain usable, and thus get replaced by suboptimal ones. The performance difference between experiment A and F could be attributed to this, that the best usable features from the blind-source separated signal are just not as effective as the best features usable on speech-only audio.

If this would indeed be the case, which we strongly believe, it would mean that the quality of the blind-source separation strongly affects the level of performance which can be achieved. More accurate blind-source separation would mean that fewer optimal features need to be replaced by suboptimal ones due to lack of robustness against the inaccuracies and noise introduced during blind-source separation. While Spleeter is perceptually very impressive, this would mean that there is still much room for further improvement, as the replacement of features due to inaccura-

Feature type	M.I. - All	M.I. - Top 250	N in top 250	Contribution to total imp. of top 250
mfcc[630]	.021	.135	74 (11.7%)	24.8%
logMelFreqBand [336]	.042	.163	86 (25.6%)	34.6%
lspFreq [336]	.018	.137	31 (9.2%)	10.5%
F0 [82]	.094	.237	31 (37.8%)	18.2%
jitter [76]	.038	.181	13 (17.1%)	5.8%
loudness [42]	.009	.116	3 (7.1%)	0.9%
voicing [42]	.03	.25	5 (11.9%)	3.1%
shimmer [38]	.03	.125	7 (18.4%)	2.2%

(A) Speech-only audio

Feature type	M.I. - All	M.I. - Top 250	N in top 250	Contribution to total imp. of top 250
mfcc [630]	-.064	.027	102 (16.2%)	37.1%
logMelFreqBand [336]	-.057	.044	53 (15.8%)	34.1%
lspFreq [336]	-.065	.022	50 (14.9%)	16.3%
F0 [82]	-.083	.013	12 (14.6%)	2.3%
jitter [76]	-.066	.017	13 (17.1%)	3.2%
loudness [42]	-.072	.019	6 (14.3%)	1.7%
voicing [42]	-.052	.029	7 (16.7%)	3.0%
shimmer [38]	-.045	.02	7 (18.4%)	2.1%

(B) Mixed audio

Feature type	M.I. - All	M.I. - Top 250	N in top 250	Contribution to total imp. of top 250
mfcc [630]	.007	.075	92 (14.6%)	31.7%
logMelFreqBand [336]	.014	.01	55 (16.4%)	25.2%
lspFreq [336]	.007	.071	49 (14.6%)	16.1%
F0 [82]	.037	.164	21 (25.6%)	15.8%
jitter [76]	-.006	.097	6 (7.9%)	2.7%
loudness [42]	.019	.068	15 (35.7%)	4.7%
voicing [42]	.003	.073	7 (16.7%)	2.3 %
shimmer [38]	.002	.063	5 (13.2%)	1.4%

(C) Blind-source separated audio

Table 5.2: Importance statistics per feature type. The mean importance, relative contribution and number of features are reported - *Note: all importance values are reported in* e^{-2}

cies results in a performance gap of around 14% to 18%. However, perfect blind-source separation might be impossible to achieve.

Fundamental Frequency features

Something interesting occurs regarding fundamental frequency (f0) features. We deemed the speech-only audio model ‘optimal’ as it used audio which was not affected by music presence. This allowed it to achieve the best performance. Therefore the features used by this model can be seen as the ‘optimal’ performing set for our dataset.

The feature importance analysis shows that fundamental frequency features are important for the ‘optimal’ speech-only model. It relies significantly on these features, as seen in Figure B.1 and B.2 and Table 5.2. Multiple fundamental frequency features are among the most important features of the model, and fundamental frequency features have a noticeably high mean importance and contribution to the total importance of the model.

Fundamental frequency features appear to be of similar importance for the model using blind-source separated audio. However, they lose almost all of their importance on mixed audio (experiment D). No fundamental frequency features can be found among the most important features depicted in B.1, and the feature type scores significantly lower on all statistics in Table 5.2. This means that fundamental frequency features become almost useless on mixed audio, but they are able to regain most of their lost importance when blind-source separation is applied.

The fact that the lowest performance is achieved on the audio type which is unable to use fundamental frequency features is likely not coincidental. The fundamental frequency is the lowest dominant frequency found in the audio. Fundamental frequency features likely become useless because the lowest frequency is related to the music rather than speech in the features extracted from mixed audio. This makes the features represent the music, and thus useless for speech emotion recognition. This means that many of the dominant fundamental frequency features can not be used by the model anymore, resulting in lower performance.

Applying blind-source separation however results in both better performance and fundamental frequency features regaining importance. We believe that this is related. Blind-source separation is able to reproduce the speech with such accuracy that the fundamental frequency features are related to the speech again, making most features usable. The usable fundamental frequency features allow the model to better classify the emotion present, resulting in better performance.

MFCC, LMFB and LSPF features

The three largest feature types in the openSMILE feature set are the Mel Frequency Cepstrum Coefficients (MFCC), Log Mel Frequency Band (LMFB) and Line Spectral Pair Frequencies features (LSPF). They contain a total of 1302 features and make up more than four-fifths of the entire speech feature set. Our feature importance analysis shows that they contain relatively many important features regardless of audio type, judging by their importance contribution and number of features in the top 250.

Their importance thus appears to be consistent between audio types, but a few observations can be made. When moving from speech-only audio to mixed audio, MFCC and LSPF features gain importance almost equally to the importance loss of fundamental frequency features. This means that the model relies more on MFCC and LSPF features to compensate for the loss of fundamental frequency features, but this still leads to lower performance.

However, when applying blind-source separation it is not the LSPF features that lose importance due to fundamental frequency features becoming important again, but mainly the LMFB features and partially the MFCC features. This suggests that LSPF features, and to a lesser degree MFCC features, become more important when music is present in audio (mixed and BSS-audio). Useful MFCC and LSPF features thus appear to be more robust against music presence than LMFB features as they gain some importance at the cost of LMFB features. However, all three remain contribute significantly to the model regardless of audio type.

Other features

There are still four feature groups to discuss, these are: jitter, loudness, voicing and shimmer features. They are small compared to the other four feature groups and contribute significantly less to the model. However, they each contain multiple important features. While their contribution is not comparable to the larger feature groups, the fact that they describe other characteristics of the audio would suggest that they are not redundant and possibly allow to model to distinguish between samples that is not possible with solely the larger feature groups, but this is mere speculation.

5.1.2 Summary

So to conclude the analysis of the speech experiments: The results showed that speech-only models can not handle mixed or blind-source separated audio, meaning that specialized models are required for speech emotion recognition in a MiSME situation. However, creating a specialized model that uses blind-source separation results in a significant performance boost over no blind-source separation, as seen in the results. We deem this improvement significant enough to say that blind-source separation should always be included in the speech model of a MiSME system.

The feature importance analysis showed us that the most important features between all audio types are highly dissimilar. We speculate that many optimal features, as used by the speech-only model, are replaced by suboptimal features when classifying other audio types because most optimal features are not robust against music presence or blind-source separation noise and inaccuracies. We also saw that the features used by the blind-source separated audio model differed strongly from the mixed audio model, suggesting that blind-source separation allows for more optimal features to be usable. We speculate that the quality of blind-source separation plays an important role in the degree to which features must be replaced, causing it to dictate the 'upper ceiling' of what performance can be achieved compared to speech-only audio classification.

Regarding feature type specific observations we saw that fundamental frequency features are important for speech emotion recognition, but they can not be used by the model on mixed audio. Luckily blind-source separation makes them usable again, resulting in better performance. We also saw that MFCC, LMFB and LSPF features account for a large majority of the important features for each model, along with fundamental frequency features. The smaller feature types have a significantly lower contribution, but all types contain multiple dominant features and should not be excluded.

5.2 Music experiments

In Section 4.3.3 we showed the difference in root mean squared error (RMSE) performance for all experiments and made observations based on their differences. However, a proper test is required to know if the RMSE performance of any two experiments is significantly different. The Diebold-Mariano test (Diebold and Mariano, 2002) was used to test this, using the Python implementation by John Tsang[1]. All p-values obtained from these tests can be found in Appendix D.1. (In)significance will be reported using an italic font. Please keep in mind that the dataset consists of a total of 7200 samples, so moderate differences in RMSE performance can already be significant due to its large size.

The large difference in RMSE and heat maps observed in Section 4.3.3, and Appendix A, already strongly hinted at significant differences, which we pointed out. Many of the observations that we made are discussed again, but now with the Diebold-Mariano scores to formally confirm the suspicion of (in)significance. The RMSE results from the six main music experiments can be found in Table 5.3, which can be referenced while reading through this section.

Experiment A, B and C The Diebold-Mariano tests (Appendix D.1) show that the difference in RMSE performance is *significant* between experiments A, B and C on both valence and arousal for both models. This means that the music-only trained models perform significantly worse when tasked to do regression on mixed and blind-source separated audio (experiment B and C) than music-only audio (experiment A). This also means that they perform significantly better on blind-source separated audio than mixed audio. The Diebold-Mariano tests also show that the RFR

[1] https://github.com/johntwk/Diebold-Mariano-Test

	Experiment		**Valence**	**Arousal**
	Training	**Testing**	**RMSE**	**RMSE**
A	*Original*	*Original*	.921	.934
B	*Original*	*Mixed*	1.048	1.087
C	*Original*	*BSS-music*	.975	1.008
D	*Mixed*	*Mixed*	.942	.931
E	*Mixed*	*BSS-music*	.965	1.043
F	*BSS-music*	*BSS-music*	.909	.928

(a) Random Forest Regressor

	Experiment		**Valence**	**Arousal**
	Training	**Testing**	**RMSE**	**RMSE**
A	*Original*	*Original*	.978	1.024
B	*Original*	*Mixed*	1.387	1.538
C	*Original*	*BSS-music*	1.097	1.132
D	*Mixed*	*Mixed*	.947	0.96
E	*Mixed*	*BSS-music*	1.037	1.131
F	*BSS-music*	*BSS-music*	.943	.979

(b) Deep Neural Network

Table 5.3: A copy of Table 4.6 with only the RMSE performance

was *significantly* better than the D-NN on all three experiments. This was all as expected as the difference in RMSE varied wildly in these experiments.

The fact that the RMSE performance of the D-NN models is close to, or worse than, chance-level in experiment B and C makes it clear that music-only trained D-NN's are unsuitable for MiSME recognition. However, the same can not be said for the RFR model. Its performance when tasked to classify blind-source separated audio was deemed quite impressive by us, not only because it was close to its baseline performance, but it was also better than the D-NN on music-only audio (experiment A). While we already know that better performance can be achieved when creating specialized models (experiment D and F), the fact that a non-specialized model can achieve this level of performance is impressive. This is different than what we saw with the speech experiments, where all non-specialized models achieved chance-level accuracy.

Experiment D Something unexpected occurred when we trained and tested the models on mixed audio. The D-NN achieved a lower RMSE on both valence and arousal compared to its experiment A counterpart, which we expected to be the experiment where all models scored the lowest RMSE. The Diebold-Mariano test shows that the performance difference is *significant* for both valence and arousal. This means that the D-NN is able to do valence and arousal prediction significantly better on audio affected by speech than audio not affected by speech.

At first this seems counterintuitive. It is a logical assumption that the presence of speech in the audio should negatively affect music emotion recognition, similar to what we saw in the speech emotion recognition experiments. However, we speculate that the D-NN is able to achieve better performance due to the speech signal affecting certain music features which are ambiguous in nature. These are music features that are useful for valence or arousal prediction on some samples, contained in the training set, but negatively affect judgment for other samples, which likely can be found in the test set. These ambiguous features coincidentally overlap with the speech in one way or another, for example frequency-wise. The presence of the speech in the audio thus makes those ambiguous features much less reliant overall, forcing the D-NN to drop them in favor of more stable and universal features which are not affected by the speech[2]. This finally results in a better performing D-NN.

This occurrence was limited to the D-NN in experiment D, but we saw the same happen to the RFR in experiment F. The reason why this only occurred to the D-NN in experiment D could be due to the differences in model architecture. The D-NN is a single model that uses all features to learn a mapping during training, while the RFR consists of many decision trees which use random subsets of the entire feature set during training. It is not far fetched to assume that the RFR model was much less affected by these ambiguous features because many trees in the ensemble only

[2] We assume that useful features are affected by speech presence as well, but the performance gain from dropping ambiguous features is higher

see a few, or none, of these ambiguous features during training. Because judgment is based on a majority vote of all trees, where most are not or only partially affected by ambiguous features, the RFR is likely much less affected overall by the ambiguous features compared to the D-NN. However, this remains speculation.

The RFR model also achieved impressive levels of performance in experiment D. The valence and arousal RMSE is similar to experiment A, but the Diebold-Mariano test show that the difference is *only significant* for valence, on which the model performed worse. This means that the RFR can predict arousal equally as good on mixed audio as music-only audio. We therefore can not say that it performs better than its baseline model, but the difference is not much. We also want to point out that the RFR achieved a lower RMSE on both valence and arousal compared to the D-NN, albeit that only the difference on arousal was *significant*.

Overall it thus seems that the models can achieve impressive levels of valence and arousal prediction on mixed audio. Better than expected even, as the D-NN significantly outperforms its experiment A counterpart on both valence and arousal prediction.

Experiment F We already mentioned that the RFR was able to outperform its experiment A counterpart in experiment F, where the models were trained and tested on blind-source separated audio. The Diebold-Mariano tests show that the RFR performs *significantly* better on the valence dimension, and *insignificantly* better on arousal. The RFR model in experiment F also scored the lowest valence and arousal RMSE of all experiments, meaning that it is not only the best model on any form of mixed audio, but it is also the best performing music model of all experiments.

As already mentioned earlier, we believe that the presence of speech forces the models to drop ambiguous features. The presence of speech also negatively affects useful features, but the performance gain from dropping the ambiguous features compared to that was higher for the D-NN in experiment D. However, we believe that the inclusion of blind-source separation drastically decreases the degree to which useful features are affected negatively by speech, while keeping the elimination of ambiguous features due to speech presence. This combination allows the RFR model to achieve better performance on blind-source separated audio than on music-only audio. That is our theory of how this could have occurred.

Regarding the D-NN on experiment F, it still performs *significantly* better than its experiment A counterpart on both dimensions. However, it performs *insignificantly* better on valence and *significantly* worse on arousal compared to experiment D. The D-NN thus appears to perform better on mixed audio than blind-source separated audio, which is unexpected.

So to summarize the observations made from the results and Diebold-Mariano tests: Experiment A, B and C showed that music-only trained D-NN models are not suitable for music emotion recognition on mixed or blind-source separated audio. While the same could be said for the RFR models, its performance on blind-source separated audio is impressive.

We also observed that the D-NN was able to achieve better performance on both valence and arousal prediction when trained and tested on mixed audio, compared to music-only audio. We attributed this to the fact that the presence of speech forces the model to drop ambiguous features. The RFR was able to do the same on blind-source separated audio. We speculate that this is because the RFR is less susceptible to ambiguous features, and that blind-source separation limits the degree to which useful features are affected by speech presence. This finally resulted in the best performing model overall, even though experiment A was expected to be the best.

Keep in mind that this is speculation. Properly finding the cause of this would require dedicated research and is therefore outside of the scope of this research. While these results raise many question, it does appear that functioning music models can be developed that perform similar, or actually even better, in a MiSME scenario with around 25% speech-music overlap compared to normal music emotion recognition.

	Top 20	Top 50	Top 100	Top 250
Music only - Mixed	6	23	42	92
Mixed - BSS-music	8	22	45	95
Music only - BSS-music	7	24	52	103
All three	4	14	29	62

Table 5.4: Shared music features among the valence feature ranking of all audio types

5.2.1 Feature importance analysis

Feature importance was calculated using permutation importance on all experiments using the RFR model. The same two sets of figures and tables were produced as for the speech feature importance analysis (see Section 5.1.1). However, feature importance was computed separately on the valence prediction and arousal prediction tasks. This means that there are separate sets of tables for the valence and the arousal part of the model.

The set of figures found in Appendix C.1 depict the twenty most important and five most unimportant features per audio type. The other set, found in Appendix C.2, depicts the twenty most strongly differing features in importance between various audio types. Table 5.5 and Table 5.7 show the importance statistics per global feature type, similar to the speech importance analysis.

Again, we limit our observations to experiment A, D and F. We also limited the statistics to the 250 and 50 most important features, as the experiments have shown us that a majority of the beneficial feature lie within that range. That means that a majority of the 2561 music features contained in Essentia are suboptimal or negatively affect the model. Including these in most statistics would skew them, making them harder to analyze.

Valence prediction

Similar to speech we will discuss the similarity between the most important features for all audio types first, followed by feature type specific observations. We will start with valence, followed by arousal.

Similarity Table 5.4 shows the number of features appearing in the *N* most important features for any two, or all three, experiments. It is immediately noticeable that there is a higher degree of similarity between the most important features for valence prediction on all three audio types compared to speech emotion prediction. We see that around one-third is shared between any two audio types when looking at only the 20 or 50 most important features. This means that quite some dominant valence features transfer from one audio type to another, which was not the case for speech. However, the shared features among all three audio types is a bit lower, as it drops to only fourteen features. Still, this means that there are dominant valence features that perform consistently on all three audio type. These fourteen shared feature were spread over multiple feature types, no one type of feature appeared to be overrepresented. They can be found in Appendix E.1.

We also observed that the four most important features on all three audio types are the same features, but they differ in order per audio type (see Appendix C.1). They have a much higher importance value than other features shown. These four features, two rhythm and two spectral-based features, thus appear to be highly dominant and useful regardless of audio type.

It is interesting to see that at least some valence features exist that are dominant on all three audio types, which was not the case for speech.

Feature type	M.I. - All	N in Top 250	M.I. - Top 250	Contr. to Top 250 imp.	N in top 50
tonal [413]	0.0095	58 (23.2%)	0.0626	12.8%	8
erbbands [405]	0.0035	23 (9.2%)	0.0615	5%	4
melbands [405]	0.0057	32 (12.8%)	0.0927	10.4%	9
gfcc [351]	0.0007	17 (6.8%)	0.0254	1.5%	0
mfcc [351]	0.0023	16 (6.4%)	0.0447	2.5%	1
barkbands [288]	0.0046	15 (6%)	0.0978	5.2%	4
spectral [252]	0.0555	77 (30.8%)	0.1807	48.9%	20
rhythm [121]	0.0299	8 (3.2%)	0.4463	12.5%	2
silence [27]	0.0004	0	NaN	0%	0
hfc [9]	0.0169	1 (0.4%)	0.1330	0.5%	1
pitch [9]	-0.0056	1 (0.4%)	0.0316	0.1%	0
zerocrossingrate [9]	0.0158	1 (0.4%)	0.0658	0.2%	0
dissonance [9]	0.0028	0	NaN	0%	0
dynamic_complexity [1]	0.0122	0	NaN	0%	0
average_loudness [1]	0.1056	1 (0.4%)	0.1056	0.4%	1

(A) Music-only audio

Feature type	M.I. - All	N in Top 250	M.I. - Top 250	Contr. to Top 250 imp.	N in top 50
tonal [413]	0.0101	76 (30.4%)	0.0583	16%	14
erbbands [405]	0.0008	25 (10%)	0.0409	3.7%	1
melbands [405]	0.0032	33 (13.2%)	0.0573	6.8%	3
gfcc [351]	0.0010	11 (4.4%)	0.0287	1.1%	1
mfcc [351]	0.0002	9 (3.6%)	0.0397	1.3%	1
barkbands [288]	0.0006	15 (6.0%)	0.0537	2.9%	3
spectral [252]	0.0321	68 (27.2%)	0.1230	30.2%	23
rhythm [121]	0.0816	8 (3.2%)	1.2385	35.8%	2
silence [27]	0.0008	0	NaN	0%	0
hfc [9]	0.0024	1 (0.4%)	0.0262	0.1%	0
pitch [9]	0.0076	1 (0.4%	0.0299	0.1%	0
zerocrossingrate [9]	-0.0023	0	NaN	0%	0
dissonance [9]	0.0147	2 (0.8%)	0.0643	0.5%	1
dynamic_complexity [1]	0.0090	0	NaN	0%	0
average_loudness [1]	0.4057	1 (0.4%)	0.4057	1.5%	1

(B) Mixed audio

Feature type	M.I. - All	N in Top 250	M.I. - Top 250	Contr. to Top 250 imp.	N in top 50
tonal [413]	0.0127	73 (29.2%)	0.0688	16.1%	13
erbbands [405]	0.0046	33 (13.2%)	0.0614	6.5%	5
melbands [405]	0.0080	36 (14.4%)	0.0839	9.7%	8
gfcc [351]	0.0006	1 (0.4%)	0.0235	0.1%	0
mfcc [351]	0.0011	4 (1.6%)	0.0299	0.4%	0
barkbands [288]	0.0037	28 (11.2%)	0.0381	3.4%	2
spectral [252]	0.0495	64 (25.6%)	0.1912	39.3%	17
rhythm [121]	0.0600	5 (2.0%	1.4387	23.1%	3
silence [27]	0.0007	0	NaN	0%	0
hfc [9]	0.0131	2 (0.8%)	0.0922	0.6%	1
pitch [9]	0.0036	0	NaN	0%	0
zerocrossingrate [9]	-0.0393	1 (0.4%)	0.0302	0.1%	0
dissonance [9]	0.0141	2 (0.8%)	0.0893	0.6%	1
dynamic_complexity [1]	-0.0007	0	NaN	0%	0
average_loudness [1]	0.0427	1 (0.4%)	0.0427	0.1%	0

(C) Blind-source separated audio

Table 5.5: Importance statistics for **valence** per feature type. The mean importance, relative contribution and number of features are reported. - *Note: all importance values are reported in* e^{-2}

Rhythm The feature type importance statistics for valence (see Table 5.5) indicate that rhythm features are highly important for the achieved levels of valence prediction on all three audio types. They have the highest mean importance on all three audio types, and it is many times higher than the second most important feature group, especially for mixed and blind-source separated audio. This indicates that, on average, the useful rhythm features have a very strong impact on the achieved levels of valence prediction, see Table 5.3.

Rhythm features almost triple in contribution to the total importance when moving from music-only audio to mixed audio. We already know that the RFR model achieved a slightly worse RMSE on valence prediction on mixed audio (from *.921* to *.942*), which the Diebold-Mariano test proved to be significant. The fact that the performance decreased and the importance of rhythm features almost tripled, means that the model became much more reliant on rhythm features when it was tasked to predict valence on mixed audio. This suggests that there are highly effective rhythm features that can handle mixed audio well compared to the other available features. This could either be due to rhythm features becoming more effective when speech is present, or other types of features becoming less effective. The latter makes more sense.

When moving from mixed audio to blind-source separated audio the importance contribution of rhythm features decreases by a third, but it is still twice as high compared to music-only audio. This higher contribution compared to music-only audio is also achieved with fewer features. We speculate that blind-source separation makes other types of features more reliant again, replacing and outperforming certain rhythm features. However, there remain a couple of extremely dominant rhythm features which the model can use for valence prediction, resulting in the high mean importance.

Spectral Spectral features also appear to be important for valence prediction regardless of audio type. Not only are they consistently the feature group with the second-highest mean importance, they contribute the most to the models of all feature types. Spectral features account for almost half of the total importance for music-only audio, almost a third for mixed audio and two-fifth for blind-source separated audio. They achieve such a high contribution, often higher than the rhythm features, because there are many spectral features within the top 250, with a high average importance.

However, spectral features seem to suffer on mixed audio. Compared to music-only audio, their contribution on mixed audio is noticeably lower. We speculate that speech presence negatively affects spectral features, making them less reliable for valence prediction. They regain most of the lost contribution on blind-source separated audio. This is likely because blind-source separation lessen this effect, making them more reliable and thus more important again.

Overall spectral features appear to be of high importance for valence prediction regardless of audio type, similar to the rhythm features.

Tonal, erbband and melband features Tonal, erbband and melband features are the three largest feature groups in Essentia. They account for almost half of the entire feature set. Tonal features appear to be relatively important for all audio types considering their importance contribution. Their mean importance is low compared to rhythm and spectral features, but this is balanced out by the fact that many tonal features are among the 250 most important features. Overall, tonal features appear to be useful for the model and a valuable inclusion, but they consist mostly of many moderately important features, unlike the rhythm and spectral feature groups. The contribution of tonal features is also higher on mixed and blind-source separated audio than music-only audio, suggesting that they are robust against speech presence and thus suitable features for MiSME recognition.

Erbband and melband features appear to be moderately useful for all three audio types. The

	Top 20	Top 50	Top 100	Top 250
Music only - Mixed	6	19	34	85
Mixed - BSS	9	18	37	91
Music only - BSS	6	18	32	86
All three	5	10	20	47

Table 5.6: Shared music features among the arousal feature ranking of all audio types

statistics however show that they are less useful on mixed audio compared to music-only and blind-source separated audio. This suggests that they are less robust against speech being present in the signal, but blind-source separation reduces that effect. They also have a sizable number of features among the top 50 most important features, suggesting that they contain dominant features for valence prediction and should not be left out.

GFCC, MFCC, silence, pitch, zerocrossingrate and dynamic complexity These feature types seem to be less useful for valence prediction. Gammatone frequency cepstrum coefficient (GFCC) and Mel frequency cepstrum coefficient (MFCC) features consist of 405 features each, but are of minor importance for mixed audio, and almost no importance for blind-source separated audio. Silence, pitch, zerocrossingrate and dynamic complexity features also seem to be unimportant for either mixed or blind-source separated audio.

Arousal prediction

We will go over the arousal feature importance in the same format as before.

Similarity Table 5.6 shows the number of features appearing in the N most important arousal features for any two, or all three experiments. We again see that the degree of shared features between audio types is much higher than what we saw during the speech feature importance analysis. The degree of similarity is even a bit higher than what we saw for valence prediction when comparing any two audio types, but it is a bit lower when comparing all three audio types. Still, this means that there are a handful of dominant arousal prediction features that transfer between audio types, but a large majority of dominant features gets replaced by other features.

The ten dominant features shared across all audio types are this time not as uniformly distributed over all feature types. The ten shared features can be found in Appendix E.2. Seven of the ten features are spectral features, which means that there are seven dominant spectral features which can be applied universally regardless of audio type.

Rhythm Rhythm appears to be a highly important feature type for arousal as well. Its importance contribution is relatively constant between audio types, and the mean importance lies closer to the other feature groups compared to valence prediction, but it is still the highest on all three audio types. Based on the fact that we do not see massive change in statistics between audio types, and that there was no significant difference in arousal performance for the RFR (see Table 5.3), we suspect that rhythm features are not strongly negatively affected by the presence of speech. However, the dominant rhythm features for each audio type are almost completely different. This suggests that the set of rhythm features contains enough features to adapt to each audio type without much loss in performance.

Spectral Spectral features are also of high importance for arousal prediction on all three audio types, similar to valence prediction. This means that rhythm and spectral features are the most

Feature type	M.I. - All	N in Top 250	M.I. - Top 250	Contr. to Top 250 imp.	N in top 50
tonal [413]	0.0080	46 (18.4%)	0.0675	11.7%	9
erbbands [405]	0.0044	27 (10.8%)	0.0534	5.4%	3
melbands [405]	0.0054	36 (14.4%)	0.0576	7.8%	5
gfcc [351]	0.0036	27 (10.8%)	0.0363	3.7%	0
mfcc [351]	0.0023	12 (4.8%)	0.0351	1.6%	0
barkbands [288]	0.0049	21 (8.4%)	0.0737	5.8%	5
spectral [252]	0.0375	58 (23.2%)	0.1567	34.2%	18
rhythm [121]	0.0587	14 (5.6%)	0.4937	26%	6
silence [27]	-0.0007	0	NaN	0%	0
hfc [9]	0.0040	1 (0.4%)	0.0391	0.1%	0
pitch [9]	0.0047	2 (0.8%)	0.0338	0.3%	0
zerocrossingrate [9]	0.0831	4 (1.6%)	0.1862	2.8%	3
dissonance [9]	0.0080	1 (0.4%)	0.0450	0.2%	0
dynamic_complexity [1]	-0.0185	0	NaN	0%	0
average_loudness [1]	0.1461	1	0.1461	0.5%	1

(A) Music-only audio

Feature type	M.I. - All	N in Top 250	M.I. - Top 250	Contr. to Top 250 imp.	N in top 50
tonal [413]	0.0084	58 (23.2%)	0.0603	13%	12
erbbands [405]	0.0081	37 (14.8%)	0.0947	13%	10
melbands [405]	0.0024	29 (11.6%)	0.0377	4.1%	3
gfcc [351]	0.0016	15 (6.0%)	0.0248	1.4%	0
mfcc [351]	0.0011	12 (4.8%)	0.0296	1.3%	0
barkbands [288]	0.0014	13 (5.2%)	0.0428	2.1%	2
spectral [252]	0.0380	66 (26.4%)	0.1449	35.6%	18
rhythm [121]	0.0625	12 (4.8%)	0.6206	27.7%	3
silence [27]	-0.0023	0	NaN	0%	0
hfc [9]	0.0075	1 (0.4%)	0.0723	0.3%	0
pitch [9]	0.0035	2 (0.8%)	0.0235	0.2%	0
zerocrossingrate [9]	-0.0006	0	NaN	0%	0
dissonance [9]	0.0241	3 (1.2%)	0.0621	0.7%	1
dynamic_complexity [1]	0.0550	1 (0.4%)	0.0550	0.2%	0
average_loudness [1]	0.1376	1 (0.4%)	0.1376	0.5%	1

(B) Mixed audio

Feature type	M.I. - All	N in Top 250	M.I. - Top 250	Contr. to Top 250 imp.	N in top 50
tonal [413]	0.0048	54 (21.6%)	0.0411	8.9%	6
erbbands [405]	0.0031	37 (14.8%)	0.0503	7.5%	6
melbands [405]	0.0006	31 (12.4%)	0.0288	3.6%	2
gfcc [351]	0.0008	8 (3.2%)	0.0270	0.9%	0
mfcc [351]	0.0010	6 (2.4%)	0.0310	0.7%	1
barkbands [288]	0.0009	20 (8%)	0.0392	3.1%	1
spectral [252]	0.0445	69 (27.6%)	0.1653	45.7%	23
rhythm [121]	0.0504	12 (4.8%)	0.4999	24%	6
silence [27]	0.0024	2 (0.8%)	0.0280	0.2%	0
hfc [9]	0.0863	3 (1.2%)	0.2613	3.1%	3
pitch [9]	0.0082	2 (0.8%)	0.0183	0.1%	0
zerocrossingrate [9]	0.0378	3 (1.2%)	0.1101	1.3%	1
dissonance [9]	0.0180	3 (1.2%)	0.0573	0.7%	1
dynamic_complexity [1]	-0.0233	0	NaN	0%	0
average_loudness [1]	0.0148	0	NaN	0%	0

(C) Blind-source separated audio

Table 5.7: Importance statistics for **arousal** per feature type. The mean importance, relative contribution and number of features are reported. - *Note: all importance values are reported in* e^{-2}

important feature groups for both valence and arousal prediction regardless of audio type.

However, spectral features seem to increase in importance on blind-source separated audio compared to the other two. The contribution of spectral features to the total importance increases significantly in this case, with almost half of the fifty most important features being spectral features, while the RMSE decreased insignificantly from *.931* to *.928*. This seems to suggest that the model becomes more reliant on spectral features for blind-source separated audio. Still, spectral features appear to be highly effective for arousal prediction regardless of audio type.

Tonal, erbband, melband and barkband-features Tonal and erbband features are a step below rhythm and spectral features regarding importance, but their contribution is still quite significant. While their contribution to the total importance is only a fraction of that of rhythm and spectral, several dominant tonal and erbband features can be found among the fifty most important features for all audio types.

Melband and barkband features follow the same pattern but contribute less overall compared to tonal and erbband features. They become slightly less important on mixed and blind-source separated audio, meaning that they synergize less with speech presence than other feature types. However, a couple of melband and barkband features are included among the fifty most important features for all audio types. This means that there are also dominant features for arousal prediction among the melband and barkband feature-set, and are therefore beneficial to include.

Other features There are some interesting things to note regarding these feature groups. It appears to GFCC features become less important for mixed and blind-source separated audio, suggesting that speech-presence negatively affects its arousal prediction capabilities. Silence features appear to be not important at all. However, we think that silence features should not be excluded if MiSME recognition is done on audio with possible segments of silence, such as television content.

We also see that high-frequency coefficient (HFC) features contain three dominant arousal prediction features for the blind-source separated audio model, which achieved the best RMSE performance. It is interesting to see that these features are only important for blind-source separated audio model. It might be the case that HFC features are robust against noise and inaccuracies introduced by blind-source separation, or that they can signal the presence of noise to the model so that it can treat other features differently. Still, this shows that HFC features are important for arousal prediction on blind-source separated audio.

5.2.2 Summary

The results of the music experiments were not in-line with what we expected. We assumed that the models would achieve the best performance on music-only audio, because there is no speech-interference in that scenario. However, the D-NN was able to outperform its music-only performance when trained and tested on mixed audio. The same happened with the Random Forest model when trained and tested on blind-source separated audio. This means that both models are able to more accurately predict valence and arousal on speech-affected audio than on music-only audio.

We speculate that this is caused by the presence of speech forcing the model to drop more ambiguous features during training in favor of more stable ones, resulting in a better generalizing model. The D-NN exhibits this benefit earlier than the RFR due the differences in architecture, which made the D-NN more prone to ambiguous features in the baseline experiments. The presence of speech also negatively affects useful features, but blind-source separation appears to decrease this effect while keeping the positive effect of isolating ambiguous features. This finally allows the RFR to achieve the best valence and arousal performance of all experiments when trained and tested on blind-source separated audio.

We also saw that the Random Forest model performed better than the D-NN on all experiments, suggesting that it is the superior music model for valence-arousal prediction in a MiSME scenario. However, we can not say for certain that blind-source separation is a beneficial for valence and arousal prediction, unlike what we saw for speech. Blind-source separation allowed the Random Forest model to perform significantly better on valence, but it caused the D-NN to perform significantly worse on arousal. While blind-source separation does not consistently improve results, it thus appears to be beneficial for a Random Forest model. The Random Forest model on blind-source separated audio was also the best performing model overall, suggesting that blind-source separation might be necessary to achieve optimal performance.

The fact that the models were able to achieve such levels of accurate valence and arousal prediction on speech-affected audio means that music emotion recognition is definitely achievable in a MiSME scenario.

Regarding the feature importance analysis, many similarities were found between valence and arousal prediction. First off, we saw that less than one-fifth of the 250 most important features are shared between all audio types for both valence and arousal prediction. This suggests that the dominant features mostly differ for each audio type, even though similar levels of performance were achieved by the RFR. However, there did exist a handful of features which were dominant on all three audio types, unlike what we saw for speech.

Rhythm and spectral features were by far the most important features types for both prediction tasks, as they accounted for a majority of the dominant features. This means that rhythm and spectral features should always be included in the feature set. Tonal and erbband features were also important for both valence and arousal prediction, but their importance was a step below that of rhythm and spectral. Melband features were found to be of similar importance for valence prediction, but not arousal prediction. Finally, among the many other feature groups we saw that most had a relatively low importance, but they often contained at least one dominant features for either valence or arousal prediction, meaning that they are not useless.

5.3 Answering the research question

We are now finally able to answer the main research question. To reiterate, it goes as follows:

> *Can a system be produced which can recognize both the emotion of speech and music in mixed audio, where both are concurrently present, significantly above chance level?*

The speech emotion recognition experiments showed that performance far above chance-level could be achieved on mixed audio, with the best performing model achieving an accuracy of 43.1%, where chance-level was 13%. However, there remained a performance gap between the best mixed audio model (43.1%) and that model on speech-only audio (61%).

Similarly, the music emotion recognition experiments showed that music emotion recognition performance could be achieved far above chance-level on mixed audio. Not only were both models able to significantly outperform chance-level performance on both valence and arousal prediction on mixed audio, but they also achieved a lower valence and arousal root mean squared error (RMSE) on either mixed or blind-source separated audio than on music-only audio. This proves that a similar level of music emotion prediction can be achieved on mixed audio[3] to music-only audio.

Based on these results we can say that we have successfully created a MiSME system that achieved above chance-level performance on both the speech and music emotion recognition task.

[3] with around 25% speech-music overlap

However, audio-specific training was required to achieve these levels of performance. Hence, the research question can be answered positively.

5.4 Reflection and future work

This research has produced various useful insights into MiSME recognition. It proves that both speech and music emotion recognition models can be created that perform well on mixed audio. It also showed that blind-source separation is strongly beneficial for speech emotion recognition, and to a lesser extent as well for music emotion recognition, along with which common features are effective on mixed audio. Before we can conclude this research there are a couple of things that should be reflected upon.

This research is the first to study the problem of recognizing both the emotion of speech and music from a single mixed audio signal to the best of our knowledge, coined by us as MiSME recognition. To be more specific, we focused on how mixed audio affects both recognition tasks without context, assuming contextual-independence between the speech and music. This allows our results to be applicable to many MiSME use cases, as the results are obtained using experiments where the models could not benefit from a contextual relationship between speech and music. However, we had to come up with many solutions for first time problems in this research. While we observe that this research produces many valuable insights, it may be the case that some of the problems and solutions used were suboptimal or have some implications regarding generalization.

Context-less approach Let us start with the aspect of this research which has the biggest impact, the decision to study MiSME recognition in an environment where there is no relationship present between the speech and music. In many, possible even every MiSME use case there is a relationship between the music and speech. Television content would be a strong example of this, as stated before. This relationship between the two often contains valuable (contextual) information. For a MiSME system to function effectively in a real use case it should be able to understand and use this relationship. However, as stated before, how the relationship between speech and music takes form is highly specific for each use case. We mentioned the example of differences between movie genres.

MiSME recognition is a fairly unexplored problem. It was unknown how both speech and music emotion recognition behave on mixed audio compared to single-source audio from a 'low-level' perspective. How are dominant speech and music features affected? Are they still reliable on mixed audio? If not, can the model use other features to achieve similar performance to single-source audio? What kind of performance loss can we expect due to mixed audio? These were all interesting questions which are not specific to a use case, and allowed us to gain a better understanding of MiSME recognition in general. Because of this lack of general knowledge the decision was made to exclude a 'context' within our experiments, which allowed us to answer these questions.

However, we must stress that this strongly affects generalization. Our research shows how speech and music emotion recognition models and features are affected by mixed audio *in a general sense*. We think that it is very important to include the contextual relationship between speech and music in a real-life application. We speculate that it would not only improve the effectiveness of the models, but also produce more descriptive emotion information. We hope that the results from our research aids others in building MiSME systems for real life application, allowing them to build a system which performs 'decently' without using the context present. The next step would be to then develop the system further, 'optimizing' it on their context, allowing the model to learn the relationship between the speech and music for their specific use case.

Dataset implications The usage of RAVDESS, DEAM and Soundtrack also has several implications regarding the results obtained. The language spoken in RAVDESS is English. This means that our models are optimized for English speech. While we think that similar levels of performance should be achievable on different languages, this can not be guaranteed. The same goes for the fact that RAVDESS only contains two different utterances[4], both quite short and similar. It is unknown how well our results and models generalize to speech of different length, or with words not seen in the dataset.

There are similar implications regarding the music emotion recognition results. While it is beneficial that DEAM consists of music of various genres and Soundtrack of movie soundtracks, it is still unsure how well the results generalize to different use cases. DEAM's music stems from the Free Music Archive, where mostly unknown artists publish their work for free. This music might be very different than actual popular music from the same genres. The same goes for the movie soundtracks used. It is hard to establish how well the results generalize to various genres of television content. These things should be taken into consideration when generalizing our results to other use cases.

Speech-music overlap We had to create a mixed speech-music dataset ourselves as we were unable to find a suitable one that excluded a contextual relationship between speech and music. Mixed audio samples were created by mixing the same speech sample into the music sample with a 4.5 second windowed interval. The actual utterance in the speech samples is between 0.8 and 1.5 second, based on perceptual evaluation. This means that around 20% to 35% of the music signal is affected by speech if the music duration fits the windows perfectly. In reality this is not the case, as the average duration of all music samples is 45.9 seconds, but there are some strong outliers. For simplicity sake we estimated the average speech-music overlap to be about 25%, a conservative guess.

While it is hard to say now, and even harder to know beforehand, but it could have been the case that 25% speech-music overlap was not enough to push the adaptability of the models to mixed audio to their limits, as similar or even better levels of valence and arousal prediction were achieved on mixed audio. However, we saw that the dominant features used to predict valence and arousal differed strongly between all three audio types. This suggests that the model had to, and was able to adapt to the speech presence.

Reflecting back, a higher overlap rate might have been better, but we think that 25% is enough to prove that music emotion recognition is possible in a MiSME scenario. Our results show that the music models can adapt to mixed audio with up to 25% speech-music overlap without (much) performance loss. We think that this amount of overlap is quite representative for many MiSME scenarios, such as some genres of television content. However, it should not be assumed that the results generalize to scenarios with higher levels of overlap.

Limited to speculation This brings us to another aspect that we want to adress, the speculative nature of some observations mentioned in Section 5.2. It is unclear how the music models were able to outperform their single-source performance. It raised many questions, as we assumed that the performance on music-only audio would have been optimal. While we speculate that the presence of speech helps the models to distinguish between ambiguous and non-ambiguous features, ultimately resulting in a better generalizing model, this remains only speculation. This occurrence was not foreseen and there was not enough time and resources left to explore the cause. However, we think that this can serve as a good starting point for future research.

[4] "Kids are talking by the door" and "Dogs are sitting by the door"

Feature analysis limitations The final thing that we want to address is the limited scope of the feature analysis, especially regarding that most observations could only be clarified through speculation. This was due to time limitations. These observations would require significant research by someone with sufficient expertise to properly answer. Speculation allowed us to at least create some 'food for thought', which hopefully motivates others to study these observations further.

Nevertheless, this research has successfully proven that a MiSME system can be created, without the use of any contextual relationship between the speech and music. Not only has it done that, which was the main motivation for this research, but we also produced additional insight into the MiSME recognition problem through the use of multiple models, an in-depth feature importance analysis, the use of blind-source separation and R128 loudness normalization, the creation of our own mixed audio dataset and more.

In our opinion this research serves as a solid starting point for understanding the problem of MiSME recognition. Our results and observations should be able to aid others, including RTL The Netherlands, in creating their own MiSME system. In addition to that this research has also produced many potential topics for future work. How well do more complex neural networks perform on various MiSME recognitions tasks? How is performance affected when there is a contextual relationship present between the music and speech, for example when developing a MiSME system for television content? Are models able to use this relationship for more accurate predictions? And what causes the music models to achieve better performance when speech is present? These are just some of the many interesting questions which remained unanswered.

Chapter 6

Conclusion

This research studied the problem of separate speech and music emotion recognition on audio where both appear concurrently, coined as MiSME recognition. This research was done in collaboration with RTL The Netherlands, as they desire to create such a system. The goal was to not only prove that a 'functioning' MiSME system can be created, but also to gain as much insight as possible into MiSME recognition, laying a foundation for others to work with in the future. The MiSME system was deemed 'functional' if it could do both speech and music emotion recognition significantly above chance-level on any form of mixed audio. The decision was to study MiSME recognition from a general perspective by excluding any contextual relationship between the speech and music. This allowed us to gain insight into how speech and music emotion recognition are affected by mixed audio regardless of context, making the results applicable to all kinds of future use cases.

The MiSME system consisted of separate speech and music emotion recognition models. A blind-source separation component was included in some version of the MiSME system because we believed it could aid the models with emotion recognition on mixed audio. The speech and music emotion recognition capabilities of the system were tested with a separate set of experiments. These experiments were various combinations of training and testing on three different audio types: speech-only or music-only audio, mixed audio and blind-source separated audio. A mixed speech-music dataset was created specifically for the experiments, with an average speech-music overlap of 25%.

The speech experiments showed that speech-only trained models were incapable of speech emotion recognition on both mixed and blind-source separated audio, as accuracy degraded to near chance level. Training models on mixed audio resulted in above chance-level performance. However, the inclusion of blind-source separation appeared to be highly beneficial, as the models saw a significant increase in accuracy when including it in their pipeline. The results thus proved that a functioning speech emotion recognition model for mixed audio can be created, but there remained a performance gap compared to speech-only audio.

The music experiments showed that music-only trained models were also incapable of music emotion recognition on both mixed and blind-source separated audio, as performance was around chance-level. However, something unexpected occurred when the models were trained and tested on either mixed or blind-source separated audio. One model, the Deep Neural Network, was able to achieve significantly better performance on mixed audio than on music-only audio. The other model, a Random Forest model, was able to do the same when blind-source separation was included. We assumed that the performance on music-only audio would have been 'optimal', as it does not suffer from speech interference, but it appears that that is not the case. We speculated that the presence of speech forces the music models to drop more ambiguous features in favor of less ambiguous ones, resulting in a better generalizable model. Regardless of this occurrence the results show that music emotion recognition can be done equal to, or even better than, music-only audio on mixed audio.

The results thus show that both speech and music emotion recognition can be done far above chance-level on mixed audio, proving that a ‘functioning’ MiSME system can be created.

In addition a speech and music feature importance analysis was performed, which included many commonly used speech and music features. It showed that the dominant features for each audio type differed almost completely, for both speech and music emotion recognition. This means that vastly different speech and music features are optimal for MiSME recognition. We also identified which type of speech and music features are of high importance for MiSME recognition, and which are not.

To conclude this research: we have proven that a functioning MiSME system can be created, as we were able to create one that performed far above chance-level on both the speech and music emotion recognition task. The knowledge produced in this research regarding MiSME recognition, through the results, multiple feature analyses, the creation of the first MiSME dataset and more, should allow others to create a MiSME system of their own, including RTL The Netherlands. However, we have only explored the surface of the MiSME recognition problem space. We can only hope that this research can serve as a starting point for future research, hopefully motivating others to study more specific or complex problems and, of course, to create MiSME systems of their own.

Bibliography

Aljanaki, A. (2016). *Emotion in Music: representation and computational modeling.* PhD thesis, Utrecht University.

Aljanaki, A., Wiering, F., and Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management*, 52(1):115–128.

Aljanaki, A., Yang, Y.-H., and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392–e0173392. 28282400[pmid].

Anagnostopoulos, C.-N., Iliou, T., and Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57.

Bachorowski, J.-A. and Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological science*, 6(4):219–224.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.

Bänziger, T. and Scherer, K. R. (2010). Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, pages 271–294.

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46.

Bhavan, A., Chauhan, P., Shah, R. R., et al. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, page 104886.

Boersma, P. and Weenink, D. (2007). Praat: Doing phonetics by computer (version 5.3.51).

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., Serra, X., et al. (2013). Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR).

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. (2019). The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States.

Bou-Ghazale, S. E. and Hansen, J. H. L. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto (2015). librosa: Audio and Music Signal Analysis in Python. In Kathryn Huff and James Bergstra, editors, *Proceedings of the 14th Python in Science Conference*, pages 18 – 24.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Univ of California Press.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology.*

Cabrera, D., Ferguson, S., Rizwi, F., and Schubert, E. (2008). Psysound3: a program for the analysis of sound recordings. *The Journal of the Acoustical Society of America*, 123:3247.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Chen, S.-H., Lee, Y.-S., Hsieh, W.-C., and Wang, J.-C. (2015a). Music emotion recognition using deep gaussian process. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 495–498. IEEE.

Chen, Y.-A., Yang, Y.-H., Wang, J.-C., and Chen, H. (2015b). The amg1608 dataset for music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 693–697. IEEE.

Clynes, M. (1977). *Sentics: The touch of emotions.* Anchor Press.

Cooke, D. (1959). The language of music.

Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1970–1973. IEEE.

Demir, C., Cemgil, A. T., and Saraçlar, M. (2010). Catalog-based single-channel speech-music separation. In *Eleventh Annual Conference of the International Speech Communication Association.*

Demir, C., Saraclar, M., and Cemgil, A. T. (2012). Single-channel speech-music separation for robust asr with mixture models. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):725–736.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

EBU-Recommendation, R. (2011). Loudness normalisation and permitted maximum level of audio signals.

Eerola, T. and Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49.

Eerola, T. and Vuoskoski, J. K. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, 30(3):307–340.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ekman, P., Friesen, W. V., and Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Gabrielsson, A. and Juslin, P. N. (2003). *Emotional expression in music.* Oxford University Press.

Grais, E. M. and Erdogan, H. (2011). Single channel speech music separation using nonnegative matrix factorization and spectral masks. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–6. IEEE.

Grais, E. M., Sen, M. U., and Erdogan, H. (2014). Deep neural networks for single channel source separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3734–3738. IEEE.

Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H., and Li, X. (2018). Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In *Interspeech*, pages 1611–1615.

Han, B.-j., Rho, S., Dannenberg, R. B., and Hwang, E. (2009). Smers: Music emotion recognition using support vector regression. In *ISMIR*, pages 651–656.

Hansen, J. H. L. and Bou-Ghazale, S. E. (1997). Getting started with susas: a speech under simulated and actual stress database. In *EUROSPEECH*.

Hasan, M. R., Jamil, M., Rahman, M., et al. (2004). Speaker identification using mel frequency cepstral coefficients. *variations*, 1(4).

Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M. (2019). Spleeter: A fast and state-of-the art music source separation tool with pre-trained models. Late-Breaking/Demo ISMIR 2019. Deezer Research.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268.

Hevner, K. (1937). The affective value of pitch and tempo in music. *The American Journal of Psychology*, 49(4):621–630.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., and Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. In *LREC*.

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. (2017). Singing voice separation with deep u-net convolutional networks.

Juslin, P. N. and Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of new music research*, 33(3):217–238.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proc. ISMIR*, volume 86, pages 937–952.

Kramarz, A. (2017). Is the idea of 'musical emotion'present in classical antiquity? *Greek and Roman Musical Studies*, 5(1):1–17.

KUCHINKE1A, L., KAPPELHOFF1B, H., and KOELSCH1C, S. (2013). Emotion and music in narrative films: a neuroscientific perspective.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*.

Langer, S. K. (1953). *Feeling and form*, volume 3. Routledge and Kegan Paul London.

Langer, S. K. (2009). *Philosophy in a new key: A study in the symbolism of reason, rite, and art*. Harvard University Press.

Lartillot, O., Toiviainen, P., and Eerola, T. (2008). A matlab toolbox for music information retrieval. In Preisach, C., Burkhardt, H., Schmidt-Thieme, L., and Decker, R., editors, *Data Analysis, Machine Learning and Applications*, pages 261–268, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Li, T. and Ogihara, M. (2003). Detecting emotion in music.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Luo, Y., Chen, Z., Hershey, J. R., Le Roux, J., and Mesgarani, N. (2017). Deep clustering and conventional networks for music separation: Stronger together. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65. IEEE.

Manilow, E., Seetharaman, P., and Pardo, B. (2018). The northwestern university source separation library.

Markov, K. and Matsui, T. (2014). Music genre and emotion recognition using gaussian processes. *IEEE access*, 2:688–697.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., and Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Neiberg, D., Elenius, K., and Laskowski, K. (2006). Emotion recognition in spontaneous speech using gmms. In *Ninth international conference on spoken language processing*.

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

Olsen, K. N., Dean, R. T., Stevens, C. J., and Bailes, F. (2015). Both acoustic intensity and loudness contribute to time-series models of perceived affect in response to music. *Psychomusicology: Music, Mind, and Brain*, 25(2):124.

Panda, R., Rocha, B., and Paiva, R. P. (2015). Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence*, 29(4):313–334.

Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1175–1191.

Roma, G., Grais, E. M., Simpson, A., Sobieraj, I., and Plumbley, M. D. (2016). Untwist: A new toolbox for audio source separation. In *Extended abstracts for the late-breaking demo session of the 17th international society for music information retrieval conference, ismir*, pages 7–11.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Salaün, Y., Vincent, E., Bertin, N., Souviraa-Labastie, N., Jaureguiberry, X., Tran, D. T., and Bimbot, F. (2014). The flexible audio source separation toolbox version 2.0.

Sato, N. and Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(3):835–848.

Schmidt, E. M. and Kim, Y. E. (2011). Learning emotion-based acoustic features with deep belief networks. In *2011 IEEE workshop on applications of signal processing to audio and acoustics (Waspaa)*, pages 65–68. IEEE.

Schmidt, E. M., Turnbull, D., and Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the international conference on Multimedia information retrieval*, pages 267–274.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (2010). The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

Scikit-learn (2007). Diagram of various kernels on the iris flower dataset. https://scikit-learn.org/stable/modules/svm.html.

Seehapoch, T. and Wongthanavasu, S. (2013). Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)*, pages 86–91. IEEE.

Seo, Y.-S. and Huh, J.-H. (2019). Visualisation of russell's circumplex model. https://www.researchgate.net/publication/330817411/figure/fig1/AS:721752380411904@1549090585250/Russells-circumplex-model-The-circumplex-model-is-developed-by-James-Russell-In-the.jpg.

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., and Yang, Y.-H. (2013). 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 1–6. ACM.

Song, Y., Dixon, S., and Pearce, M. T. (2012). Evaluation of musical features for emotion classification. In *ISMIR*, pages 523–528. Citeseer.

Stoller, D., Ewert, S., and Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*.

Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 4:1667.

Swain, M., Routray, A., and Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120.

Teagar, H. and Teagar, S. (1990). Evidence for nonlinear production mechanisms in the vocal tract. In *Speech Production and Modelling*. Kluwer.

Tian, L., Moore, J. D., and Lai, C. (2015). Emotion recognition in spontaneous and acted dialogues. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 698–704. IEEE.

Tits, N., Haddad, K. E., and Dutoit, T. (2019). The theory behind controllable expressive speech synthesis: a cross-disciplinary approach. https://www.researchgate.net/profile/Noe_Tits/publication/336550610/figure/fig4/AS:814109289349123@1571110188235/Diagram-describing-voice-production-mechanism-and-source-filter-model.ppm.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

Uhlich, S., Giron, F., and Mitsufuji, Y. (2015). Deep neural network based instrument extraction from music. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Victor, Z. (2019). Decision tree example. https://victorzhou.com/media/random-forest-post/decision-tree2.svg.

Wang, K., An, N., Li, B. N., Zhang, Y., and Li, L. (2015). Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75.

Wikipedia, the free encyclopedia (2019). Diagram of k-fold cross-validation. https://en.wikipedia.org/wiki/Cross-validation_(statistics)#/media/File:K-fold_cross_validation_EN.svg.

Wundt, W. M. and Judd, C. H. (1897). *Outlines of psychology*, volume 1. Scholarly Press.

Xu, J., Li, X., Hao, Y., and Yang, G. (2014). Source separation improves music emotion recognition. In *Proceedings of International Conference on Multimedia Retrieval*, page 423. ACM.

Yang, X., Dong, Y., and Li, J. (2018). Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24(4):365–389.

Yang, Y.-H. and Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, 16(2):448–457.

Zeng, Y., Mao, H., Peng, D., and Yi, Z. (2019). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 78(3):3705–3722.

Zheng, W., Yu, J., and Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 827–831. IEEE.

Appendix A

Music experiments - Hexbin heatmaps

A set of six valence- and arousal heatmaps were produced that plot the predictions against the actual ground truth values. The heatmaps depict the Random Forest- and Deep Neural Network model on experiment A, D and F (see Table 4.6) and the Random Forest model using the top 50 most important valence features and top 100 most important arousal features, the optimal feature set.

The bin-size was set to 0.25. The color mapping was capped to 120 samples, but there are three bins with outliers. Using the full range would reduce the readability of these heat-maps, as they would make the lower bins less visually distinctive. These outliers are reported separately instead.

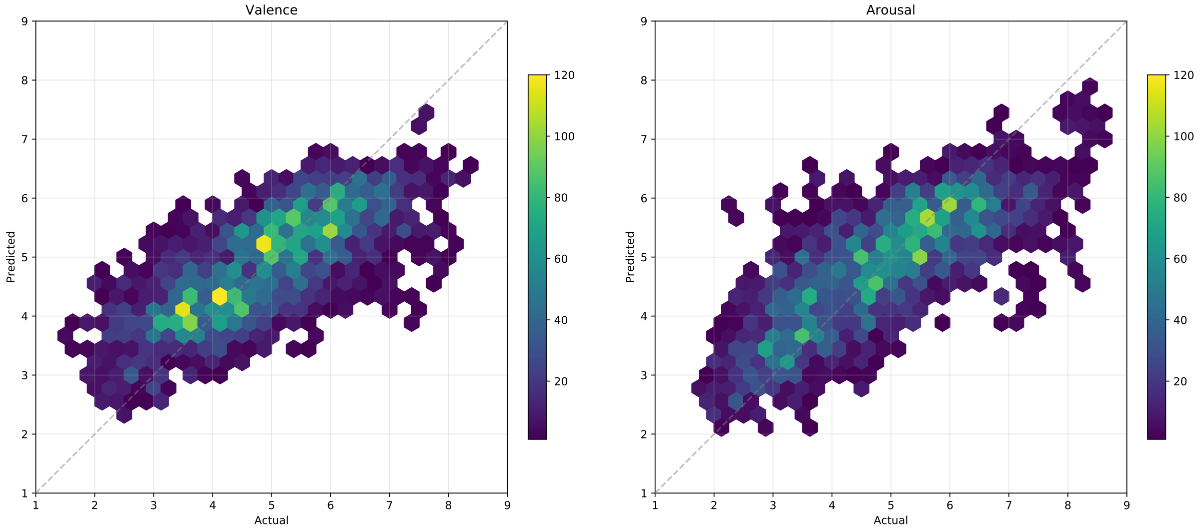
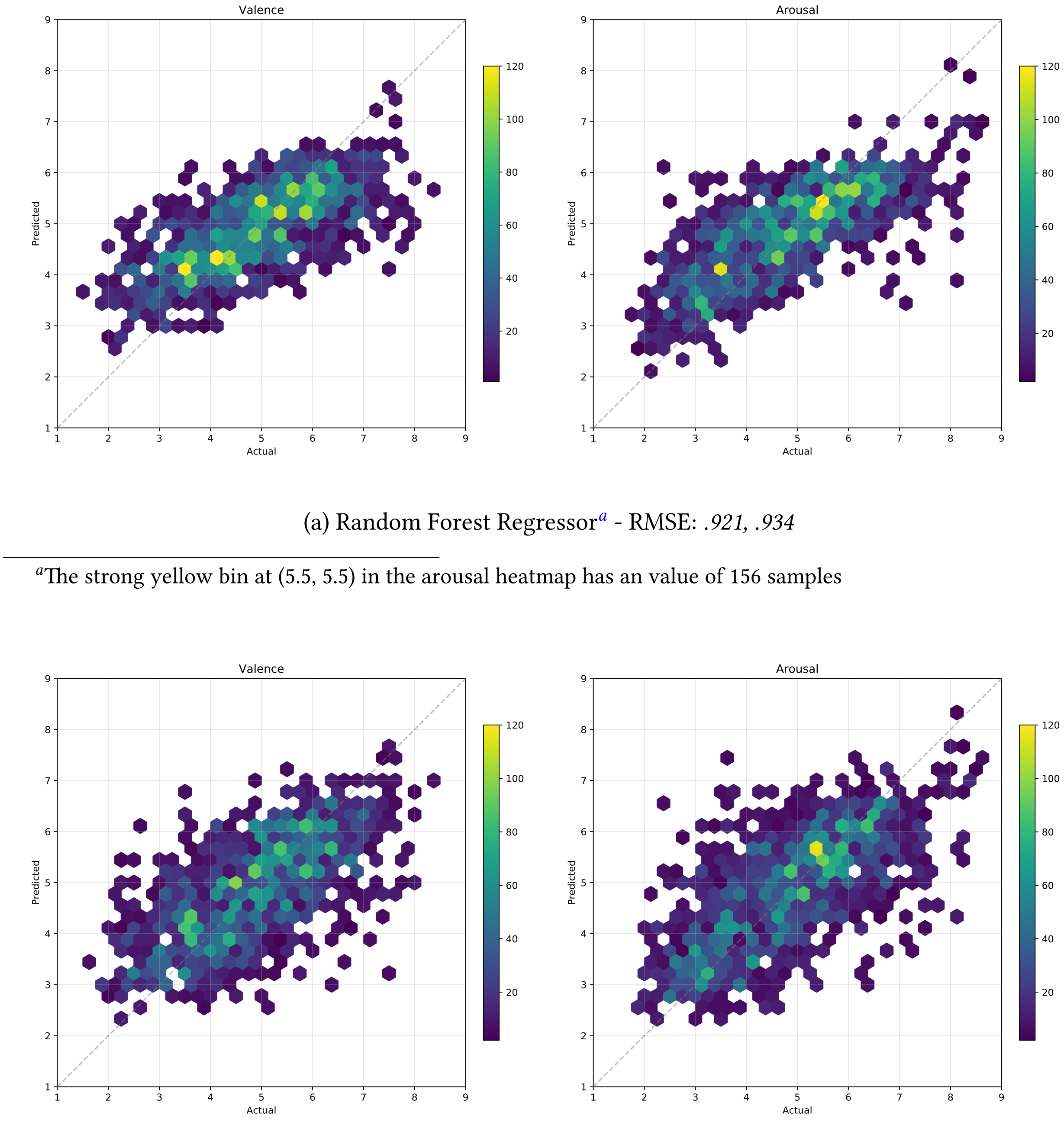


Figure A.1: The RFR using the 50 most important valence features and 100 most important arousal features on BSS audio - RMSE: *.84, .857*

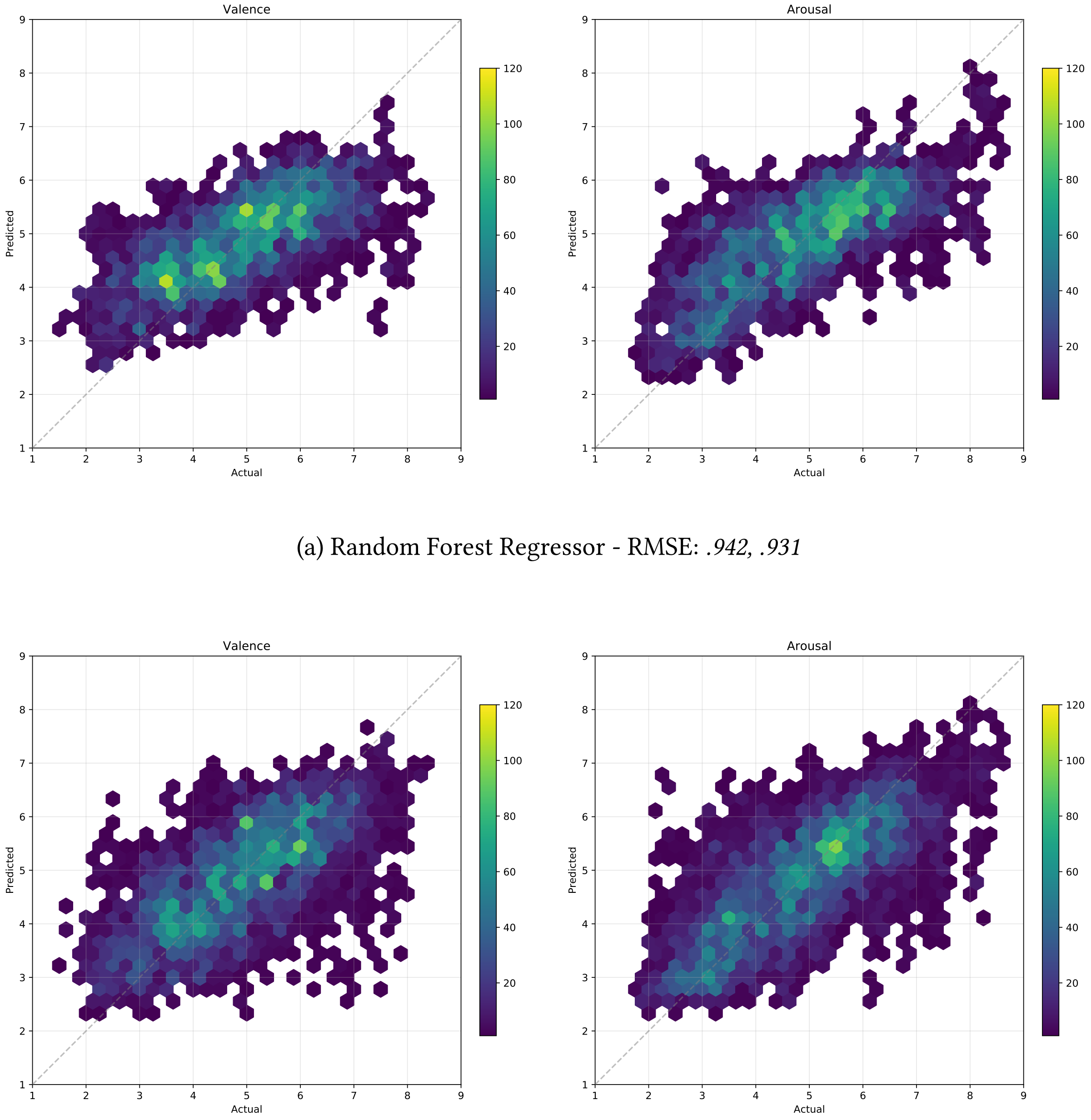


(a) Random Forest Regressor[a] - RMSE: *.921, .934*

[a] The strong yellow bin at (5.5, 5.5) in the arousal heatmap has an value of 156 samples

(b) Deep Neural Network - *RMSE: .978, 1.024*

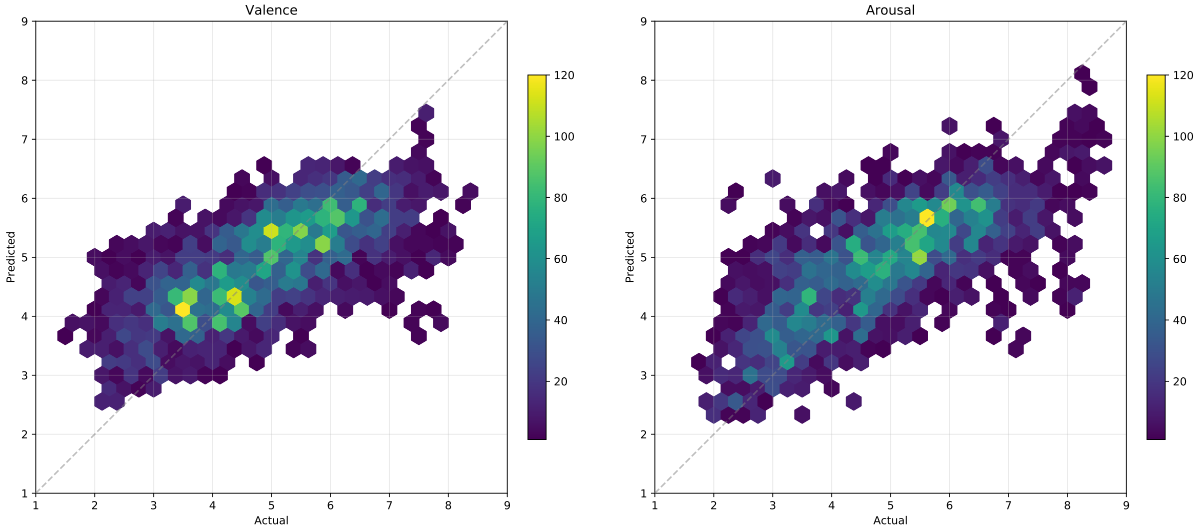
Figure A.2: experiment A (Table 3.1)



(a) Random Forest Regressor - RMSE: *.942, .931*

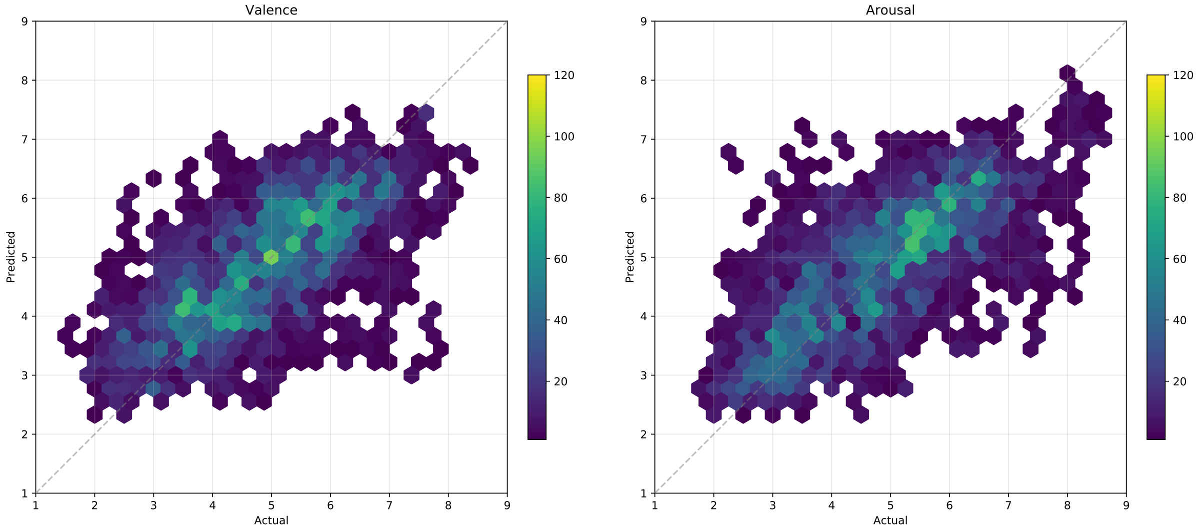
(b) Deep Neural Network - RMSE: *.947, .96*

Figure A.3: experiment D (Table 3.1)



(a) Random Forest Regressor[a] - RMSE: *.909, .928*

[a] The strong yellow bin in the valence heatmap (3.3, 4.2) has an value of 136 samples and the strong yellow bin in the arousal heatmap (5.6, 5.6) has an value of 130 samples, both outliers

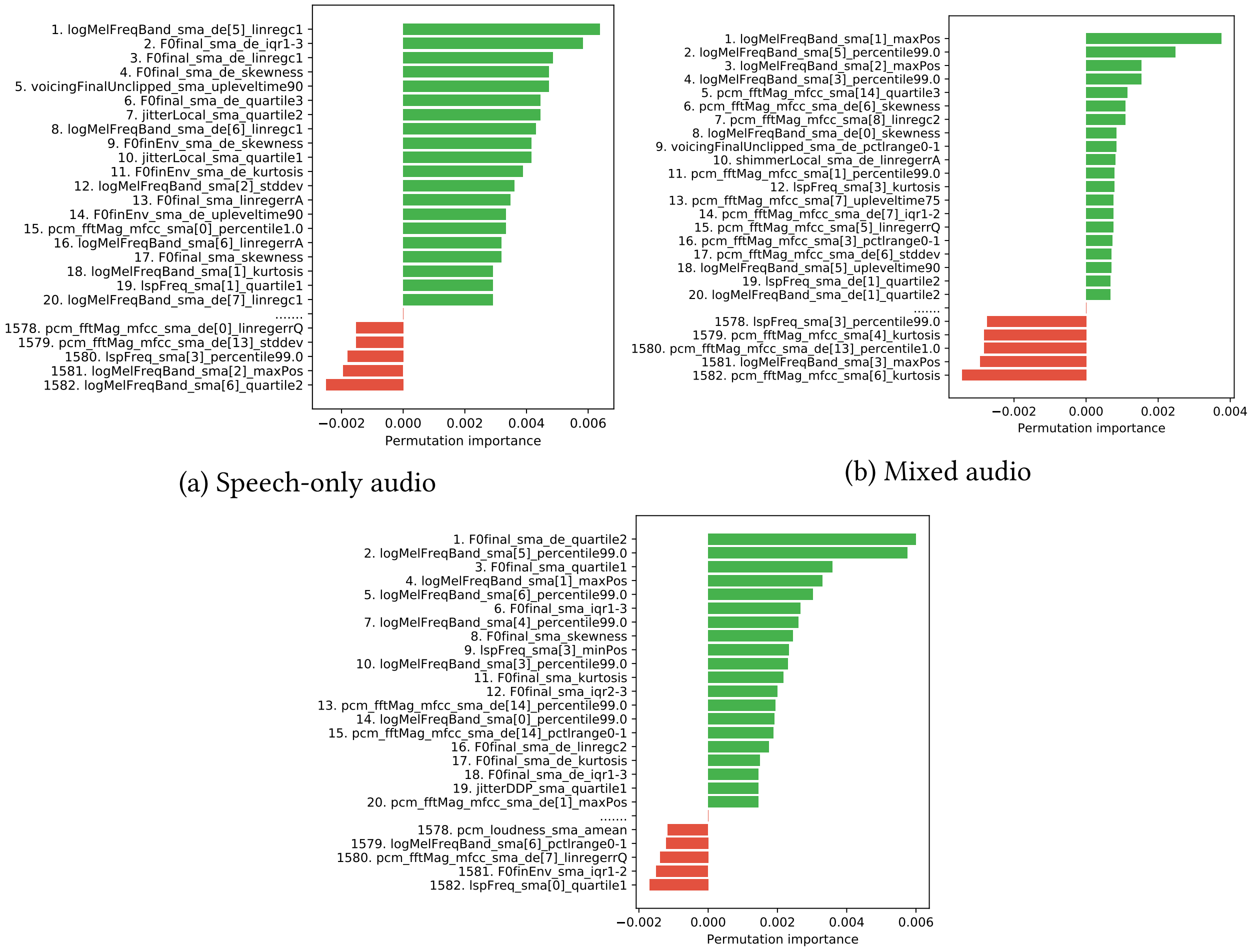


(b) Deep Neural Network - *RMSE: .943, .979*

Figure A.4: experiment F (Table 3.1)

Appendix B

Feature importance analysis - Speech

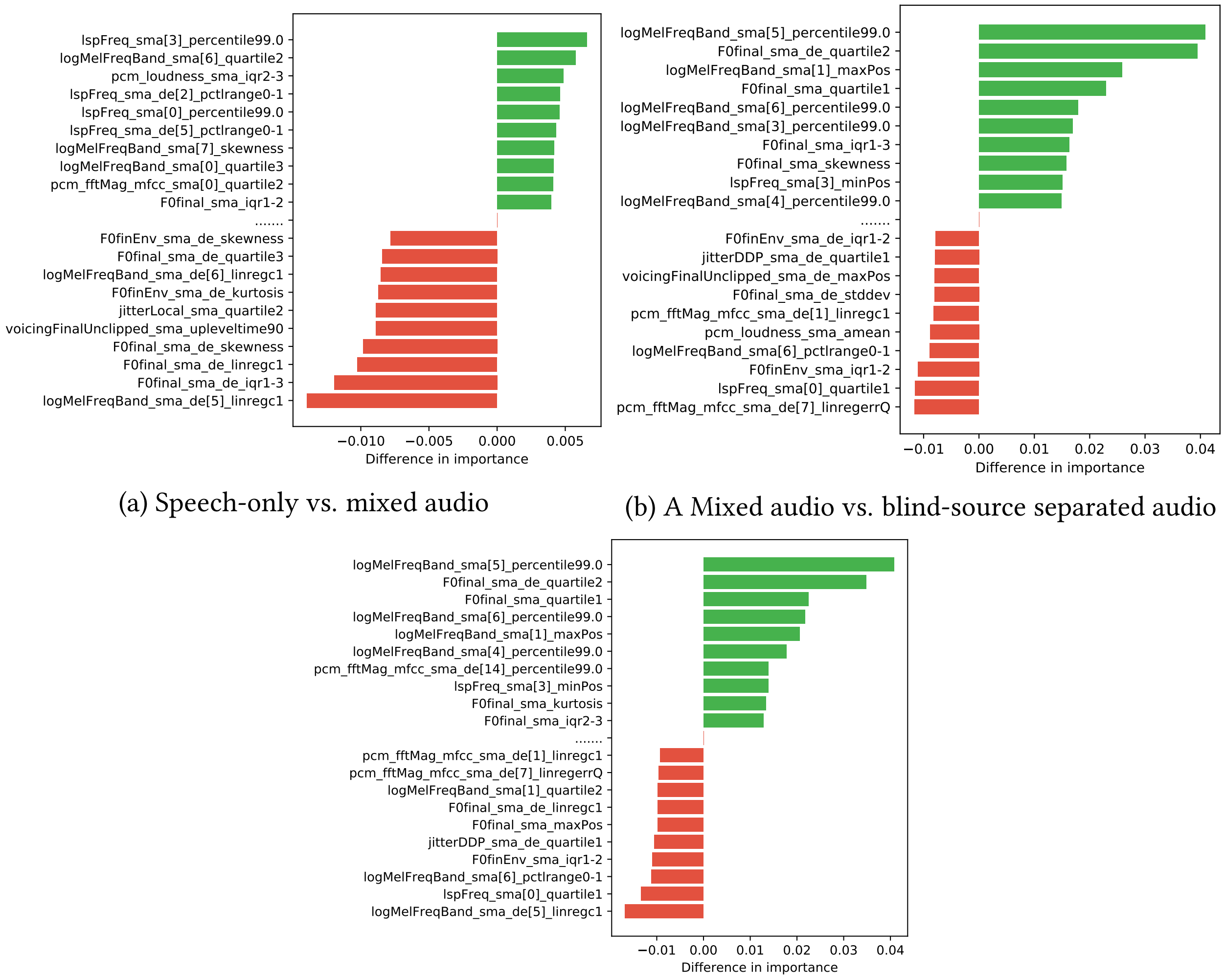


(a) Speech-only audio

(b) Mixed audio

(c) Blind-source separated speech audio

Figure B.1: The most important speech features for the Random Forest Model on all three audio types



(a) Speech-only vs. mixed audio

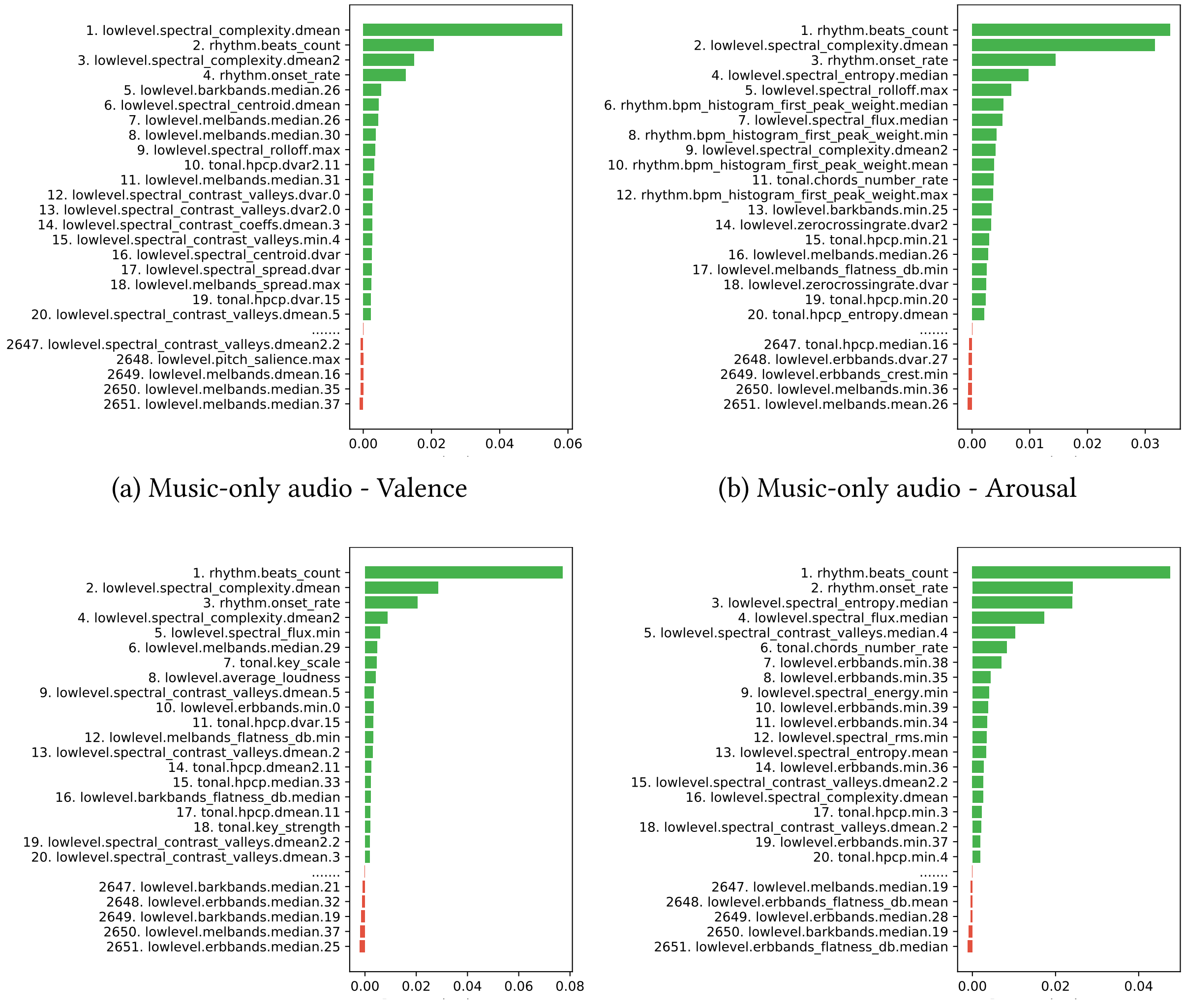
(b) A Mixed audio vs. blind-source separated audio

(c) Speech-only vs. blind-source separated audio

Figure B.2: The most strongly differing speech features importance-wise between any two audio types for the Random Forest model

Appendix C

Feature importance analysis - Music



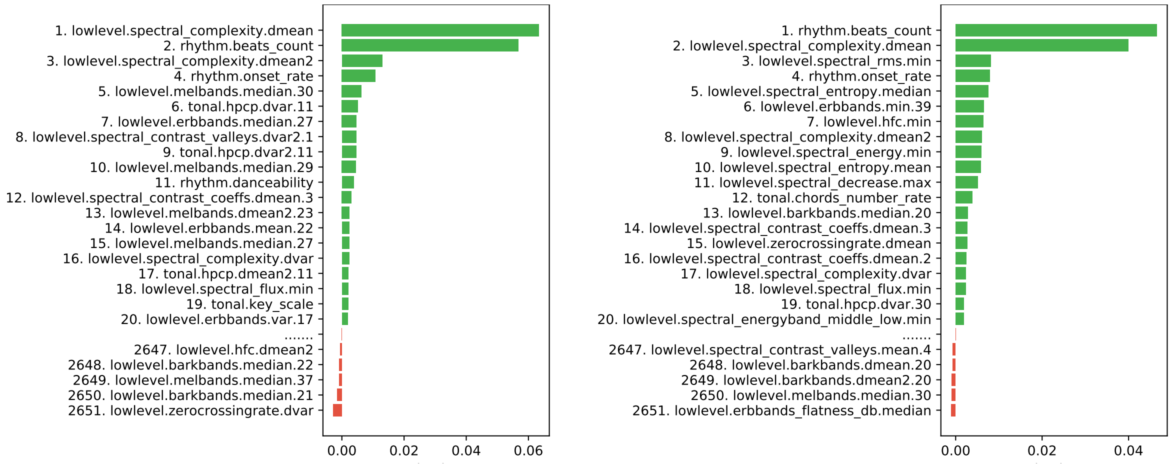
(a) Music-only audio - Valence

(b) Music-only audio - Arousal

(c) Mixed audio - Valence

(d) Mixed audio - Arousal

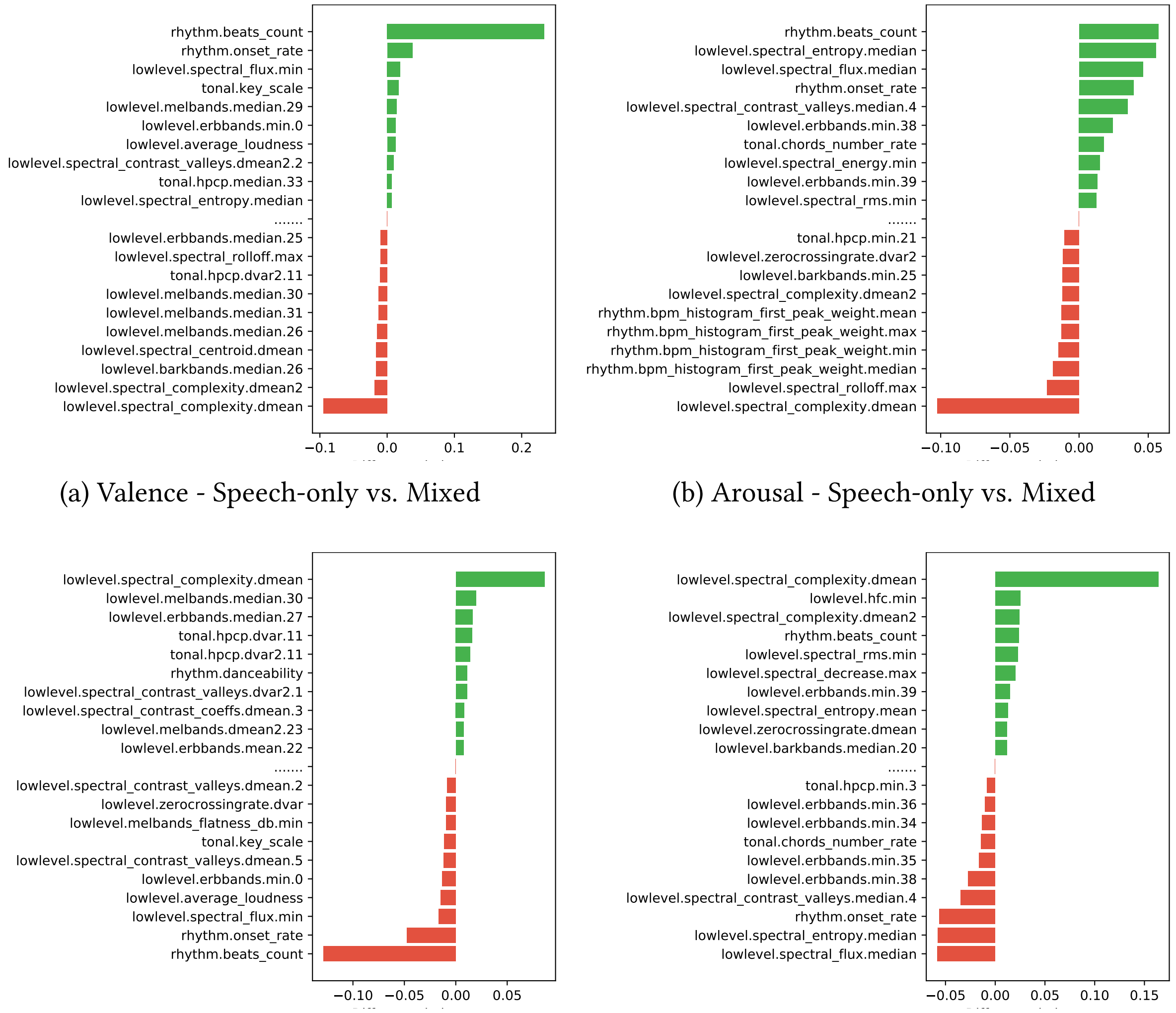
Figure C.1: The most important- and unimportant music features for valence- and arousal prediction for the Random Forest Model on all three audio types



(e) Blind-source separated audio - Valence

(f) Blind-source separated audio - Arousal

Figure C.1: Continuation of Figure C.1



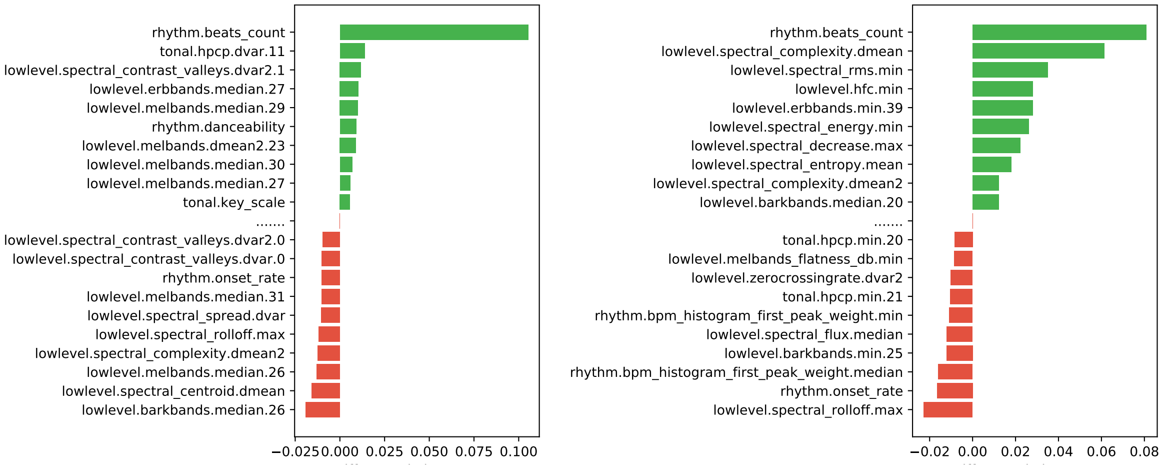
(a) Valence - Speech-only vs. Mixed

(b) Arousal - Speech-only vs. Mixed

(c) Valence - Mixed vs. BSS

(d) Arousal - Mixed vs. BSS

Figure C.2: The most strongly differing music features importance-wise between any two audio types for the Random Forest model regarding valence- and arousal prediction



(e) Valence - Speech-only vs. BSS

(f) Arousal - Speech-only vs. BSS

Figure C.2: Continuation of Figure C.2

Appendix D

Music experiments - Diebold Mariano tests

Experiment A	Experiment B	P-value	
Model	**Model**	**Valence**	**Arousal**
RFR - SS (A)	*DNN - SS* (A)	6.76e-14	1.58e-26
RFR - SS (A)	*RFR - SM* (B)	1.19e-131	1.15e-142
DNN - SS (A)	*DNN - SM* (B)	1.33e-197	7.05e-195
RFR - SS (A)	*RFR - SB* (C)	7.44e-45	9.47e-76
RFR - SM (B)	*RFR - SB* (C)	2.8-50	6.44e-45
DNN - SS (A)	*DNN - SB* (B)	1.42e-32	1.91e-19
DNN - SM (B)	*DNN - SB* (C)	1.51e-138	2.12e-147
RFR - SS (A)	*RFR - MM* (D)	6.28e-07	0.475
DNN - SS (A)	*DNN - MM* (D)	3.374e-12	1.429e-29
RFR - SS (A)	*RFR - BB* (F)	0.001	0.142
RFR - MM (D)	*RFR - BB* (F)	8.124e-14	0.467
DNN - SS (A)	*DNN - BB* (F)	6.782e-15	3.231e-21
DNN - MM (D)	*D-NN - BB* (F)	0.526	0.005

Table D.1: Diebold-Mariano tests done on the results obtained from the six main music experiments, as described in Section 4.3.3. P-values < 0.005 are considered *significant*, meaning that the RMSE performance between the two compared experiments is significantly different from each other considering the dataset used. Each experiment is notated using the following format: <*model type*> <*letter indicating training audio type*><*letter indicating testing audio type*> (Experiment from Table 3.1), for which the following letters are used: S = single-source, M = mixed source, B = blind-source separated. E.g., RFR - SB (C) means the Random Forest model trained on single-source audio, but tested on blind-source separated audio. That would be experiment C, see Table 3.1

Appendix E

Music experiments - Shared features between all three audio types

- lowlevel.melbands_flatness_db.median
- lowlevel.melbands_flatness_db.min
- lowlevel.melbands.median.29
- rhythm.onset_rate
- rhythm.beats_count
- lowlevel.barkbands_flatness_db.median
- lowlevel.spectral_contrast_valleys.dmean.2
- lowlevel.spectral_contrast_valleys.dmean.3
- lowlevel.spectral_contrast_valleys.dmean.1
- tonal.hpcp.dmean2.14
- tonal.hpcp.dvar.15
- tonal.hpcp.dmean2.11
- lowlevel.spectral_complexity.dmean2
- lowlevel.spectral_complexity.dmean

Figure E.1: The fourteen shared features among the 50 most important features for valence prediction on all audio types

- rhythm.beats_count
- rhythm.onset_rate
- tonal.chords_number_rate
- lowlevel.spectral_complexity.dmean
- lowlevel.spectral_entropy.mean
- lowlevel.spectral_entropy.median
- lowlevel.spectral_contrast_valleys.dmean2.2
- lowlevel.spectral_contrast_valleys.dvar.1
- lowlevel.spectral_contrast_valleys.median.4
- lowlevel.spectral_flux.median

Figure E.2: The ten shared features among the 50 most important features for arousal prediction on all audio types