Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

End-to-End Music Emotion Recognition: Towards Language-Sensitive Models

Ana Gabriela Pandrea

Supervisor: Juan Sebastián Gómez Cañón

Co-Supervisor: Perfecto Herrera

September 2020

Contents

Acknowledgement

I would like to express my sincere gratitude to:

- My supervisor, for his constant support, implication and positivity
- My co-supervisor, for his in-depth and professional remarks
- My classmates, for constantly inspiring me with their passion

Abstract

Music, like any other art, represents an expression of emotions and moods, thus, a natural question that can be raised for both human and computer is 'Which specific emotion was meant to be transmitted in a certain music excerpt?'. The Music Emotion Recognition (MER) field has been exploited for several years, but it still faces some challenges because emotion is a very subjective aspect. One problem could be that because each language has its particularities in terms of sound and intonation, and implicitly in terms of associations that are made upon them, we expect the observed emotions to be different from one culture to another. To address this issue, we choose a more natural, human-like approach towards emotion detection and propose a language sensitive supervised model that learns to tag emotions from music datasets with lyrics in different languages. Other studies have shown that emotion interpretations seem to differ in terms of the valence of emotion, that is how positive or negative it is perceived. We aim to investigate this phenomenon by training a novel end-to-end model independently for music in English, Mandarin and Turkish. The architecture is called SincNet and was initially proved to be successful for the task of speaker recognition. Our results with SincNet for MER are not very good when trained on individual datasets, but show promising results under several transfer learning set-ups, where general cues are learned from all datasets, but fine-tuning on the target test set gives more sensitive results.

Keywords: Music Emotion Recognition; SincNet; End-to-End Learning; Audio Classification

Chapter 1

Introduction

Music is perhaps one of the most spread and loved art in the world. In fact, many cultures see it as a necessary part of the normal life, without calling it art. As technology evolves, music becomes increasingly accessible and diverse, such that it reaches the soul of each and every one of us. This happens due to its capacity of triggering strong feelings and changes in our internal states, i.e., emotions. The subject of identifying patterns and reasons behind the connection of music and emotions is an open question for multidisciplinary research. Particularly relevant would be to determine which characteristics of music influence the various emotions we perceive.

In this thesis, we aim to explore the field of Music Emotion Recognition in an attempt to improve the current state-of-the-art methods. We will consider some approaches that have already been implemented in other contexts and we will also propose a set of experiments of our own. While the topic itself could be considered a human task, that is a topic of psychology and emotion studies, here we will focus on the automatic music emotion recognition problem. In this context, a very important role is played by Machine Learning theory and algorithms. By this means, we teach the machine how to identify emotions in music excerpts by training it with many human annotated excerpts. However, there are various ways to tune these algorithms and also to pre-process the input data, therefore the perfect solution has not been

identified yet.

Music Emotion Recognition has developed as part of the Music Information Retrieval field, with applications in music recommendation and search, playlist creation, but also in therapy and marketing. The main challenges of the field come from the ambiguity and subjectivity of emotions, where ambiguity refers to the interpretability and lack of exact delimitations for different emotions and subjectivity stands for the lack of agreement between different annotators. There is also a confusion between the induced and perceived emotions, in the sense that what we personally experience might be different to the emotions evoked by music only. Different people can experience different states based on our their own previous experiences and correlations, whereas the composer along with the interpreter express a single state, based on their personal emotions and feelings. In their study, Holzapfel et al. [1] also identified several issues that are general for the Music Information Retrieval field: the diverse cultural biases, the questionable value of the widely spread datasets, as well as the remoteness of MIR research from actual musicians and music specialists. Thus, there are many challenges to be defeated that also include figuring out the best way to take advantage of the extra-musical information, for example culture or education. Various pre-extracted feature sets were proposed in order to leverage these issues, as well as various deep learning architectures that directly retrieve information from spectrograms.

Despite all these, the field has not been approached too much with the end-to-end learning strategy that encodes and learns from the raw waveform input, therefore we propose it here as a relatively novel contribution. The main feature of the end-to-end approach is that all the music analysis happens inside the model, which is fed with the raw audio waveform and outputs the appropriate emotion category. In this way, we classify music segments under a mixture of dimensional and categorical taxonomy, in one of the four emotion classes determined by the Valence-Arousal plane. The two axis of the plane mainly emphasize how positive or negative music sounds like and also how much energy it expresses.

One aspect that will be emphasized in our study is the role of language and cul-

ture in perceiving emotions in music. This is a challenging task since most of the available studies and datasets are made with only Western music and annotators. Dataset creation and annotators agreement are some big issues of the field and we are going to explore these by performing experiments with three datasets with music of different cultures, more exactly English, Chinese and Turkish. Our results show that the three datasets indeed behave differently throughout multi-cultural experiments, suggesting a need for context-based models. Perhaps more data and human annotators would be required in order to approve these inferences.

It is worth mentioning that our results will have been presented at the Audio Engineering Society (AES) Machine Learning (ML) Symposium[1], as part of the breakout sessions, and also at the 21st International Society for Music Information Retrieval (ISMIR) Conference, as a Late-Breaking/Demo (LBD) Contribution. The AES précis will be provided in Appendix B just like it was submitted and accepted to the symposium, while the extended abstract accepted at ISMIR[2] will be provided in Appendix C.

1.1 Motivation

Music Emotion Recognition (MER) has grown to be an important part of the Music Information Retrieval field. One of its goals is to narrow down the so-called "semantic gap" between the physical properties of the audio signal and the semantic concepts characteristic to humans, in this case emotions.

Music Emotion Recognition is valuable in the scientific background in terms of both business applications and academic discoveries. In this last sense, we believe that identifying emotions in music brings us a step closer to building emotion-sensitive machines and robots and it also brings us closer to simulating our brains and understanding ourselves better. While the human mind is extremely complex and made of so many variables, we aim to add another layer of understanding by bringing to the table the cultural component. This is because the sounds we are trained to listen

[1] https://www.aes.org/events/2020/learning/
[2] https://ismir.github.io/ISMIR2020/

to in our own language, along with the music we grow up with, are believed to play important roles in modelling our taste and emotional experience in music.

In terms of real life applications, emotion recognition would be very relevant to music recommendation systems, search systems and emotion categorized playlists. These can be considered for individual listeners and music enthusiasts, but also for big industries like gaming or film. A further appealing application would be on-the-spot emotion detection in human and playing appropriate music to match the current mood. Moreover, it can also have various applications in therapy and emotion regulation and due to its ability of making customers emotionally involved, music can even be used for marketing purposes, based on the states it transmits.

A personal motivation for the project stands behind the fact that music is capable of moving people to a very deep level and it is only fascinating to discover how and why that happens. Coming from a scientific background, this type of curiosity only pushes us further to create models and experiments to find explanations for the things that matter to us, as humans. Moreover, training the computer to make sense of emotional content is a breakthrough that can make it our friend even more than it already is in our daily tasks.

1.2 Objectives

In order to fulfill the existent academic desire, but also the gaps that exist in the MER field, we established several objectives for our research. These are related to several technical aspects and possibilities, but the observations, analysis and discussions around these also play a significant role.

Our main purpose with this study is experimenting with a new end-to-end machine learning approach and concluding whether this is a promising approach for Music Emotion Recognition. The neural network we are considering is called SincNet and was previously successful in several speech related tasks [2]. We aim to discover whether such a language related architecture would be useful in the automatic identification of emotions in music.

In addition, we also consider and compare a few baseline models that involve traditional machine learning techniques and feature extraction as a pre-processing step. The goal of these experiments is to provide a strong baseline model for our three datasets, compare these results to previous similar attempts and evaluate the performance of the SincNet deep learning architecture. In comparison to the baseline models that use certain selected features as input, SincNet uses directly the raw waveforms of the music excerpts, therefore another aim is to observe the behaviour of this approach in terms of input format.

Last but not least, this project investigates how relevant language and cultural considerations are when building Music Emotion Recognition models. We aim to provide some trustworthy results based on cross-dataset and transfer learning experiments, with regard to the differences that might exist between cultures and argue about the validity of our assumption.

1.3 Structure of the Report

This thesis report has five chapters followed by the list of figures, bibliography and appendices. The following chapter provides a summary of the Music Emotion Recognition journey so far and emphasizes current state-of-the-art methods. Chapter 3 describes the materials and methods we propose, along with an in-detail presentation of our datasets, baseline experiments and proposed deep learning architecture. The fourth chapter presents our results and comparisons between algorithms and cultures, while in the final chapter we discuss the interpretation of our results, our final conclusions and further work that could to be done to extend our work in the field.

Chapter 2

State of the Art

Research regarding emotions in music started with psychological studies in terms of what emotions are, how and why we experience and perceive them. In order to identify and differentiate between emotions, several types of definitions and divisions were proposed, therefore automatic emotion recognition also started to expand on slightly different paths. Moreover, analysis techniques, particularly machine learning, advanced under various architectures and approaches. Some of these are related to the structure of the input-output formats, that is the data attributes that go into the algorithm and their corresponding tags that go out as numerical values. Other variations are concerned with technological progress in terms of deep neural network design and fine-tuning. In this chapter we will provide an outline of all these existent paths and attempts to MER, with a particular emphasis on the most effective methods so far, i.e. state-of-the-art models.

2.1 Music and Emotions

The concept of 'emotion' is quite complex since our feelings and perception are very relative and subjective, therefore there are several variations on its scientific definition. For this study, we are going to assume the following definition that is quite general but appropriate for Music Emotion Recognition:

Emotions are relatively brief, intense and rapidly changing responses to potentially important events (subjective challenges or opportunities) in the external or internal environment, usually of a social nature, which involve a number of sub-components (cognitive changes, feelings, physiology, expressive behaviour, and action tendency) which are more or less "synchronized" during an emotion episode [3].

In our investigations of emotions, one important distinction that can be made is between the emotions that are perceived and those that are felt. We can identify certain emotions as being transmitted in a song, the general atmosphere they create - the perceived ones, but we do not need to fully experience them ourselves [4]. This is generally based on what the composers and artists meant to express and transmit, irrespective of the listener. Research in Music Emotion Recognition generally targets these ones because agreement in this direction can be reached to some extent. On the other side, what we personally experience - the felt emotions, will depend significantly on previous experiences, memories and connections within the brain, therefore these are much more difficult to generalize. However, even though they inherit this difficulty to be tackled, there is also research on MER about that.

Current studies on emotion consider one of the two main taxonomies: categorical and dimensional [5]. The categorical one divides emotions in various numbers of clusters comprising similar emotion words (e.g. happy, sad, angry, fear, relaxed, surprise, etc.). However, some results show that this might not be the most optimal approach as clusters tend to overlap [6]. For example, in the community of Music Information Retrieval Evaluation eXchange (MIREX), where five mood clusters are used, 'fun' and 'cheerful' in Cluster 2 could easily be confused for 'humorous' and 'silly' in Cluster 4. In addition, due to emotional granularity across languages and their concepts, meaning that not all emotion words in a language can be accurately translated into another, it would be a very sensitive and exhausting matter to create valid divisions within all the different emotion words that exist in the world [7].

On the other side, in the dimensional approach, emotions are identified on the basis of their location in a space with a small number of emotionally-relevant dimensions. It has the advantage of being able to detect more details and shades of emotions,

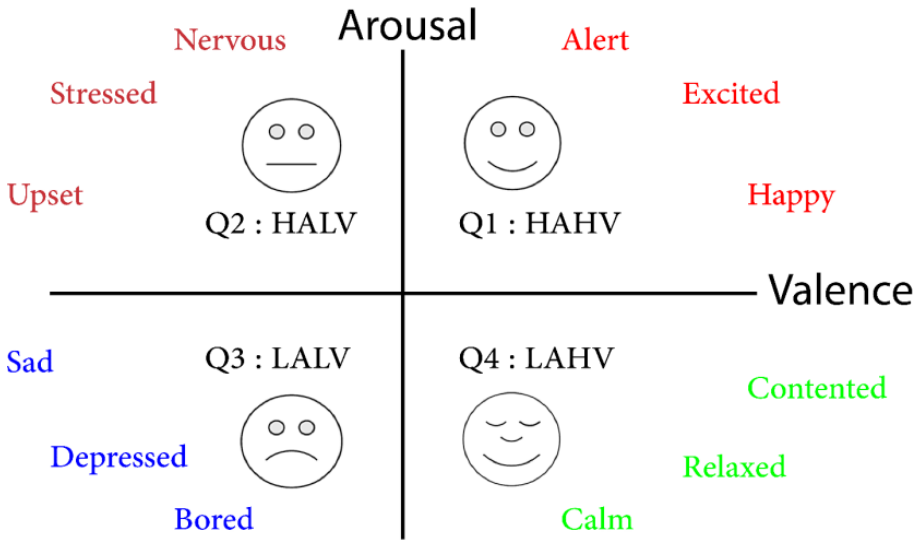

Figure 1: Valence-Arousal plane with some basic emotions on it [11]

while maintaining a universal framework. The most famous and accepted model is Russell's Valence and Arousal (VA) plane [8], as depicted in Figure 1, where valence refers to how positive or negative the emotion is and arousal to how intense it is in terms of energy level. The plane classifies emotion in one of the four quadrants which would suggest, in part, some sort of emotions as following:

1. Q1: high arousal and positive valence – happiness, enthusiasm;
2. Q2: high arousal and negative valence – anger, stress;
3. Q3: low arousal and negative valence – sadness, bitterness;
4. Q4: low arousal and positive valence – calm, serenity.

Psychologists state that for the VA model, while the arousal dimension is related to tempo (fast or slow), pitch (high or low), loudness level (high or low) and timbre (bright or soft), the valence dimension is more related to mode (major or minor) and harmony (consonant or dissonant) [9]. Some findings also show that the arousal dimension is more universal, while the valence characteristic tends to be more culture specific [10], which makes it easier to identify the right level of arousal than the perceived valence of a song.

2.2 Music Emotion Recognition

After identifying the characteristics of the music-emotions relationship, it only comes natural to think about the logic and reasoning behind why certain emotions are evoked by certain songs. Extracting emotional information from music makes sense because we assume that there are certain aspects of it (or combinations of more aspects) that contain particularly relevant information. These can be related to pitch, loudness, rhythm, etc. or even the shape of the raw waveform and thus, we aim to be able to detect emotion based on such characteristics.

In order to better define some rules and identify patterns in the music-emotion associations, people started using Machine Learning under the assumption that computers are able to analyze larger corpora of data, along with longer lists of characteristics. The development of a MER algorithm usually has several steps, described as following:

- Data collection and annotation - Annotations can either be discrete emotion classes or continuous values of dimensional axes and for these, we discriminate two machine learning approaches: classification and regression. In the case of classification, the model predicts discrete classes represented by integers and it can be single-label, where classes are mutually exclusive, or multi-label, where multiple tags are available for one song. On the other hand, regression predicts continuous values of different emotions or emotionally relevant axes, either as floating point numerical-values or continuous probability distributions.

- Feature extraction and selection - Very common for music analysis is the feature extraction from spectrograms, but recently, raw waveforms have also been used [12][2]. In addition, the considered features can be low-level related to audio signal processing or high-level related to human perception.

- Model training - The model can be a traditional machine learning algorithm, that was perhaps optimized for the task, or it can be a specifically designed

deep learning architecture. The deep learning system can have some intermediary pre-processing steps or it can be an end-to-end system.

- Model evaluation - Evaluation techniques are different for classification and regression. The former is most commonly rated in terms of classification report (accuracy, f-score, recall, precision) and confusion matrix, while the latter in terms of root mean squared error (RMSE).

2.2.1 Datasets

In order to train machine learning algorithms to accurately determine emotions from music, it is required to feed it with a large corpora of human-annotated music. Because of the fact that the effectiveness of machine learning algorithms tends to improve with the size of the train set, the creation of appropriately-sized datasets, that also maintain a good level of integrity, is a challenging task. Building appropriate datasets and output classes remains a challenging task since emotions are so abstract and personal, therefore the problem is still open.

In addition, another challenge for the music analysis industry are the rights associated with songs, especially popular songs. There are several open source sound and music collections, but these are usually limited. Some researchers use their own private music collections and only distribute several extracted features or results, but the impossibility to fully reproduce their experiments makes it harder to advance the MER field.

During the past years there were indeed several datasets that were proposed and open to everyone. Panda et al. [13] created a 900 30-second clips dataset, annotated in terms of Russell's emotion quadrants. The dataset was called 4Q-Emotion and mostly contains popularly consumed English music. It was collected through the AllMusic API by selecting emotion tags from the original AllMusic Tags and intersecting them with the Warriner's adjectives list [14]. Finally, a manual blind validation is conducted with subjects in order to validate the annotations.

Aljanaki et al. proposed the MediaEval Database for Emotional Analysis in Music

(DEAM) [15], in the context of the 'Emotion in Music' task at MediaEval Multimedia Evaluation Campaign. It contains royalty-free music from several sources: freemusicarchive.org (FMA), jamendo.com and the medleyDB dataset, that was crowdsourced with manual annotations through the Amazon Mechanical Turk engine. It consists of 1802 excerpts and full songs annotated with valence and arousal values both continuously (per-second) and over the whole song, comprising a variety of Western music genres. The authors also tried to leverage the issue of disagreement between annotators by building an annotation experiment where the mood of the subjects at the time of listening along with the time of the day were taken into consideration [16]. They concluded that the time of the day and workers' reported "energetic" mood had a small but significant effect on the ratings.

Hu et al. created the Moods MIREX dataset [6] for the first Audio Mood Classification (AMC) evaluation as part of the Music Information Retrieval Evaluation eXchange (MIREX) challenge. It is made of a selection of the libraries of Associated Production Music (APM) and the pieces were rated by 3 persons with only a subset of agreement of 2 out of 3 being kept. Music is classified in 5 clusters defined as following: Cluster 1 (passionate, rousing, confident, boisterous, rowdy), Cluster 2 (rollicking, cheerful, fun, sweet, amiable/good natured), Cluster 3 (literate, poignant, wistful, bittersweet, autumnal, brooding), Cluster 4 (humorous, silly, campy, quirky, whimsical, witty, wry), Cluster 5 (aggressive, fiery, tense/anxious, intense, volatile, visceral). However, the defined taxonomy lacks support from music psychology and clusters show several overlaps.

Yang and Chen [17] proposed a ranking-based emotion annotation method requiring subjects to annotate pairs of songs in relation to each other. This was aimed to ease the pressure on the subjects and enhance the accuracy of annotated data. However, noise still exists, therefore methods of finding noise in data should be further investigated.

Among other not so popular datasets, there is the Soundtracks dataset [18], comprising film soundtrack music and designed to overcome familiarity discrepancies since they contain not that well known examples. NTWICM: Now That's What I Call

Music is another dataset [19] that represents very well most music styles which are popular today ranging from Pop and Rock music over Rap, R&B to electronic dance music as Techno or House. For this, 4 raters gave static annotations for about 2500 complete songs for arousal and valence in a discrete range from negative 2 to positive 2. Finally, Emotify [20] is another music emotion dataset focused on induced rather than perceived emotions, but this is beyond our scopes.

2.2.2 Features

We have seen so far that various tags and types of datasets exist for MER, but there are also various types of features. This is because there are many possibilities to go from an audio segment to a specific emotion word. This path can depend on the low-level sound descriptors or on the higher level human perceivable characteristics like harmony or rhythm, that might also depend on longer sequences of sounds. The study of features began even before the development of computational techniques such as machine learning, when people started to look by themselves after correlations between various aspects of music and the perceived emotions. Studies show that combinations of features like major mode, small tempo variability, wide pitch range and staccato articulation are related to happiness and combinations like large tempo variability, minor mode, dissonance and large sound level variability trigger fear [21]. A distinction that was also made is between features related to composers, like mode or melody progression, and features related to performer, like tempo or timbre. Juslim argues that, while performance features are more related to the non-verbal aspects of emotional speech, composer features are likely to reflect characteristics of music as an art that follows its own intrinsic rules and that varies from one culture to another.

As research and technology evolved, people started to approach the problem systematically and computationally, and the correlations between various feature sets were made by machine learning algorithms. Yang, Dong and Li [9] review a set of MER methods and papers from the perspective of three aspects of music: features only, ground-truth data only and their combination, which is the most common and

used supervised methods. By ground-truth data we refer to the tags and types of tags that are directly associated to music by one or more human annotators. Yang et al. affirm that different emotional states are usually associated with different music features and an increase in the feature dimension simply cannot improve the performance effectively, and thus, we need means of selecting the most suitable features, such as the principal component analysis (PCA) [22] or factor analysis [23]. They also show that the single-label task, as well as valence prediction in general, give better results with source separation considerations, while the multi-label task outperforms by considering correlations between labels. The main challenge for multi-label classification is the high dimensionality of label space, especially when few train samples are available.

There were a lot of investigations in what concerns the specific features to consider for these algorithms. While Yang, Dong and Li [9] present the purely sound-based and statistical low-level features: Mel-frequency Cepstrum Coefficients (MFCCs), Octave-based Spectral Contrast (OSC), Statistical Spectrum Descriptors (SSDs), Chromagrams and Daubechies Wavelet Coefficient Histograms (DWCH) as the most commonly used descriptive features of MER, Soleymani and Aljanaki [24] reason for more perceptual and meaningful-to-human features: Melodiousness, Articulation, Rhythmic Stability, Rhythmic Complexity, Dissonance, Tonal Stability and Modality ('Minorness').

A current benchmark in terms of feature sets for MER is the Interspeech 2013 Compare Feature Set [25], tested with the DEAM dataset. This feature set was created with the Munich open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolkit [26], a modular and flexible feature extractor for signal processing and machine learning applications. The set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness.

In terms of higher level features, there can be more usual music features like pitch, timbre or rhythm, but also more specific and nuance related. Panda et al. [13]

investigated higher level features related to musical form, texture and expressive techniques within the 4Q-Emotion dataset. Their SVM classification gives better results with these novel features, especially after applying feature selection. Moreover, while better features are needed to discriminate the low arousal quadrants, several characteristics were observed as discriminative for each quadrant: tone colour, rhythm, musical layers and texture for Q1, roughness, rolloff (i.e. amount of high frequencies), MFCCs and vibrato for Q2, musical layers, dissonance, inharmonicity and tremolos for Q3, skewness, spectral flatness, dissonance and vibrato for Q4. A general conclusion that was obtained is that the majority of available features are related with tone colour or timbre (63.7%), which could look a bit surprising given the typical association of major/minor modes with positive-negative moods in Western listeners.

Another study approached the multi-label classification task by considering only timbre related features, leaving rhythmic patterns aside [27]. After applying two learning algorithms, Random k-Labelsets (RAkEL), based on random projections of the label space, and Multi-Label k-Nearest Neighbours (MLkNN), an adaptation of the usual kNN, similar results are obtained with an accuracy of about 80%. In addition, the standard deviation of MFCC-12 is concluded to be very relevant and influential in the listeners emotions.

Researchers also experimented with many types of approaches in terms of machine learning algorithms for MER and as technology progresses there are certainly new ones that are yet to come. It is interesting to note how MER models evolved from traditional techniques like Support Vector Machines (SVMs), which are very popular in this domain, to deep learning techniques like Convolutional Neural Networks (CNNs), that proved to be especially good for speech and other sound related tasks.

One paper that investigates the performances of several traditional machine learning approaches based on previously extracted features shows that the most efficient algorithm was the Gaussian Naive Bayes with a 40.33% accuracy [28]. However, this result discourages further use of such methods, especially with very general datasets and for very general purposes like the MIREX challenge in this case.

So far we have seen that, experiments were conducted with several label types, then with several feature spaces along with various learning algorithms. While many configurations were fed to the systems, results were not satisfying enough and deep learning methods started to be employed. In this way, the hand-crafted features described so far are being replaced with 2D representations of sounds.

2.2.3 Deep Learning

In the Deep Learning field, Deep Neural Networks (DNNs), composed of several layers of neurons, are employed in order to solve complex problems. Each neuron or node in the network represents one aspect of the whole and together they provide a full representation of the input. Each node or hidden layer is given a weight that represents the strength of its relationship with the output and as the model develops the weights are adjusted accordingly [29]. A hidden layer is located between the input and the output of a neural network, performs nonlinear transformations of the entered inputs and directs them through an activation function to the output. The activation function defines the output of a node based on its inputs. The weights of the neurons are updated through gradient descent, an optimization algorithm used to minimize a loss function by iteratively moving in the direction of the steepest descent as defined by the negative of the gradient.

The biggest advantage of Deep Learning algorithms is that they try to learn high-level features from data in an incremental manner. This eliminates the need of domain expertise and hardcore feature extraction, which could free researchers from the eternal quests of finding the best feature set for the problem. End-to-end learning is a hot topic in this field, in which features are an outcome of the whole process, instead of being computed on purpose and under musical or signal-processing insights. This means that the input for an end-to-end music emotion recognition network is the raw audio waveform. The end-to-end learning process must collect all the parameters jointly from this, while a usual deep learning process can collect the parameters either jointly or step by step.

In the case of sound related tasks, Convolutional Neural Networks (CNNs), spe-

cific to analyzing visual imagery, are employed as learning is based on the image representation of sound, which in turn, is represented as a 2-dimensional array of numbers, known as pixels. In comparison to normal DNNs, the CNN layers are organised in 3 dimensions: width, height and depth and the neurons in one layer do not connect to all the neurons in the next layer, only to a small region of it. The final output is reduced to a single vector of probability scores, organized along the depth dimension. CNNs have two component parts: feature extractor and classifier. Inside the feature extractor, where the hidden layers are, the main convolution operation is performed, that is the mathematical combination of two functions to produce a third function, merging two sets of information. Convolution is performed by sliding a filter or kernel over the input data to produce a feature map. Numerous convolutions with different filters are performed, resulting in different feature maps that make the final output of the convolution layer. After a convolution layer, it is common to add a pooling layer to continuously reduce the dimensionality and reduce the number of parameters and computation in the network. The classification part consists of fully connected layers, meaning that the neurons have full connections to all the activations in the previous layer, usual to normal DNNs as well.

There are two types of input formats for CNNs, that define two types of networks: spectrogram and waveform-based. The former performs 2D convolutions (across time and frequency), while the latter performs 1D convolutions (across time). Another difference is that waveform-based models do not discard the phase and use the raw signal as it is. If this is an advantage or not, this is still to be determined for different tasks.

Pons et al. [12] provide an effective comparison between two types of end-to-end models for the task of automatic music tagging: an assumption-free CNN – using waveforms as input with very small convolutional filters – and another CNN that relies on domain knowledge – using log-mel spectrograms and designed to learn timbral and temporal features. Although the task is not exactly the same with MER, their lessons should be taken into consideration in the sense that no model assumptions are required when operating with large amounts of data and the wave-

form input liberates some limitations, allowing more information to be fed to the network. These conclusions were raised after training on more datasets of different sizes, where the biggest one is a private dataset made of 1.2 million songs. However, results show that for the biggest dataset publicly available – the Million Song Dataset [30] actually containing only 240 000 songs, the spectrogram model outperforms the waveform model suggesting that the lack of larger public datasets could be limiting the outcomes of deep learning research for music auto tagging and perhaps for other MIR tasks as well.

A simple attempt to MER with a Convolutional Neural Network (CNN) that only considers the spectrogram input and shows promising results can be found in the work of X. Liu at al. [31]. The dataset used here contains multilabel annotations. Another CNN that was built in a similar manner, but in the VA plane can be found in the work of T. Liu at al. [32]. Here, the authors show that the CNN model learns considerably better than the SVM model.

As the use of spectrograms with domain knowledge proved to be more accessible and efficient for the given data, many researchers experimented with various music features and assumptions when building their deep learning models. Schmidt and Kim [33] used deep belief networks (DBN) to learn the sparse features of music and Chen et al. [34] employed the deep Gaussian process (deep GP) to achieve the classification of nine music emotions in the VA emotion space. In comparison to the classification results of SVM and standard Gaussian process, the deep GP results were slightly better, but there is still room for improvement [9].

Chowdhury et al. [35] propose a VGG-style deep neural network – a uniform CNN architecture with many small filters, that learns to predict emotional characteristics of a musical piece together with (and based on) human-interpretable, mid-level perceptual features. They compared a black box audio-to-emotion architecture to an audio-to-mid-level-features-to-emotion and further proposed a single network with two outputs to predict both the emotion and the features used. Despite a small loss in performance when going through the mid-level features, they argue that it is justified by the gain in explainability of the predictions, which is desirable in MIR

applications.

In an attempt to provide more complete musical explanations, the same authors propose an extension of the model in which listenable explanations for the mid-level features are created, based on local interpretable model-agnostic explanations (LIME) [36]. By listenable explanations, they mean to gain some human interpretable intuition in regard to the decisions of the model. To derive an explanation for an instance, LIME trains a simpler, interpretable model on a set of perturbed samples by using the original model prediction as the label. Explanations are read through patterns in the output spectrograms.

Another approach showed some good MER results by considering the chroma spectrogram as input [37]. Deep visual features are extracted from different layers of the the pre-trained deep neural networks, AlexNet and VGG-16 and used to train and test SVM and Softmax classifiers. After performing data augmentation on their own four-emotion dataset (happy, sad, angry, relaxed) with Turkish music, the best classifier success was obtained from the Fc7 layer of the VGG-16 with the SVM classifier with 89.2% accuracy. The fact that most of their tests were about changing the train-test data splits raises some questions, but the approach of using pre-trained networks remains promising.

From another perspective, Orjesek et al. [38] propose an architecture that takes as input the raw signal and is created by stacking a one-dimensional CNN (1D-CNN), a Time-Distributed Fully-Connected (TD-FC) layer and a Bidirectional Gated Recurrent Unit (BiGRU). The last layer is the Maxout Fully-Connected (MFC) layer and provides mapping into the continuous two-dimensional VA scale. This method outperforms other systems trained on the DEAM dataset and emphasized in the MediaEval benchmark initiative [15]. It also gained slightly better performance while using significantly less parameters than the state-of-the-art system recently published by Malik et al. [39], where the same architecture was trained with extracted features as input. A significant improvement over this last one was achieved for valence prediction, demonstrating that some commonly used pre-determined features do not efficiently represent valence.

Finally, a very recent and perhaps one of the most complex approaches was developed by Dong et al. [40] through a new bidirectional convolutional recurrent sparse network (BCRSN), evaluated again on the DEAM dataset, but also on MoodSwings Turk (MTurk) [41], that was created in a similar manner to DEAM [41]. Similar to the work of Orjesek et al. [38], sequential information is presented as critical for continuous music emotion recognition, but instead of directly stacking the CNN and the Recurrent Neural Network (RNN), a CNN is used to replace the connection between the input and hidden layers of a RNN. In addition, to reduce the high computational complexity caused by the numerical-type ground truth, Dong et al. propose a weighted hybrid binary representation method that converts the regression prediction process into a weighted combination of multiple binary classification problems. Results imply that the model outperforms current state-of-the-art methods and that the learned features are robust to genre, timbre and noise variation and sensitive to the more precise human-perceived emotion.

Therefore, the era of deep learning started to raise the standards for Music Emotion Recognition. While several systems continued to explore feature extraction configurations, some of them for the sake of obtaining classification explanations, the general trend is towards letting a deep network discover relevant information by itself. We have also seen that the design and fine tuning of a specific network could be just as promising as transfer learning, with previously trained networks on other tasks.

2.3 Language and Culture

Since music and speech are relatable to some extent and music usually has lyrics as well, it is worthy to take into consideration some research made on speech. Things like intonation curves in speech might resemble melodic contours in music, stress patterns in speech could be reflected in different rhythm meter preferences and the noisiness of certain languages could also be connected with the noisiness of music. Bowling et al. [42] show how the perceived consonance of chords is predicted by their relative similarity to voiced speech sounds. From this we could further infer

that slightly different distributions of speech sounds in different languages could influence the way listeners perceive music. This means that if consonance perception is motivated by a given language, this could also be the case for emotion perception. In their paper about speech, music and sounds, Weninger et al. [43] also discover a high degree of cross-domain consistency in encoding both dimensions of affect, valence and arousal, and argue that this may be attributable to the co-evolution of speech and music from multi-modal affect bursts, including the integration of nature sounds for expressive effects.

Gomez-Cañón et al. [44] attempted to discover whether language and lyrics comprehension are relevant to emotion perception. Results show that basic emotions will have higher universal agreement, while complex ones will show the opposite. Moreover, lyrics comprehension (LC) improves agreement for emotions in quadrants Q1(A+V+) and Q3(A-V-) and decreases it for quadrants Q2(A+V-) and Q4(A-V+), suggesting that better understanding could lead to similar emotions in some cases and finer judging criteria in others.

Such studies would further reason for building transfer learning models from speech emotion recognition to music emotion recognition. Transfer learning means that the model trained on a dataset is then transferred to a target task, leading the knowledge learned in a large dataset to be transferred into a possibly small dataset. This is also why Coutinho and Schuller [45] investigated this issue and found that there is a substantial overlap between the acoustic codes for emotional expression in music and speech. Their results show an excellent cross-domain generalisation of time-continuous estimations of emotional Arousal and Valence in music and speech. Also, knowledge transfer from speech to music was more successful than in the opposite direction and in cases of small datasets like it is usually the case with music emotions, additional speech data could enhance the models.

Recent studies have tried to show that culture might play a role in the way we perceive music and therefore, they built models trained on music of different cultures and languages. Hu and Yang [10] provided an experiment with two datasets of music in English, annotated by Chinese and Western people respectively and one dataset

with both Chinese music and annotators. They found that within-dataset predictions out-performed cross-dataset predictions. While a common cultural background in the datasets is important for predicting the valence dimension, the annotation reliability level seems to be the most important factor for cross-dataset generalizability of models on arousal prediction.

Another cross-cultural study [46] investigated Korean and Western music by building a collection of K-pop songs with mood annotations collected from both Korean and American listeners, based on three different mood models. It was observed that mood judgments and the level of agreement are dependent on the cultural background of the listeners, thus different behaviours were found between the annotations of the two groups. Similarly, Sangnark et al. [47] attempted to study the differences between Western and Thai emotion tagging. Interesting is that for Thai music, they retrieved the valence and energy – representing arousal – values with the Spotify API. The final dataset was small and unbalanced, unfortunately, but they also conclude that different models should be used for different cultures.

By using an unsupervised adversarial domain adaptation method, Chen et al. [48] built a neural network that aims to adapt itself to better predict new datasets, in this case music from another culture. The adapted model does improve results for valence prediction but not for arousal, as arousal is more generalizable across datasets. It also performs better than the method proposed by Hu and Yang [10], i.e., support vector regressor, for both valence and arousal prediction. It is to note that the approach only accounts for cultural differences in music features and not in emotion perception.

Delbouys et al. [49] provide a multimodal music mood prediction based on audio signal and lyrics, based on 18 000 tracks with associated continuous arousal and valence. They compare a late fusion of unimodal predictions, identified by two separate CNNs based on audio-mel spectrograms and embedded lyrics respectively, to a bimodal deep learning model that implements a mid-level fusion between the two. The later model proves to be more successful, particularly when it comes to predicting valence, by unveiling and using mid-level correlations between audio and

lyrics.

A similar research was conducted by Zhou et al. [50] by applying a three-stage strategy made of unsupervised feature learning, regression model training and testing, on a 3000-song-dataset of Chinese music, annotated with PAD (Pleasure-Arousal-Dominance) values. Results with unimodal learning show again that lyrics data contributes more to pleasure while audio data to arousal and dominance because of its energy-related information. Multimodal learning with the Bimodal Deep Auto Encoder (BDAE) also proves the effectiveness of learning shared representations. Finally, unimodal enhancement worked only for lyrics emotion prediction after performing feature extraction from both domains.

In the context of lyrics and speech, it is worth mentioning some existent studies for speech emotion recognition. Umamaheswari and Akila [51] propose a hybrid of Pattern Recognition Neural Network (PRNN) and K-Nearest Neighbour (KNN) classifier, along with a cascaded system of Mel Frequency Cepstral Coefficient (MFCC) and Grey Level Co-occurrence Matrix (GLCM) for feature extraction, in order to classify speech samples into neutral, anger, happiness, sadness, surprise and fear.

On the other hand, Trigeorgis et al. [52] perform an end-to-end spontaneous emotion prediction on raw input signal from a French speech dataset. The architecture combines CNNs with Long Short-Term Memory (LSTM) networks and seems to outperform state-of-the-art approaches based on designed features. By further studying the activations of different cells in the recurrent layers, they found some interpretable cells, which are highly correlated with prosodic and acoustic features that were assumed to convey affective information in speech, such as the loudness and the fundamental frequency.

All of the above things considered, it seems that research in the MER field should indeed take into account cultural and language considerations, at least to some extent, for instance when determining valence as this has been found to be more context-based. We have seen that music and speech indeed share similar emotional content and also that lyrics considerations usually improve MER systems. The

general issue with these multi-cultural investigations is the lack of data from other cultures than Western, but there is also the challenge of finding the best way to include these considerations in the learning process. We have seen that in such contexts when hand-crafting features is challenging, deep learning comes as a useful and efficient solution when analysing both music and speech. A summary of most of the approaches presented in this chapter can be seen in Figure 2.

2.4 Our Proposal

As we have seen, differences in terms of input format are being studied, as well as in terms of design for deep neural networks. Most studies use spectrograms and other pre-processed features as input to emotion classifiers, but more recent studies started to experiment with raw waveforms as they could be more representative for human related concepts. In addition, it was shown that deep learning models can prove to be more efficient at directly discovering appropriate features than humans who need to hand-craft them firstly.

An interesting architecture that takes the raw audio waveform as input and was successful on speech related tasks is called SincNet [2]. This novel CNN that encourages the first convolutional layer to discover meaningful filters, is based on parameterized sinc functions, which implement band-pass filters in the frequency domain. SincNet could be a state-of-the-art model for speaker recognition and verification, but it also seems a promising candidate for other sound related tasks as it offers a very compact and efficient way to derive customized filters specifically tuned for the desired application.

For instance, Zeng et al. [53] have recently managed to improve and propose a SincNet-based classifier, called SincNet-R, which consists of three convolutional layers and three deep neural network (DNN) layers in order to classify emotional electroencephalography (EEG) signals. There were only three output classes: Positive, Negative, Natural. Although EEGs are out of our scopes here because they are correlated with felt or induced emotion, their results show that the SincNet-R model

Approach	Technical aspects	Reference
4-quadrants classification MER	Expressive features, SVM	[12]
Multi-label classification	Timbre related features, RAkEL vs. MLkNN	[26]
MIREX clusters MER	Inception v3, transfer learning on mid-level perceptual features	[23]
Music tagging	End-to-end CNN vs. CNN with log-mel spectrogram	[11]
Multilabel MER	CNN, spectrogram input	[30]
V-A MER	CNN	[31]
MER	DBN, sparse features	[32]
Classification of 9 emotions in VA plane	Deep GP	[33]
Numeric emotion ratings along 8 dimensions	VGG, mid-level perceptual features, LIME	[34], [35]
MER, Turkish (4 classes)	Chroma spectrogram, transfer from Alexnet vs. VGG-16 + SVM vs. Softmax	[36]
Continuous V-A MER, DEAM	Raw signal, 1D-CNN + TD-FC + BiGRU + MFC	[37]
Continuous V-A MER, DEAM	Extracted features, 1D-CNN + TD-FC + BiGRU + MFC	[38]
Continuous V-A MER, DEAM	BCRSN (CNN + RNN), weighted hybrid binary representation	[39]
Continuous V-A, speech & music	LSTM-RNN, transfer learning with denoising auto-encoders	[44]
Cross-cultural MER, Chinese vs. Western	SVM, feature extraction	[9]
Cross-cultural Thai vs. Western, valence and energy	Multiple linear regression, KNN, feature extraction	[46]
Culture adaptive emotion recognition	Unsupervised adversarial domain adaptation	[47]
Mood prediction with lyrics, continuous V-A	Audio & lyrics, late fusion unimodal – 2 x CNNs, bimodal DL with mid-level fusion	[48]
Emotion recognition with lyrics, pleasure-arousal-dominance, Chinese	Unsupervised feature learning, audio & lyrics, unimodal, multimodal – BDAE, unimodal enhancement	[49]
Speech emotion (6 classes)	PRNN + KNN, feature extraction	[50]
Speech emotion, French, spontaneous	CNN + LSTM	[51]
Speaker recognition	End-to-end, SincNet	[1]
EEG-classification (positive, negative, neutral), induced emotion	End-to-end, SincNet-R	[52]

Figure 2: Summary of the most recent and relevant studies related to our work.

has higher classification accuracy and better algorithm robustness than other deep learning methods, which reasons for further investigations of the SincNet architecture.

In addition, since lyrics and culture considerations seem an appealing direction for the study of music, we propose a new way of extracting emotions from music with SincNet, trained on music from different cultures. We will not consider the exact lyric words, but the architecture will indirectly consider the speech sounds and timbre since it was built for speaker recognition. We plan to compare this approach to a baseline model with pre-extracted features and then, we will perform a set of cross-dataset and transfer learning experiments in order to observe how the architecture behaves under different cultural considerations.

Furthermore, we will observe how effective SincNet is in comparison to other learning methods, especially since the end-to-end learning has not been approached with MER so far. The spectrogram input could be a limitation since it is a pre-processing step for the music excerpt and that is why we will experiment here with the raw waveform input, in an attempt to allow the model to discover more meaningful information. The original SincNet architecture will be described in detail in the next chapter, along with its adaptation to our task.

Our overall goal in this culturally motivated research is building MER models that are able to adapt as much as possible to the human being and how it perceives emotions, while also maintaining some inter-human generalizability. Our aim is to improve on the current state-of-the-art approaches by proposing a new methodology that, not only uses a new CNN model, but also takes into consideration language and culture.

Chapter 3

Materials & Methods

In this chapter, we introduce the three datasets that we investigated and the various experiments that were performed on them. The music was mapped in all the cases to one of the four quadrants of the Valence-Arousal plane, which means that our approach towards Music Emotion Recognition used a mixture of dimensional and categorical taxonomy. We chose this approach because the Valence-Arousal plane has a significant psychological support and we used it with discrete classes because it is easier to deal with, at least for the beginning of this novel approach. In this way, music is classified in one of the four following classes: high arousal and positive valence, corresponding to emotions like happiness and excitement (first quadrant - Q1), high arousal and negative valence suggesting anger or fear (second quadrant - Q2), low arousal and negative valence as in sadness and nostalgia (third quadrant - Q3) and finally low arousal and positive valence representing relaxation and peace (fourth quadrant - Q4).

At first, we conducted experiments with standard Machine Learning algorithms and two different feature sets in order to determine a reasonable baseline model, that uses the exact same data as the proposed model. Then the SincNet end-to-end learning architecture [2] was employed in an attempt to explore a language oriented model for music emotion recognition. In the last subsection, we will also present the cross-dataset and transfer learning set-ups that we investigated with the two selected

models. The code for all our experiments is open source and can be accessed on GitHub as SincNet-MER[1].

3.1 Datasets

For the purpose of this study, we used three datasets with Western (English), Chinese (Mandarin) and respectively Turkish music, because we wanted to investigate the level of similarity for emotion perception in various cultures and also how context-based considerations impact a MER system. The Western set, mostly based on popularly consumed English music is called 4Q-Emotion and was derived through the AllMusic API [54] [13]. The retrieved mood metadata, i.e., the AllMusic tags, are intersected with the Warriner's list [14] containing many English words with VA ratings. Only songs that clearly belong to a single quadrant were kept, based on the distribution of tags. Because several clips were observed to be inadequate, a manual blind validation was also employed, where subjects were given sets of randomly distributed clips to be rated in the four quadrants of Russell's plane. The final dataset contains 30-second excerpts from 900 different songs, equally split between the four quadrants of the Valence-Arousal plane (225 excerpts for each category). About 724 songs have actual lyrics in English.

The Chinese dataset, CH-818 [10], is made of 818 30-second excerpts from Chinese pop songs, released in Taiwan, Hong Kong and Mainland China. The set was initially annotated with continuous real values between [-10,10] in the Valence-Arousal plane which were then mapped to single quadrant tags. The authors selected the 30-second segment with the most emotionally relevant content within the whole song, such that it has the highest score in terms of the sum between the squared two values (arousal, valence). Each clip was annotated by three music experts who were born and raised in Mainland China and thus had a Chinese cultural background. Despite the small number of annotators, the authors argue that annotions were considerably consistent. This dataset was very unbalanced, therefore after balancing, we got only 89 songs of each category (356 in total).

[1] `https://github.com/ana-pandrea/SincNet-MER`

Dataset	Language	Initial annotation	Original size	After balancing	Segments	Train	Test
4Q-EMOTION	English	4 quadrants	900 (30s)	900	3600	2880	720
CH-818	Mandarin	2 dimensions	818 (30s)	356	1424	1140	284
TR-MUSIC	Turkish	4 categories	400 (~30s)	400	1600	1280	320

Figure 3: Summary of the 3 datasets with music of different cultures, that were used in our study.

The Turkish TR-MUSIC dataset [37] is made of 400 song excerpts from instrumental and with lyrics music of different genres of Turkish music. It was originally split by the categorical taxonomy into four classes: happy, angry, sad, relaxed, which we mapped to the four quadrants: Q1, Q2, Q3 and respectively Q4. In the annotation experiment there were 13 participants who were asked to label a random 30-second excerpt from each song. The chosen tag was voted by the majority of annotators. The dataset is balanced, therefore we got 100 segments for each quadrant.

For all three datasets, we segmented the 30-second clips into four approximately 8-second parts as this is a reasonable excerpt length for emotional analysis and it also allows for more training data than in the original configuration. It is known that the shorter the segment, the more homogeneous the emotion will be, thus making the evaluation results more consistent [9]. Despite the popular tendency of using 25-30 second excerpts, results of a study investigating the ideal length of a segment in MER show that this is optimal when it is between 8 and 16 seconds, thus we believe the trade-off in this experiment could be feasible. In this process, we had to ensure that segments belonging to the same song stay in only one of the train-test data splits, therefore the split was performed before segmentation. Test data represents 20% of the whole dataset for all datasets and all experiments. The information about these datasets is summarized in Figure 3.

3.2 Baseline Model

In order to determine a reliable baseline model to be used for comparison for our end-to-end approach, we implemented a few different algorithms and feature extractors. These were compared to each other and we identified the set-up that was generally the best performing with our task and data. These comparisons were run using Google Colab, along with the appropriate Python libraries.

3.2.1 Comparison of Features

In order to choose a strong baseline model to reflect the performance of our end-to-end network proposal, we compared two different feature sets, extracted with two different tools. The first set is computed with Essentia[2] [55] from which we only selected the low-level scalar acoustic descriptors and the second is computed with the the openSMILE extraction tool[3] [26] and provides low-level descriptors with more emotionally-relevant and task-specific content.

The Essentia music extractor is an open-source configurable command-line feature extractor that can compute a large set of spectral, time-domain, rhythm, tonal and high-level descriptors. The extractor is suited for batch computations on large music collections and was used within AcousticBrainz [56], a project that aims to crowd source acoustic information for all music in the world and to make it available to the public. For simplicity and also for the sake of performing a comparison between similar feature sets, we only kept the low-level descriptors from Essentia, which are represented with a single scalar value. These descriptors are related to loudness, silence rates, spectral energy, mfccs, barkbands, erbbands, dissonance, etc. along with several statistics: mean and standard deviation. These add up to a total of 84 features.

The second feature set is a well-evolved set for automatic recognition of paralinguistic phenomena – the INTERSPEECH 2013 Computational Paralinguistics ChallengE

[2] `https://essentia.upf.edu/streaming_extractor_music.html`
[3] `https://www.audeering.com/opensmile/`

4 Energy Related LLD	Group
Sum of Auditory Spectrum (Loudness)	Prosodic
Sum of RASTA-Style Filtered Auditory Spectrum	Prosodic
RMS Energy, Zero-Crossing Rate	Prosodic
55 Spectral LLD	**Group**
RASTA-Style Auditory Spectrum, Bands 1−26 (0−8 kHz)	Spectral
MFCC 1−14	Cepstral
Spectral Energy 250−650 Hz, 1 k−4 kHz	Spectral
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90	Spectral
Spectral Flux, Centroid, Entropy, Slope, Harmonicity	Spectral
Spectral Psychoacoustic Sharpness	Spectral
Spectral Variance, Skewness, Kurtosis	Spectral
6 Voicing Related LLD	**Group**
F_0 (SHS & Viterbi Smoothing)	Prosodic
Probability of Voicing	Sound Quality
Log. HNR, Jitter (Local, Delta), Shimmer (Local)	Sound Quality

Figure 4: IS13 ComParE acoustic feature set: 65 low-level descriptors (LLD) [25].

feature set (IS13 ComParE) [25]. This was extracted using the Munich open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolkit [26], a modular and exible feature extractor for signal processing and machine learning applications. It includes energy, spectral and voicing related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. Although the set predominantly contains low-level descriptors, just like our configuration for Essentia, it was specifically designed and well-evolved for automatic recognition of emotion from audio, as part of the Interspeech Computational Paralinguistics Challenge taking place every year from 2009. The set contains a total of 260 features, out of which 65 features are means, 65 features are standard deviations and the 130 first derivatives of these. The main 65 low-level descriptors are summarized in Figure 4.

For each of the feature sets, we normalized the feature values between 0 and 1 and balanced the training data. Then we selected the best 50 features, using the SelectKBest algorithm from Scikit-Learn[4] [57], that is based on univariate feature

[4] `https://scikit-learn.org/stable/index.html`

selection and works by selecting the best features based on univariate statistical tests. The function takes as input a scoring function that returns univariate scores, in our case the chi-squared dependence between stochastic variables, that computes chi-squared statistics between each non-negative feature and class. We chose the best 50 features based on experimental results with both fewer and more features. Moreover, we compared the best 10 features that were selected with this method so that we can observe whether different cultures create different distinctions between emotions in music. These will be observed in subsection 4.3.1 and emphasized in Figure 11.

A further preliminary experiment that we conducted involved combining the two feature sets in order to select the best 50 features from both sets. However, many low-level descriptors are similar within the two sets and this approach might not necessarily be efficient. As it will be emphasized in the next chapter, the selected baseline model only uses the IS13 ComParE set.

3.2.2 Comparison of Algorithms

For this part, we considered several classifiers when training individual datasets, in order to select the one that gives the best performance with our task of music emotion recognition. The traditional machine learning algorithms that we implemented were built with the Scikit-Learn library [57], generally with their default configurations. We compared the classification results for each dataset for the following algorithms: K-Nearest Neighbors (KNN) with number of neighbours n = 3 (determined experimentally) and uniform weights, Support Vector Machine (SVC) with linear kernel and regularization paramenter C = 0.025, Support Vector Machine (SVC) with Radial Basis Function, Gaussian Process Classifier with Radial Basis Function, Multi-Layer Perceptron (MLP) with 1 hidden layer, regularization 1 and 1000 maximum iterations, Gaussian Naive Bayes with default variance smoothing of 1e-9 and Random Forest Classifier (RFC) with 100 trees, maximum depth of 15 (determined experimentally) and random state 0.

We ran a classifier comparison function on the set of seven classifiers, with 10-fold

cross-validation on the training set, in order to compute some general statistics for the classifiers. Then we selected the best performing one, tested it on the test set and computed the accuracy, precision, recall, f1-score and support, along with the according confusion matrix. Several differences in terms of the best performing algorithms were found within the different datasets as well as in term of the different feature sets, but we managed to establish a baseline model such that is works reasonably well for all configurations. Results will be further provided in Section 4.1.

We found that the Multi-Layer Perceptron (MLP) is the algorithm that performs well with all configurations of features and datasets. The perceptron is a simple classification algorithm that makes its predictions based on a linear function combining a set of weights with the input vector. A perceptron only has two layers, input and output, while a multi-layer perceptron has at least one hidden layer and it is also a feedforward network, meaning that there are no loops involved (like in the case of RNNs). In our case, the algorithm was called with the L2 regularization parameter 1, maximum number of iterations 1000 and the rest were default settings. The regularization term is added to the loss function that shrinks model parameters to prevent overfitting. The model has one hidden layer with 100 neurons and uses the Rectified Linear Unit (ReLu) activation function. ReLU is a linear function that outputs the input directly if it is positive, otherwise, it outputs zero and it is very popular because it overcomes the vanishing gradient problem (when the parameters become very high, towards infinity or very low, close to 0), allowing models to learn faster and perform better. We believe that the choice of this baseline model is good since it is basically the simplest neural network and it is a comparable starting point for a deep learning architecture. We will present all these results in the following chapter.

3.3 End-to-End Model

The end-to-end learning model that we proposed for this MER task is SincNet [2]. This architecture was successful for tasks like speaker recognition and verification,

but the authors argue that it is promising for other tasks as well, like emotion detection. SincNet is a novel CNN architecture for processing raw audio samples that encourages the first convolutional layer to discover more meaningful filters. The first layer or the feature selector layer was selected to be modified from a normal CNN because of its high dimensional inputs and their relevance to the task. This layer is usually more affected by vanishing gradient problems, where the gradients of the loss function approach zero, making the network hard to train. The full SincNet architecture can be observed in Figure 5.

CNNs became the most popular models for raw speech processing since weight sharing, local filters and pooling help discover robust and invariant representations. However, filters in normal CNNs are noisy and take incongruous multi-band shapes. In contrast, SincNet optimizes this by adding constraints on the shape and convolving the waveform in the time domain, with sinc functions that implement rectangular band-pass filters, instead of usual convolutions. The only two parameters that are learned are the low and high cut-off frequencies. This offers a very compact and efficient way to derive a customized filter bank specifically tuned for the desired application, in this case emotion recognition.

In their paper, Ravanelli and Bengio [2] proved that SincNet converges faster and achieves better end-task performance than standard CNNs, especially with minimal training data and short test sentences. This could be an advantage for our experiments since our datasets are not very big either. Their model was able to learn low level speech representations and capture narrow-band speaker characteristics, like pitch and formants. It also has the advantage of interpretability as its parameters have a clear physical meaning, while also being able to derive customized filter banks tuned for different applications.

Because this model takes as input the raw audio waveform and only considers small fragments at a time, the architecture becomes sensitive to the various phonemes of a language. In this context, we believe that associations between sounds in a particular language and emotions could enhance general MER results. As SincNet was originally designed for speaker recognition, its sensitivity to timbre might also con-

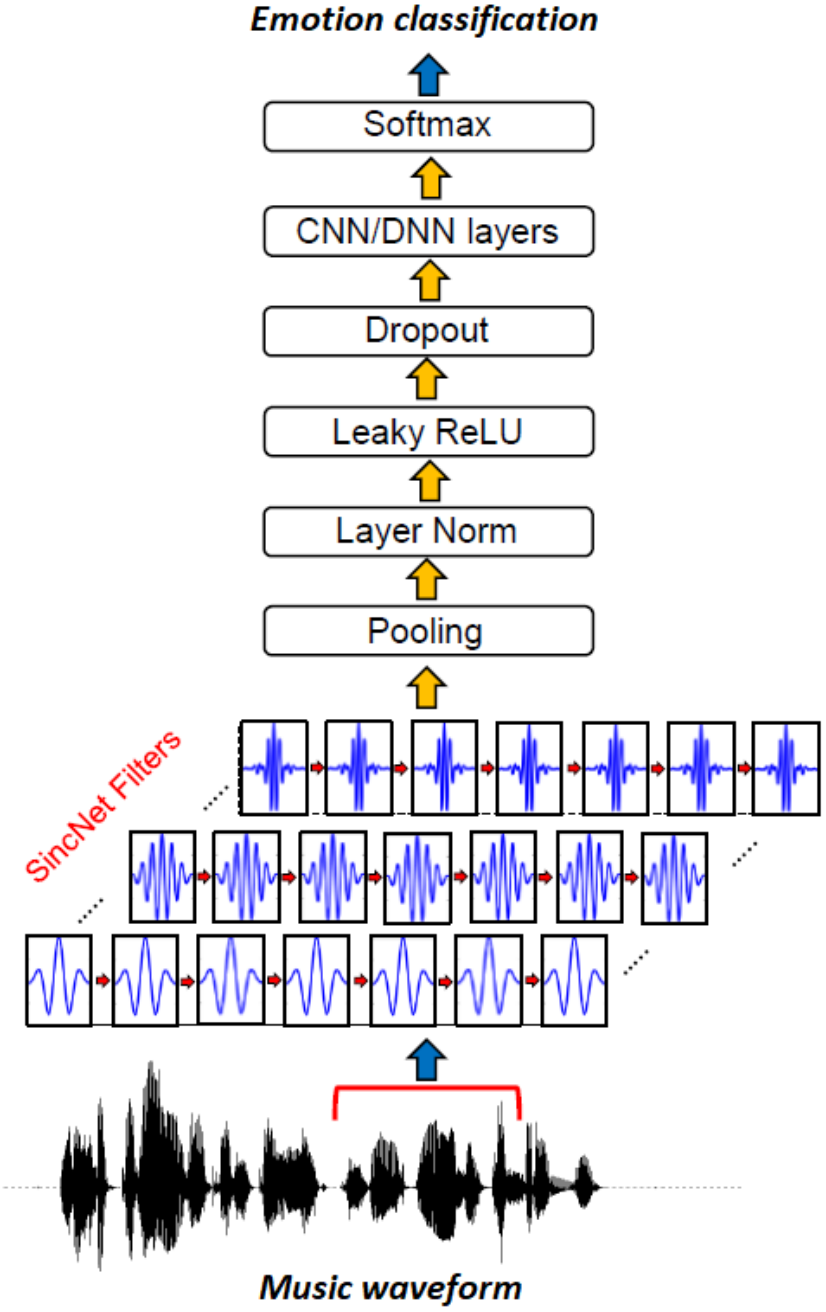

Figure 5: The SincNet architecture [2]: the feature extractor is made of sinc filters, followed by pooling, layer normalization, leaky ReLU activation and dropout, then the fully connected CNN layers are employed for classifying with the Softmax algorithm.

tribute to good emotion recognition results since timbre was found to be a relevant emotional feature in some previous studies [13].

It can be argued that a drawback for using SincNet on MER is the very small chunk size that is tagged when the algorithm is applied, that is of about 200 milliseconds. As it will be described in the next subsection, the whole excerpt will be tagged based on the majority of such frame ratings. While it can be questionable whether emotion can be detected so fast, we should keep in mind that certain nuances of audio can be detected this fast by both human and computer. For tasks like speaker recognition and verification or sound detection this window frame was widely used and effective. At the same time, humans can recognize familiar songs in less than 200 milliseconds [58], which sends to the idea that emotions could also be recognized this fast. Although the chunk is not big enough to emphasize melodic lines or rhythm, deep learning aims to discover underlying patterns that are less expectable. It might also be the case that this frame length is more appropriate for tasks where the output is almost unique to the input, that is the chance of an input to be a specific output to be very low. For our four class prediction, that is not the case, thus the issue should be considered and investigated in future studies.

3.3.1 SincNet Set-up

In our experiments, we mostly kept the original SincNet set-up with only a few modifications in terms of the number of output classes and the optimization parameters. We first fed the model with the tagged 8-second music excerpts, that were split into chunks of 200 ms with 10 ms overlap. The first layer of the network performs sinc-based convolutions using 80 filters of length L = 251 samples. The architecture then employs two standard convolutional layers, both using 60 filters of length 5. Layer normalization was also applied for all convolutional layers, including the SincNet input layer. Next, three fully-connected layers composed of 2048 neurons and normalized with batch normalization were employed. All hidden layers use leaky-ReLU activation functions, that allow a small, positive gradient if the input is negative. The parameters of the sinc-layer were initialized using mel-scale cutoff frequencies,

while the rest of the network was initialized with the Glorot initialization scheme, providing each weight with a small Gaussian value with mean = 0.0 and variance based on the fan-in and fan-out of the weight.

Frame-level emotion classification was obtained through the softmax classifier, providing a set of posterior probabilities over all classes. Note that softmax activations are necessary for single label classification so that one single label can be chosen. A song-level classification was then derived by averaging the frame predictions and voting for the quadrant which maximizes the average posterior probability. Training used the Root Mean Square Propagation optimizer, with an adapting learning rate starting at lr = 0.004, alpha = 0.95, eps = 1e-8 and minibatches of size 128. For all experiments, 100 epochs were enough since there were no improvements beyond that. However, for other configurations, more epochs might be beneficial.

At the same time, we believe that experimenting with larger frame sizes might be useful, given the fact that the 200 millisecond window could be too small for emotion detection. We actually tried out a 500 millisecond frame size, but we had to cut down on the batch size, from 128 to 32, because of the technical limitations of the machine we used for training.

3.4 Cross-Dataset Experiments

In order to determine how relevant context - in this case language - is for emotion extraction algorithms, we made several experiments in a cross-dataset train-test manner. This means that different combinations of datasets were used to train and test the models, in order to observe what the proposed algorithms are able to learn and how well. The settings that we are going to describe were analyzed for both the established baseline model, i.e., the Multi-Layer Perceptron with pre-extracted features, and the proposed deep learning architecture, i.e., SincNet. Where the same experiments were conducted for both models, we will refer to either one of them as "the main model" and this means that the setting was considered for both. It is also to note that the same train-test splits were maintained throughout our

investigations, so that results remain comparable.

After exploring how the main model behaves with each individual dataset, which we identify as within-dataset evaluation, the first cross-dataset experiment we conducted was to train the main model, i.e., firstly the MLP and then SincNet, on music in one language and test it with music in another language. The purpose of this is to analyze the evaluation statistics of the different set-ups in comparison to the original within-dataset train-test configuration. Since the datasets are completely different, it is expected that results are poorer across different datasets, suggesting that only a part of the learned features from music in one language are relevant to music in another language.

The next experiments explore some mixed dataset learning set-ups, where the training is made on data from all three available datasets and results are observed individually for each test set that was used for all tasks. Our assumption in this case states that if emotions are learned from music irrespective to language and culture, then feeding more data to the machine learning algorithms should improve the sensitivity of the model. Thus, training on all the three datasets should give improved results for each individual test set, compared to the independent within dataset training. To make sure the experiment is not biased, we fed the main model with data from all sets in a random order. If this was not the case and music was fed one entire dataset after another, then the model would be mostly tuned according to the last samples, which would be similar to what a transfer learning method does. One thing that should be noted here is that for the MLP model, the mixed dataset was balanced in terms of samples belonging to different languages, as well as in terms of different classes from each language, that is a total of roughly 3000 training samples. On the other hand, with SincNet we employed the full training sets, balanced only in terms of different classes, summing up a total of roughly 5000, in the idea that the end-to-end learning should perform better as the amount of data increases.

Furthermore, with SincNet, we also employed the transfer learning methodology, where we used one of the sets, on turns, as a target set. The initial training had two steps, according to the remaining two datasets, before training on the target set.

Therefore, two sets were used as source sets in each transfer learning experiment. The model parameters firstly adapt to the first source language, then, based on these learned parameters, the model further adapts to the second language and finally, everything that was learned so far in this order, is used as initialization for the fine-tuning step. In this last step, the parameters are optimized with the target train set, such that the general cues that were learned so far from the first models are extended by context-based cues. The two source sets were employed on turns first and second, therefore we had six different configurations with different orders for our three datasets. This set-up was built under the assumption that a general music emotion recognition engine can be fine-tuned with some train data from the same context as the target data in order to gain more insights about the music we are interested in. This method would be feasible for more applications since it does not require a huge amount of data from a new context or culture in this case, as we have seen that dataset creation and verification is a challenging task and most data that is available belongs to the Western culture.

Chapter 4

Results

In this chapter we will present the outcomes of the previously described experiments, from establishing a baseline model to employing an end-to-end architecture on MER, under cultural considerations. We will report classification scores like accuracy, precision, recall, f-score, along with some confusion matrices. Two concepts are identified in this context: positives, that is samples that were classified as the target class, and negatives, that is samples that were classified as other than the target class. Precision refers to the ability of a classifier not to label as positive a sample that is negative, recall is the ability of the classifier to find all the positive samples and f-score is defined by the harmonic mean of the precision and recall, with values between 0 and 1, where 1 suggests the best value. Accuracy refers to the percentage of samples that were correctly classified from the test set, but it is usually not very descriptive for classification problems. Accuracy can become a problem when the data is unbalanced or when false positives are more desirable than true negatives; however, this is not our case so we will still report here, along with the other measures. The support will also be provided in classification reports and it refers to the number of occurrences of each class in the true test tags. More in-detail results will be placed in the first Appendix section.

Note that the scores are reported as weighted averages because, although we balanced the datasets, they got slightly unbalanced after performing the train-test split.

The training sets were rebalanced so that the learning is not biased. The sizes of the train sets are 2880 samples of English music, 1130 samples of Chinese music and 1280 samples of Turkish music. The test sets contain 720 samples of English music, 288 of Chinese and 320 of Turkish.

4.1 Baseline Model

In order to determine a baseline model, we examined two feature sets, predominantly made of low-level descriptors. The first one was extracted with the Essentia feature extractor and the second with the openSmile feature extractor. We will present here the classification results for training and testing with each of these, along with more traditional machine learning approaches. To establish the best algorithm and feature set, we analysed results for all three datasets and chose the set-up that works reasonably well for all.

After extracting and selecting the best features with Essentia, we trained and tested seven different classifiers from Scikit-Learn, that were generally used with default parameters, unless otherwise stated in the previous chapter: K-Nearest-Neighbors (KNN) with n = 3 (determined experimentally), Support Vector Machine (SVC) with linear kernel, Support Vector Machine (SVC) with Radial Basis Function (RBF), Gaussian Process Classifier, Multi-Layer Perceptron (MLP), Gaussian Naive Bayes and Random Forest Classifier (RFC) with maximum depth of 15. By trying these independently for each dataset, we discovered that the best three algorithms are, in descending order, MLP, Random Forest and KNN for English, KNN, RBF SVM and Random Forest for Chinese, Gaussian Process, MLP and Random Forest for Turkish. However, there are only small differences between these.

As it was observed that the Random Forest classifier was among the best ones for all three datasets, we report its results here, along with its confusion matrices in Figure 20. We will use these to compare it to the best algorithm trained on the IS13 ComParE features. From what we see in Figure 6, the model learns something, especially with English and Turkish music, but the scores are not very high. Also,

	English	Chinese	Turkish
Precision	61%	27%	67%
Recall	58%	25%	66%
F-score	58%	26%	65%
Accuracy	58%	25%	66%

Figure 6: Results for Essentia features and Random Forest classifier.

	English	Chinese	Turkish
Precision	65%	23%	74%
Recall	63%	30%	71%
F-score	62%	23%	71%
Accuracy	63%	30%	71%

Figure 7: Results for IS13 ComParE features and Multi-Layer Perceptron classifier.

the Chinese set seems problematic, with scores at around 25% which is very close to a random assignment of quadrant tags.

The same process was applied with the IS13 ComParE feature set. Best three algorithms in this case were MLP, Random Forest and KNN for all three datasets. Since the MLP had the top score for both English and Turkish and the score for the Chinese set was very close to the very first in its case, we selected the MLP as the best algorithm for this second feature set. We report the summaries of the individual classification reports of this set-up in Figure 7 and the corresponding confusion matrices in Figure 8. In addition, best 3 algorithms for each feature set and each language, in terms of accuracy, are reported in Figure 19, in the Appendix.

From these first two tables, it was inferred that the most promising baseline configuration is the IS13 ComParE feature set trained with the Multi-Layer Perceptron. Scores are not much better, they are only a few units higher for English and Turkish, while for Chinese they are very comparable. We will keep these final results as reference for analyzing the behaviour of the SincNet architecture on our three

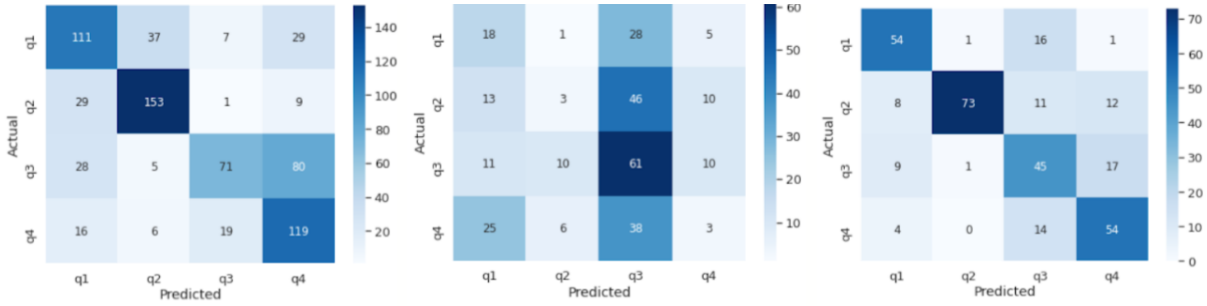

Figure 8: Confusion matrices for within-dataset baseline classification with IS13 ComParE and MLP. Darker blue means that more samples were classified that way.

datasets and also for the cross-dataset and transfer learning set-ups.

We also performed some extra experiments with the combination of the two feature sets on the MLP, where we selected again the best 50 features, but this time from the combined set. The distribution among these was almost equal, interestingly with slightly more features from Essentia. Best 10 features extracted with this model can be seen in Figure 18, in the Appendix, while the appropriate confusion matrices can be observed in Figure 21. With this combination of features, we obtained better results with the English and Turkish sets, in comparison to the IS13 ComParE established baseline, with an increase of 3 and 6 percent points respectively in accuracy, but also a decrease of 12 percent points for Chinese. If we disregard the Chinese set, we can observe that the combination of the two is beneficial, therefore it can be further explored. In figure 22, we can also regard the MLP results with all the three feature sets and combinations considered.

4.2 End-to-End Model

The first research on SincNet was made independently for each dataset, in order to observe how it compares to the traditional machine learning approach established as baseline. The SincNet architecture was employed as described in the previous chapter and results can be seen for each of the three datasets in Figure 9, along with their corresponding confusion matrices in Figure 10. The results that are emphasized for SincNet are generally worse than those from the baseline model in Figure 7 and many issues appear in the third and fourth quadrants. Although we proposed this

	English	Chinese	Turkish
Precision	59%	11%	68%
Recall	57%	27%	63%
F-score	52%	16%	58%
Accuracy	57%	27%	63%

Figure 9: SincNet results within-dataset.

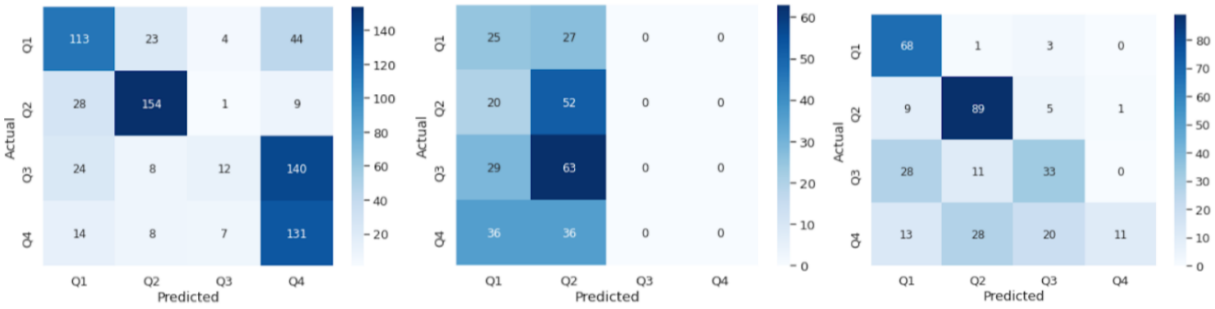

Figure 10: Confusion matrices for within-dataset classification with SincNet. Darker blue means that more samples were classified that way.

network in the idea that it could be better than the baseline traditional machine learning, it seems that it actually performs worse. However, we should keep in mind that it has been trained on waveform signals only, which could represent an optimized approach for some tasks and a challenging task for others.

In the attempt to overcome the issue of optimization, we also experimented with a bigger frame size for the within-dataset configuration, that is 500ms instead of the original 200ms window. These results tend to be better than with the original SincNet and this can be seen in Figure 23, in Appendix A, along with the corresponding confusion matrices in Figure 24. We perform the rest of the experiments with the original 200ms frame size, but this new set-up could be further explored.

4.3 Cross-Dataset Evaluation

After performing all experiments and comparisons with independent datasets, where the models were tested with music in the same language as the train set, we decided

to analyze results in a cross-dataset manner. By this, we mean that the train data is completely or partially in a different language than the test data. Our initial assumption was that MER models should be different for different languages, therefore we expect that these cross-dataset evaluations to give poor results. This assumption was developed from the study of Hu and Yang [10] where experiments were made on more datasets with different language speaking annotators and results showed in their case that models that had culturally correlated music and annotators were more effective.

4.3.1 Baseline Model

A first investigation we conducted in order to discover the various differences between the three datasets coming from three different cultures was to observe the selected best features for each of them. More exactly, we analyzed the similarity of the best 10 features, extracted with the SelectKBest function from Scikit-Learn, in the context of both proposed feature sets. In the case of Essentia, features like spectral centroid mean, melbands spread mean, spectral skewness standard deviation and zero-crossing rate mean were found to be relevant and common for English and Turkish music. The Chinese set is mostly influenced by various silence rates and spectral energybands statistics. These can be seen in Figure 17. However, as it was observed, results with this set are not as good as with the other feature sets, so maybe these features are not that relevant either.

With the IS13 ComParE features, we similarly observed many similarities between English and Turkish music, suggesting at a first impression that the machine can not differentiate too much between emotions extracted from music in different languages, especially with this feature-based approach. We added the best 10 features to the table in Figure 11. We highlighted features that were found to be common or somehow related to two or all three sets. In yellow we can see statistics related to spectral roll-off, spectral centroid or spectral variance, that are common to English and Turkish. We also highlighted in red what was common to all three sets, that is the logHNR (logarithmic harmonic-to-noise ratio), while in blue we have the only

English	Chinese	Turkish
logHNR_sma_stddev pcm_fftMag_spectralRolloff90.0_sma_amean pcm_fftMag_spectralVariance_sma_amean mfcc_sma[1]_amean pcm_fftMag_spectralRolloff75.0_sma_amean pcm_fftMag_spectralFlux_sma_amean audspec_lengthL1norm_sma_amean pcm_fftMag_spectralCentroid_sma_amean logHNR_sma_amean pcm_zcr_sma_amean	logHNR_sma_de_stddev logHNR_sma_stddev pcm_fftMag_fband1000-4000_sma_stddev F0final_sma_stddev pcm_fftMag_spectralHarmonicity_sma_de_stddev pcm_fftMag_spectralSkewness_sma_de_stddev pcm_fftMag_fband1000-4000_sma_de_stddev pcm_fftMag_fband1000-4000_sma_amean mfcc_sma_de[1]_stddev shimmerLocal_sma_amean	logHNR_sma_stddev pcm_fftMag_spectralCentroid_sma_amean pcm_fftMag_spectralRolloff75.0_sma_amean pcm_fftMag_spectralRolloff90.0_sma_amean pcm_fftMag_spectralRolloff50.0_sma_de_stddev pcm_fftMag_spectralVariance_sma_amean pcm_zcr_sma_amean pcm_fftMag_psySharpness_sma_amean audspec_lengthL1norm_sma_de_stddev pcm_fftMag_spectralRolloff50.0_sma_amean

Figure 11: Best 10 IS13 ComParE features selected for each dataset. In yellow are spectral features common for English and Turkish, in red harmonic-to-noise levels common for all three, in blue the common MFCC-related feature.

Train\Test	English				Chinese				Turkish			
Score(%)	P	R	F	A	P	R	F	A	P	R	F	A
English	**65**	**63**	**62**	**63**	**29**	19	17	19	53	48	48	48
Chinese	22	24	22	24	23	**30**	**23**	**30**	21	25	22	25
Turkish	51	49	50	49	25	23	21	23	**74**	**71**	**71**	**71**

Figure 12: Results of cross-dataset experiments on the baseline model: the model is trained with music in the language on the left column and tested with music in the language on the first row. In bold are the best scores for each language, mostly within-dataset configurations.

MFCC related feature common to English and Chinese.

Another thing that we wanted to explore was how the established baseline model behaves when trained with music in one language and tested with music in another, that is the comparison between within-dataset and cross-dataset evaluation. In Figure 12 we can observe the classification summaries with Precision (P), Recall (R), F-score (F) and Accuracy (A) of all nine experiments on the Multi-layer Perceptron model with IS13 ComParE input features. As expected, cross-dataset results are worse than within-dataset, suggesting a possible need of context oriented models. We also report their confusion matrices in Figure 26, as well as results for the same set-up, but with Essentia features, in Figure 25.

Furthermore, we created a mixed training dataset made from all the three training sets we used so far, which we randomly shuffled so that results are not biased. In the case of the mixed training for the baseline model, we also considered an

	English	Chinese	Turkish
Precision	60%	22%	67%
Recall	57%	23%	64%
F-score	56%	22%	64%
Accuracy	57%	23%	64%

Figure 13: Results of mixed training and individual testing on the baseline model.

equally distributed train set, in terms of numbers of samples belonging to the same quadrant from each dataset, which means that the same amount of music from each of the three cultures was fed to the model. Best 10 features in this case can be seen in Figure 27 in the Appendix. We trained the baseline model and evaluated it separately for each of the three test sets in English, Chinese and Turkish. As it can be seen in Figure 13, scores are slightly worse than those in Figure 7, where the same model configuration was employed, but with independent within-dataset training. The confusion matrices can also be seen in Figure 29. This suggests that training with more music, from more cultural backgrounds is indeed noisier than maintaining the same background. Differences are not very big, but the sensitive models we desire would probably benefit from cultural considerations. Interestingly, mixed training with Essentia provides slightly better results than the selected baseline model, as it can be seen in Figure 28.

4.3.2 End-to-End Model

For the end-to-end model, i.e. SincNet, we performed three different investigations in which multiple datasets were involved. These are the cross-dataset, mixed training and transfer learning evaluations. For cross-dataset we trained the network on music from one culture and tested it with music from another, under all six possible combinations, just like with the baseline model. For mixed training, we created a big training set made of all available training instances and examined results with each of the individual test sets. Note that in this case, the amount of data from each language was not balanced and the English dataset is twice as bigger than the

Train\Test	English				Chinese				Turkish			
Score(%)	P	R	F	A	P	R	F	A	P	R	F	A
English	**59**	**57**	**52**	**57**	14	20	14	20	57	46	41	46
Chinese	9	23	13	23	11	**27**	16	**27**	21	39	27	39
Turkish	48	41	38	41	**24**	24	**24**	24	**68**	**63**	**58**	**63**

Figure 14: Results of cross-dataset experiments on the end-to-end model. In bold are the best scores obtained for each language; most of these are from within-dataset set-ups.

other two. Finally, for transfer learning, all three independent training sets were fed to the same model, but on turns, in specific orders, such that six configurations were obtained. All of these were proposed so that we can analyze the relevance of language and cultural considerations in MER.

The cross-dataset evaluations with SincNet are reported in Figure 14 in a similar format to the cross-dataset evaluations on the baseline model. Also, the cross-dataset confusion matrices are reported in Figure 30. Again, it can be observed that scores from within-dataset are better than cross-dataset and confusions are significant under this confuguration.

A further test that was previously conducted for the baseline model as well, is the mixed training, where all training data is combined, shuffled and fed to the model. Particularly in the case of an end-to-end deep learning model, more training data would usually turn into better classification results. However, with music in mixed languages and considering the fact that there were twice as many Western music instances than Chinese and Turkish, this aspect is debatable. Unlike the mixed dataset experiment with the baseline model, where all scores are clearly lower than within-dataset evaluation, here there are three different behaviours for each language. By comparing the mixed training results in Figure 15 (and their confusion matrices in Figure 31) to the initial results of SincNet within-dataset from Figure 9, we can observe that for English, results are very similar, with an even higher f-score when mixed, for Chinese, results are better in terms of only precision and accuracy,

	English	Chinese	Turkish
Precision	58%	26%	61%
Recall	57%	23%	51%
F-score	56%	21%	50%
Accuracy	57%	23%	51%

Figure 15: Results of mixed training and individual testing on the end-to-end model.

while for Turkish all scores appear to be significantly worse when mixed training is employed. In the case of English, results are understandable since the main part of the mixed training was made on the same dataset as within-dataset evaluation. For Chinese, it seems that as within-dataset results are very poor, considering features from other reliable sets could be beneficial. Finally, within dataset Turkish results were very good from the beginning, thus adding noisier sets to the model, especially the Mandarin one, and also the big amount of English data, concludes with poorer outcomes.

In what concerns transfer learning, we decided to adopt this method after we have already seen that combining more datasets leads to classifications with lower performance in some cases and higher in others. However, because the noise is relatively small, we believe that fine-tuning on the appropriate language dataset could enhance the quality of a pre-trained classifier, that was initially fed with similar data, but in a different language. In this context, we decided to employ a double transfer learning approach, in which SincNet is firstly trained in one language, then tuned on a second and finally fine-tuned on the third target language. Thus, we decided to compare the results of this model on the target test set to the results of SincNet trained and tested on the same dataset.

There are six such transfer learning configurations and their results are reported in Figure 16, along with the other experiments we conducted, such that it is easier to observe differences and patterns. Note that our main goal was to observe the results when testing on the target set, but tests on all three datasets were performed. In

Figure 16, we underlined the scores from the best configuration for each dataset and we also highlighted in bold the best SincNet configuration for each set. It can be seen that in some cases, our assumption is indeed valid and results are better when transfer learning is applied. For example, note that the best of all set-ups for Turkish music is the 4Q-CH-TR transfer learning, or that the best SincNet set-up for English music is the TR-CH-4Q transfer learning. What is indeed interesting is that the relatively good (compared to the others) SincNet set-up for Chinese music is the mixed dataset one (especially in terms of f-score) which could suggest some good influences from the other languages. We report the confusion matrices evaluated on these target sets in Figure 33.

We also tried to test these six transfer learning set-ups with the two source sets. It seems that several cues are learned and therefore these are probably common to the three datasets. On the other hand, results are poorer than within-dataset, suggesting that the last fine-tuning added noise to the source training as it is made on another language. These results are provided in Figure 32 from Appendix A.

As we observed that English and Turkish had more in common, we explored some extra set-ups with transfer learning with only these two datasets. Surprisingly, results are not better than our initial three-step transfer learning, as depicted in Figure 34, suggesting that some cues are also learned from the Chinese dataset. This partially contradicts the hypothesis that more languages only bring noise and leads to the possibility of a more general MER model, that is only context-based fine-tuned.

Model		Baseline: MLP				End-to-end: SincNet				S
Configuration	Data (train / test)	P	R	F	A	P	R	F	A	
Within dataset	4Q / 4Q	65%	63%	62%	63%	59%	57%	52%	57%	720
TL 1 - finetune 4Q	CH-TR-4Q / 4Q	-	-	-	-	57%	56%	51%	56%	720
TL 2 - finetune 4Q	TR-CH-4Q / 4Q	-	-	-	-	**61%**	**60%**	**57%**	**60%**	720
Mixed dataset	ALL / 4Q	60%	57%	56%	57%	58%	57%	56%	57%	720
Within dataset	CH / CH	23%	30%	23%	30%	11%	**27%**	16%	**27%**	288
TL 1 - finetune CH	4Q-TR-CH / CH	-	-	-	-	10%	24%	12%	24%	288
TL 2 - finetune CH	TR-4Q-CH / CH	-	-	-	-	28%	26%	17%	26%	288
Mixed dataset	ALL / CH	22%	23%	22%	23%	26%	23%	**21%**	23%	288
Within dataset	TR / TR	74%	71%	71%	71%	68%	63%	58%	63%	320
TL 1 - finetune TR	4Q-CH-TR / TR	-	-	-	-	75%	75%	75%	75%	320
TL 2 - finetune TR	CH-4Q-TR / TR	-	-	-	-	72%	71%	71%	71%	320
Mixed dataset	ALL / TR	67%	64%	64%	64%	61%	51%	50%	51%	320

Figure 16: Summary of all results from our experiments. In red are experiments evaluated on the English set, in purple Chinese and in blue Turkish. TL 1 and TL 2 are transfer learning set-ups with two source sets in the order provided in the second column. P is precision, R is recall, F is f-score, A is accuracy, S is support. Underlined is the best of all configuration. In bold in the end-to-end columns are the best configurations on SincNet.

Chapter 5

Discussion

In this chapter we will provide our inferences and conclusions on the previously presented results. Some of our assumptions were partially met, but further investigations would be recommended in order to clarify some of them.

One inconvenient aspect typical to MER that we had to deal with was the questionable quality of datasets, which is usually due to the ambiguity of emotion and the low agreement between annotators. Our main problems were raised around the Chinese dataset, possibly because of the segmentation we performed on it. Many MER studies use 30-second excerpts and that is why our datasets were initially built this way, but we split these into four parts because 7-8 seconds seemed a reasonably long time to detect emotion and also because we thought that more training instances would be beneficial. What we observed by experimentally listening to a few samples from the Chinese set, is that there are several samples that seem to evoke the same emotion but are annotated differently or the other way around. At a more in-depth analysis of the Chinese culture by asking native speakers to listen to some excerpts in the Chinese dataset, we found out that emotions in their culture are based on lyrics and poetry much more than on sounds, therefore the original annotations might be related to that more than to audio features. Moreover, the songs in the set seem to belong to different dialects and different time periods and that is why our models were not too good in finding patterns when fed with music

from this dataset. As emphasized in the previous tables, results with this set often fall below 25%, corresponding to chance guessing, which means that the models are not able to learn much from training or they even learn mistaken correlations.

Our general results are not particularly good, therefore the configurations we considered can only act as comparison references. Our best results with the 4Q-DATASET were obtained with the baseline model under the within-dataset configuration, followed by one of the transfer learning approaches with SincNet. The Chinese set, CH-818, performed poorly in all tests, but the best one seems to remain the baseline MLP within-dataset. In the case of TR-MUSIC, best results come from a double transfer learning model, followed by some very close scores for with within-dataset baseline again, and also the other transfer learning set-up.

Although we expected that the problem of four-class classification to be easier and more approachable since only some general cues should be learned, it might also be the case that specific boundaries are more difficult to establish. For instance, in the case of the Mandarin dataset, we observed that many music segments were initially annotated with values very close to 0 for either arousal or valence. This means that when two samples that are relatively close to each other in the VA plane, but delimited by one or two axes, are classified as different emotion classes, the algorithm might get a bit confused as it receives contradictory training examples that are similar in terms of features. This could also be the reason why continuous valence-arousal predictors tend to be more efficient when an appropriate dataset is provided, e.g., DEAM, like in the works of Dong et al. [40] or Orjesek et al. [38]. Contrary to our results, Orjesek et al. conclude that CNNs have a great potential to learn feature representations from raw audio signal, providing state-of-the-art results for continuous valence-arousal prediction. Despite the smaller number of parameters, their end-to-end model shows improvements in the valence domain when compared to feature-engineered solutions.

Cowen et al. also provide an analysis in the support of a rich array of distinct categories bridged by smooth gradients [59], showing that several distinct dimensions of subjective experience were found to be equally relevant for the Western and

Chinese cultures, which contradicts our assumption. Associations of music with specific feelings such as amusement or desire were better preserved across cultures than associations with valence and arousal, raising some questions about this type of ground-truth. Smooth gradients between categories are proposed as a more accurate cross-cultural representation. In addition, Yang and Chen [60] created a ranking-based emotion recognition system that tags songs based on their similarity to an already tagged corpus, in order to improve the reliability of the ground-truth data and move research towards a wider continuous space of emotion representation.

For most of the remarks we are going to make, we will refer to Figure 16, as it contains relevant information for most of our experiments. Note that there was no transfer learning approach for the baseline model, i.e. the MLP, but this could be included in further investigations. We believe that for our scopes here, we can argue that the transfer learning on SincNet is enough to observe the general tendencies. This is because firstly, SincNet results are quite close to the baseline results, therefore it can be trusted to some extent. Also, the architecture is constructed to pay attention to language and speech characteristics, thus it should be more accurate when fine-tuning it with different languages.

5.1 Baseline Model

In this section we are going to comment on our experiments on the baseline model, in particular the within-dataset ones. Results show that the baseline Multi-Layer Perceptron generally performs better than the other approaches that were considered with the nature and amount of data that was available.

5.1.1 Features

The first aspect we investigated on the baseline model were the selected features for training the machine learning classifiers. Two different feature sets were compared and the fact that, after feature selection, we obtained different feature sets for the different training datasets considered, suggests that music - emotion correlations are indeed different from one culture to another, or at least machines identify them as

different. We also considered and compared the most relevant 10 features, which were also distinct (Figure 11, 17). What is interesting is that English and Turkish had several common features as opposed to the Chinese set, which was completely different. This could be reasoned by the fact that the cultures are closer geographically and historically speaking, therefore in terms of language influences as well, which would support the context-based emotion theory.

In terms of the combined feature set we experimented with, that is with features extracted with both Essentia and openSMILE, results were considerably improved in some cases. The fact that many of the best selected features come from Essentia and not IS13 ComParE which was established as baseline, could suggest that the most relevant low-level descriptors from Essentia might be more robust (Figure 18). However, since IS13 ComParE had initially more descriptors to select from and it was developed for several years in this specific direction of discovering affective content, it scored better overall when the two were compared. On the other hand, the features from IS13 ComParE were designed for speech and speaker related task (which makes the baseline even more comparable to SincNet), while Essentia might be more focused a bit more on music events, that is why it could be more robust in this context where we work with music. For this combined features set-up, we also need to take into account that several low-level descriptors are common within the two initial sets, therefore some redundancies might exist.

Looking at the selected baseline model, i.e., the IS13 ComParE feature set trained on the MLP algorithm, emotions depicted in English and Turkish music seem to be discriminated by means of various spectral features, while in Chinese music, most relevant cues are more about the statistics of frequency bands, as reflected in Figure 11. One possible argument for this is that emotions in their culture could be perceived under different considerations, not that much in terms of what chords and harmonies are used as in the height of most frequencies. It would be interesting to further study this possibility and also discover whether this can relate to the fact that Chinese is a tonal language. The feature that was found to be common to the three languages is the Harmonic-to-Noise Ratio (HNR), i.e. the ratio between

periodic components and non periodic components. This could probably depict several timbre characteristics, that were found to impact emotions considerably in previous studies [13].

5.1.2 Algorithms

In terms of machine learning methods, there were several algorithm configurations that seemed to perform better than the others, independently from the feature set as well as from the language of the dataset. This means that a good music emotion recognition solution should detect an appropriate classifier for the task. In our experiments with the baseline Scikit-Learn algorithms, we found that the best results were found with the Multi-Layer Perceptron, the Random Forest classifier and the K-Nearest Neighbours classifier (Figure 19). These were found to be competitive with both feature sets that were employed, reinforcing the fact that they could be appropriate for the task, at least to some extent.

The main two configurations that competed for the chosen baseline were the Random Forest with Essentia and the MLP with IS13 ComParE. We chose the MLP to represent our baseline as it had few percent points in addition to the others, and we also considered it a reasonable and comparable starting point for the proposed deep learning model as it works based on the same neural network principle. The MLP also has the advantage that it could also be improved by adjusting the parameters more and adding more layers and neurons.

5.1.3 MLP Results

In terms of prediction performance, the Turkish data generally scored best, with about 71% of test data correctly classified in terms of accuracy and f-score in the case of the baseline MLP (Figure 7). This was observed despite the fact that the English dataset is considerably bigger and supposed to allow for better learning. That is why we can infer that the emotionally discriminating features in this culture are more obvious and rigorously described or at least easier to learn computationally. Perhaps it is also the case that the Western culture comprises a wider variety of

influences and styles, which could make it more difficult to generalize. In this sense, Hult et al. [61] support similar ideas after conducting a study that explores the transferability of MER systems given the quality and size of the datasets. They conclude that training a prediction model on a homogeneous dataset with carefully collected emotion annotations yields a better foundation for future predictions than training on a larger, more varied dataset, with less reliable annotations.

By looking at the confusion matrices provided in Figure 8, we observe that with the English dataset there is a big confusion of Q3 for Q4, suggesting that, for this particular culture, we need more discrimination between sad and relaxed music, i.e. the valence of low arousal songs. This is not very surprising since past research has also found that arousal is easier to predict, while valence is more subjective, cultural-specific and personal. A hand-crafted option for this problem could be to give more credit to whether the key is major or minor for the case of tonal music, but our aim here is to automatically detect such associations. Moreover, we note that the English and Turkish datasets behave similarly in terms of confusions that were made, for instance Q1 and Q2 are well classified for both. On the other hand, the Chinese model seems to be biased into classifying many samples as Q3, even though the training set was balanced.

Our results with the 4Q-EMOTION dataset align with the previous results with this dataset by Panda et al. [54], in the sense that they also conclude that emotions in songs with higher arousal are easier to differentiate. Q2 was very well classified in their case, too, as it belongs to genres such as punk or heavy-metal, which have very distinctive, noise-like, acoustic features. As they are the ones who created the dataset, the authors report that subjects also had more difficulty distinguishing valence for songs with low arousal, thus it is explainable why the machine also had issues. They used SVMs with RBF and carefully selected parameters and report a better f-score - 76%, compared to our 62%.

The CH-818 dataset was previously used in the study of Hu and Yang [10] in a similar cross-cultural and cross-dataset manner, but the valence and arousal dimensions were predicted separately as continuous values. In the case of their within-dataset

evaluation, results show an R-squared score of 0.81 for arousal and 0.19 for valence (where the R-squared measure refers to how close the data is to the fitted regression line), which means that they also had issues with the valence dimension. A further cross-cultural study by Chen and Yang [48] that uses a 1D-CNN on the same dataset achives even better R-squared scores of 0.82 for arousal and 0.24 for valence, arguing that the timbre feature is the one to consider for better valence prediction.

In the case of the Turkish set, TR-MUSIC, experiments were previously made by Er and Aydilek [37] with both the original dataset and an augmented version of it. The best classification success before data augmentation is obtained with transfer learning from the VGG-16 pre-trained network along with the Softmax classifier - 76% accuracy. After the data augmentation, the best classifier success was obtained from the same network with a SVM classifier - 89.2% accuracy. Compared to our 71% accuracy with this set, with the MLP with pre-extracted features and without augmentation, their results recommend data augmentation, as well as transfer learning as two processes to be taken into account in further studies, especially because employing pre-trained networks is perhaps faster than handcrafting features or employing a new model.

5.2 End-to-End Model

Our end-to-end proposal for music emotion recognition with the SincNet CNN architecture proved not to be very effective under the set-up and variables presented here, especially for within-dataset evaluations. Unlike in the case of its original application in speaker recognition where results with one of the datasets achieved a classification error rate of 96%, our best results with SincNet only show a 63% accuracy for within-dataset prediction and a 75% accuracy with transfer learning. We will analyze the results of various training configurations in this and the next sections.

5.2.1 SincNet Results

Our first goal with the end-to-end training with SincNet was to compare it to the established baseline model in order to determine whether the architecture is appropriate for the task and whether its higher level of complexity is worthy. In this case, we observe that the model does not outperform the baseline (Figure 9) and it also takes a lot more time to train. One reason for the poorer results could be that deep learning usually requires very large amounts of data in order to reach its full capacity. It was previously shown that deep architectures start to outperform traditional machine learning algorithms only after a big number of training samples [12]. When relatively small datasets are available, domain assumptions and feature engineering is more effective; assumption-free networks, i.e. with raw waveform input, were truly useful in the case of music tagging only after 1 million songs were provided.

By looking at the confusion matrices obtained from within-dataset SincNet (Figure 10), we observe that the Chinese model never assigns Q3 or Q4 tags, possibly suggesting some issues with SincNet and the arousal values learned from this dataset. For Turkish there are also few samples classified as Q3 or Q4, but some correctly identified instances exist. Finally, the considerable confusion of Q3 with Q4 for the 4Q-EMOTION dataset, stresses on the issue that the annotations might not be very reliable and consistent, given a relatively diverse music corpus.

One thing that should be considered when analysing the SincNet evaluations is the way it actually tags segments and learns from them. The 7.5-second segment is fed to the architecture, which splits it again into very small chunks of 200 milliseconds, with 10 milliseconds overlap. These are independently tagged and then a voting is performed based on all chunk tags in order to select the most probable class. This chunk size was originally designed for speaker recognition, but we would like to explore the architecture with longer time frames in the future, as they might be more appropriate for this type of system. We provide here some preliminary results with 500ms frame size, but we had to cut down on the batch size due to technical

limitations. Results are partially improved, but only with a few percent points that do not outperform the baseline and not in the case of Mandarin music (Figure 23). It seems that in this last case, all test data is tagged as Q1, which clearly makes larger frames subject to biases (Figure 24). On the other hand, for English and Turkish there are less confusions between Q3 and Q4, which could suggest a possible good impact on the issue of low arousal valence prediction.

5.3 Cross-Dataset Evaluation

In this section, we provide our comments and inferences in regard to all experiments that involved more than one dataset. Interestingly, we observe that SincNet is much more competitive to the baseline in this case, with some outperforming set-ups, suggesting that the architecture might indeed capture some speech related emotion characteristics.

5.3.1 Cross-Dataset

Our hypothesis gets confirmed with the cross-cultural experiments, where cross-dataset testing proves to be less accurate than within-dataset. In Figure 12 we can observe that the lowest results were obtained when either training or testing with the Chinese dataset, which could imply that this culture is significantly different from the other two. In fact, we already know that the Chinese language is completely different from the other two languages as it is a tonal language. Because we argued previously about the shared acoustic codes with emotional meaning of speech and music [45], it could also be the case here through the fact that Chinese people are trained to associate certain positive or negative words to certain tones. However, since we did not find any reasonable results with this dataset, we can not state too much about cross-dataset results either.

In a similar setting to the cross-dataset analysis on the baseline model, we ran cross-dataset experiments on SincNet. These are reported in Figure 14 and it was again observed that within-dataset evaluation is more accurate, with very poor results when training or testing with the Chinese set. Looking at the cross-dataset

confusion matrices for both baseline and end-to-end architectures (Figure 26, 30), we made several inferences. When trained with English music, thus from a Western perspective, most of the Chinese music excerpts were classified as Q4, that is relaxed and peaceful music. Also, when training in Turkish and testing with English, it seems that a majority of Q2 samples are mistaken for Q1, reiterating the issue of valence prediction cross-culturally. We should keep in mind that Q2 was a well classified quadrant for both datasets.

Note that these cross-dataset experiments have similar outcomes to those of Hu and Yang [10], where cross-dataset evaluations with English and Chinese music give poor results in comparison to within-dataset. That is also why Chen and Yang [48] propose a better solution for cross-dataset transferability, that of an adaptive neural network, that actually outperforms the former study.

5.3.2 Mixed Training

When training the baseline model on the equally distributed mixed dataset, results show some similar patterns to those discovered through individual experiments, like the very good classification of Q2 in the English and Turkish test sets (Figure 29). Q3 also seems to be the weakest class among these two datasets, in this experiment but also in the first within-dataset set-up. On the other hand, with Chinese it seems that Q3 is classified relatively well, suggesting that Q3 could be a more subjective class depending on cultural background and language.

An interesting observation was made upon the fact that although the IS13 ComParE feature set performed better than Essentia at within-dataset evaluations, in this mixed dataset set-up, results with features from Essentia show better model performance (Figure 28). One inference could be, again, that features from Essentia might be more general and musically motivated while the other set was designed to detect speech related cues, therefore when languages are combined, generalisation is better achieved by not taking into account many speech cues.

By looking at the confusion matrices from the baseline (Figure 29) and the end-

to-end (Figure 31) models, we observe a common trend of classifying most Chinese music as either Q3 or Q4, suggesting that the CH-818 dataset probably has music with lower arousal than the other two sets. Scores are generally similar within the two mixed training set-ups (Figure 13, 15), apart from Turkish, where we hypothesize that the bigger amount of English data that was used for mixed training with SincNet had a negative impact, due to a wider variety and lack of reliability (as suggested in [61]). Scores within-dataset remain better or at least comparable for both baseline and SincNet, therefore more music did not increase the sensitivity of either model.

For the mixed dataset configuration on SincNet, we used the full training sets from the three cultures. Note that results might be influenced by the fact that the English set which had twice as much samples than the other two which were of comparable sizes. We performed these experiments in order to observe how the end-to-end network behaves when fed with as much data as possible. However, results were not very good, especially for English and Turkish, which replicate the results from performing mixed dataset learning with traditional ML. Nevertheless, if emotional response and features were similar enough within the three datasets, the evaluation results should increase as we fed the last model with more data. But this was not the case, with some even worse results, thus, we can infer that different cultures have different ways to discriminate and interpret emotions. Culture could be an influential factor in emotion prediction and context-based models are desired or should at least be further studied.

5.3.3 Transfer Learning

Another thing that was observed is that, even though the end-to-end model is not better than traditional ML, it is interesting to see that results appear close or even better when using transfer learning from one dataset to a second and to a third target set. This strengthens the idea that more data enhances the end-to-end model. This effect was seen under various circumstances for all three datasets.

With English music, the best set-up for SincNet was pre-training on Turkish, then

Chinese and fine-tuning with English. This reasons for some common characteristics to all three datasets, that can help each other improve. The best results remain with within-dataset Multi-layer Perceptron classification with pre-extracted features, as emphasized in Figure 16. This suggests that further work is needed in order to adapt the SincNet architecture to Music Emotion Recognition, as it will be suggested in the last section. The transfer learning confusion matrices (Figure 33) maintain the previously found patterns for testing with Western music.

With the Chinese dataset, best results remain as well for within-dataset baseline. We can also note some patterns in the scores in the sense that precision and f-score are considerably lower than recall and accuracy but this happens because we are reporting weighted average scores and results can be that different. As precision is more a measure of quality and recall of quantity, this means that the quality of our learning system is indeed bad, even though it manages to get some samples right. In contrast to previous confusions made for this dataset (where Q3 and Q4 were usually not predicted at all), with the two set-ups considered here, most of the test instances are classified as Q2. In contrast to the mixed learning set up, where many instances were predicted as Q4, this time instances were predicted as the exact opposite quadrant. This suggests that although useful features can not be learned, the Mandarin music is found by our models to be quite homogeneous in terms of emotion expression. This might align to the fact that emotions in Chinese music are believed to be more about lyrics comprehension than acoustic events.

In what concerns Turkish music, best results from all experiments were obtained with transfer learning from English to Chinese to Turkish. In fact, these are the highest scores within all datasets and configurations, with precision, recall, f-score and accuracy at 75%. These are slightly better than the baseline and significantly better than training SincNet with Turkish music only. Results with the other transfer learning set-up with fine tuning on Turkish also provides good results, which suggests that perhaps the English dataset is enhancing the SincNet training. If we recall the feature analysis we conducted, these two datasets had many similar features among the best 10, therefore we can infer that either the two cultures are

not very different or that context-based models are not that helpful as we expected. However, we can also see that with mixed training performed at the same time and with a random order of samples, results drop significantly for both baseline and end-to-end mixed configurations. Therefore, we can conclude that while there are certain common characteristics that can build a universal MER model, there are also specific aspects that can fine-tune the model and provide better classification results.

Given that we achieved this top result with the TR-MUSIC dataset, we wanted to compare our results to the previous MER study made on it by Er and Aydilek [37]. By extracting chroma spectrograms and employing some intensively pre-trained CNN architectures, i.e., AlexNet and VGG-16, their best score with this dataset and without data augmentation was an accuracy of 76%. The fact that we managed to achieve about the same score - 75% - with a brand new architecture that was only used for speaker recognition so far, and also by using the raw waveform input format makes our approach very promising and worthy of further investigations.

Although some transfer learning configurations achieved competitive results for the dataset they were fine-tuned on, we made several observations with testing on the other two sets, as the models were constructed partly on them, too (Figure 32). For instance, the two set-ups with the last fine-tuning on the Mandarin set,i.e., 4Q-TR-CH and TR-4Q-CH, achieved their best results when tested with Turkish music. This is not very surprising since the dataset was the easiest to predict. The fact that the Western dataset achieved lower scores, especially with 4Q-TR-CH, suggests that its relevant features are not too robust across more datasets. Last but not least, the fact that English and Turkish tested relatively well when the fine-tuning was made on the other of the two, implies that they indeed share some similarities in terms of musical emotions.

That is why we also performed a transfer learning analysis with only the English and the Turkish datasets. Interestingly, these do not outperform the previously described successful set-ups, i.e., TR-CH-4Q for English and 4Q-CH-TR for Turkish, which means that the Chinese dataset had its own generalisation role as a second train set

for the model. However, these new results come second in terms of evaluation scores from all configurations with the two sets, reiterating a real similarity between them, just like previously depicted. This comes partly in contrast to our context-based assumptions, but we should keep in mind that this is still transfer learning, which means that only several features are common in the end.

5.4 Conclusions

In conclusion, it was observed that the best results from all our experiments were obtained with the Turkish dataset, with 75% in f-score for transfer learning with SincNet, but also with consistent better scores in comparison to the other datasets. With the English set, some results were also relatively good, but the limitations are thought to origin in its variability and reliability in terms of music, annotations, but also ground-truth data. The Chinese dataset performed worst with many results falling below the chance score of 25% because of several possible reasons like association with lyrics, homogeneous data or ground-truth data. We believe that one limitation for our research was the fact that the ground-truth was too general. Other studies emphasized that continuous values along different dimensions or smoother gradients between more specific tags could be more effective in reaching annotator agreement and reliable datasets.

From experimenting with the baseline model based on handcrafted features, we found that different subsets of features were best for music in different languages, therefore each culture might have its own emotion discriminators. English and Turkish were found to be more similar on this aspect. A feature set comparison between low-level descriptors extracted with Essentia and the IS13 ComParE set was conducted and, while the established baseline model, the Multi-Layer Perceptron, used the later, preliminary examinations also show that the combination of the two feature sets might be beneficial.

A common limitation of several MER models that we also identified in our experiments was the confusion between the third and the fourth quadrants, that is the

valence of the low arousal music, especially with Western music. On the other hand, the first and second quadrants were generally well classified for English and Turkish data.

In terms of the proposed SincNet architecture, it does not outperform the established baseline for none of the three datasets that were considered under the within-dataset set-up. This means that the model was not able to extract as much emotionally relevant information from language and culture as we expected. One could argue that the model was not trained with enough data in order to make its results comparable to the traditional ML approach, keeping in mind that deep neural networks and especially end-to-end architectures are very data-thirsty. In addition, better tuning of the model is also more than welcome, for instance the consideration of larger frame sizes, since the architecture was initially created for a different task.

Other observations that we were able to make were that for both SincNet and baseline, within-dataset experiments outperform cross-dataset experiments and also, mixed training was not very successful either since cultural biases do exist. Although these set-ups were usually not as accurate as within-dataset training and testing, results with transfer learning proved to be more competitive. When all sets were employed but in specific orders, evaluating on the last set was a very successful SincNet set-up. While state-of-the-art models were not outperformed with this configuration either, results with the Turkish dataset are very promising. This is because the end-to-end SincNet model was able to achieve the same accuracy result as a more complex pre-trained CNN where chroma spectrograms were also extracted.

Therefore, our most important conclusion is that all cultures contributed to learning in a good way with transfer learning and in order to achieve more sensitive MER systems, we can establish a universal pre-trained model that we further fine-tune with the specific considerations of the target set. In this way, our work stands for a context-based fine-tuning of a general and culturally diverse system.

Our contribution to the MER field was significant through the proposal of one of the first end-to-end systems that learns to extract emotions from songs. As a first

attempt of the task with such a novel model, it is understandable that it is not very well performing in all cases, but more investigations could be conducted in the future. In addition, we also provided a unique comparison between models trained on music with lyrics in three different languages, in the idea that MER systems could possibly be improved by cultural-specific considerations.

5.5 Further Work

The MER problem was not solved, therefore there are endless options to extend our work. Further work firstly includes fixing the data issues in what concerns annotations, size of dataset, but also ground-truth considerations and reliability. Balancing of test sets and the mixed training dataset used with the SincNet configuration could also be considered. In terms of balancing, equal amounts of time should also be measured and preserved for different quadrants. Perhaps data augmentation would be a good solution in order to reach a more appropriate dataset size for end-to-end learning, especially with the Chinese and Turkish sets. However, for the case of the Chinese dataset, serious investigations should be conducted to verify the quality of the annotations and perhaps also experiment with segments of bigger sizes like 15 or even 30 seconds.

In terms of improvements on our feature extractors and analysis, one could extract more that the current 84 low-level features and statistics from Essentia, as it contains other types of features like tonal or rhythm descriptors. Given that many features in the combined features experiment belong to Essentia suggests that these are quite promising. At the same time, other feature selection algorithms could be employed, but this should not have too much of an impact. Another extension would be to extend and further analyze the combination of the two feature sets we compared in this study, those extracted with Essentia and the IS13 ComParE set, in the attempt to let the selector algorithm keep what is relevant from each.

In terms of the SincNet architecture we proposed for the task of Music Emotion Recognition, we should explore more in-depth modifications than just playing with

the optimization parameters. One relevant consideration would be to make the 200 millisecond chunks bigger, to one or a few seconds. Our preliminary results with the 500 millisecond frame size indicate that such lines of investigation are feasible. Most MER models use music with 30-second-long segments to be tagged, so training a model that only processes frames of 200 milliseconds might be biasing the system at the moment. Also, in the event of some modifications to the present architecture, more epochs and perhaps other batch sizes could also be tried.

A further comparison that could be made in order to determine the effectiveness of SincNet on the task is by also establishing a baseline model made of a typical CNN, that can take as input either the same waveform or even the corresponding spectrogram. In addition, since we saw that transfer learning works better than other set-ups made with SincNet, it would also be interesting to observe and compare our results to a transfer learning approach on the baseline model. This option is open for both the traditional machine learning baseline or the suggested CNN baseline.

In terms of cross-cultural extensions, more research could be conducted if other music emotion datasets are created or found for other languages than the ones considered so far, with the goal of approving or disproving our conclusions and inferences made here. A novel Indonesian dataset for music emotion recognition [62] was recently created and this would be an interesting study as the Indonesian culture is also Asian and some similarities and some differences might be observed. As transfer learning proved to be helpful in our case, an eventual goal would be to pre-train a model on a universal dataset, made of emotionally annotated music from as many cultures as possible, that will be further fine-tuned on the desired target cultures. Such a study could actually evaluate our results and draw some useful conclusions.

Therefore, many possibilities exist, the research on Music Emotion Recognition is relatively young and many questions arise every day. We managed here to maybe answer some of them or confirm some previous responses, but we also raised some new ones that are to be investigated in future work.

List of Figures

Bibliography

[1] Holzapfel, A., Sturm, B. & Coeckelbergh, M. Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval* **1**, 44–55 (2018).

[2] Ravanelli, M. & Bengio, Y. Speaker recognition from raw waveform with sincnet (2018).

[3] Juslin, P. Music and emotion: Seven questions, seven answers. *Music and the mind: Essays in honour of John Sloboda* 113–35 (2011).

[4] Gabrielsson, A. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae* **5**, 123–147 (2002).

[5] Grekow, J. *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces* (2018).

[6] Hu, X., Downie, J., Laurier, C., Bay, M. & Ehmann, A. The 2007 mirex audio mood classification task: Lessons learned. 462–467 (2008).

[7] Barrett, L. F. *How emotions are made: The secret life of the brain* (Houghton Mifflin Harcourt, 2017).

[8] Russell, J. A circumplex model of affect. *Journal of Personality and Social Psychology* **39**, 1161–1178 (1980).

[9] Yang, X., Dong, Y. & Li, J. Review of data features-based music emotion recognition methods. *Multimedia Systems* **24**, 1–25 (2017).

[10] Hu, X. & Yang, Y.-H. Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs. *IEEE Transactions on Affective Computing* **8**, 1–1 (2016).

[11] Zangeneh Soroush, M., Maghooli, K., Setarehdan, K. & Motie Nasrabadi, A. Emotion classification through nonlinear eeg analysis using machine learning methods. *International Clinical Neuroscience Journal* **5**, 135–149 (2018).

[12] Pons, J. *et al.* End-to-end learning for music audio tagging at scale (2017).

[13] Panda, R., Malheiro, R. & Paiva, R. P. Musical texture and expressivity features for music emotion recognition (2018).

[14] Warriner, A., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods* **45** (2013).

[15] Aljanaki, A., Yang, y.-h. & Soleymani, M. Developing a benchmark for emotional analysis of music. *PLOS ONE* **12**, e0173392 (2017).

[16] Soleymani, M., Caro, M., Schmidt, E., Sha, C.-Y. & Yang, y.-h. 1000 songs for emotional analysis of music. 1–6 (2013).

[17] Yang, y.-h. & Chen, H. Ranking-based emotion recognition for music organization and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on* **19**, 762 – 774 (2011).

[18] Eerola, T. & Vuoskoski, J. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* (2011).

[19] Schuller, B., Dorfner, J. & Rigoll, G. Determination of non-prototypical valence and arousal in popular music: Features and performances. *EURASIP Journal on Audio, Speech, and Music Processing, Special Issue on Scalable Audio-Content Analysis* **2010** (2020).

[20] Aljanaki, A., Wiering, F. & Veltkamp, R. Studying emotion induced by music through a crowdsourcing game. *Information Processing Management* **76** (2015).

[21] Juslin, P. & Laukka, P. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research* **33**, 217–238 (2004).

[22] Yang, y.-h., Lin, Y.-C., Su, Y.-F. & Chen, H. A regression approach to music emotion recognition. *Audio, Speech, and Language Processing, IEEE Transactions on* **16**, 448 – 457 (2008).

[23] Mcadams, S. & Giordano, B. *The Oxford handbook of music psychology*, 72–80 (2009).

[24] Aljanaki, A. & Soleymani, M. A data-driven approach to mid-level perceptual musical feature modeling (2018).

[25] Schuller, B. *et al.* Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge. *Computer Speech Language* **53** (2018).

[26] Eyben, F., Wöllmer, M. & Schuller, B. opensmile – the munich versatile and fast open-source audio feature extractor. 1459–1462 (2010).

[27] Martín-Gómez, L. & Cáceres, M. Applying data mining for sentiment analysis in music. 198–205 (2018).

[28] Suresh, R. & A, S. Different machine learning classifiers for music emotion recognition. *International Journal of Recent Technology and Engineering (IJRTE)* **8** (2019).

[29] Mahapatra, S. Why deep learning over traditional machine learning? (2019). URL `https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063#:~:text=IntraditionalMachinelearningtechniques,tolearningalgorithmstowork.&text=Usually,aDeepLearningalgorithm,tolargenumberofparameters.` Accessed: 2020-05-01.

[30] Bertin-Mahieux, T., Ellis, D., Whitman, B. & Lamere, P. The million song dataset. 591–596 (2011).

[31] Liu, X., Chen, Q., Wu, X., Liu, Y. & Liu, Y. Cnn based music emotion classification (2017).

[32] Liu, T., Han, L., Ma, L. & Guo, D. Audio-based deep music emotion recognition. vol. 1967, 040021 (2018).

[33] Schmidt, E. & Kim, Y. Learning emotion-based acoustic features with deep belief networks. 65–68 (2011).

[34] Chen, S.-H., Lee, Y.-S., Hsieh, W.-C. & Wang, J.-C. Music emotion recognition using deep gaussian process. 495–498 (2015).

[35] Chowdhury, S., Vall, A., Haunschmid, V. & Widmer, G. Towards explainable music emotion recognition: The route via mid-level features (2019).

[36] Haunschmid, V., Chowdhury, S. & Widmer, G. Two-level explanations in music emotion recognition (2019).

[37] Er, M. & Aydilek, Music emotion recognition by using chroma spectrogram and deep visual features. *International Journal of Computational Intelligence Systems* **12** (2019).

[38] Orjesek, R., Jarina, R., Chmulik, M. & Kuba, M. Dnn based music emotion recognition from raw audio signal. In *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 1–4 (2019).

[39] Malík, M. *et al.* Stacked convolutional and recurrent neural networks for music emotion recognition (2017).

[40] Dong, Y., Yang, X., Zhao, X. & Li, J. Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for music emotion recognition. *IEEE Transactions on Multimedia* **PP**, 1–1 (2019).

[41] Speck, J., Schmidt, E., Morton, B. & Kim, Y. A comparative study of collaborative vs. traditional musical mood annotation. 549–554 (2011).

[42] Bowling, D., Purves, D. & Gill, K. Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences* **115**, 201713206 (2017).

[43] Weninger, F., Eyben, F., Schuller, B., Mortillaro, M. & Scherer, K. On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in psychology* **4**, 292 (2013).

[44] Gómez-Cañón, J. S., Cano, E., Herrera, P. & Gómez, E. Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification. In *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)* (2020).

[45] Coutinho, E. & Schuller, B. Shared acoustic codes underlie emotional communication in music and speech—evidence from deep transfer learning. *PLoS ONE* **12** (2017).

[46] 0001, X. H., Lee, J. H., Choi, K. & Downie, J. S. A cross-cultural study on the mood of k-pop songs. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 385–390 (ISMIR, Taipei, Taiwan, 2014). URL `https://doi.org/10.5281/zenodo.1417993`.

[47] Sangnark, S., Lertwatechakul, M. & Benjangkaprasert, C. Thai music emotion recognition based on western music. *Journal of Physics: Conference Series* **1195**, 012009 (2019).

[48] Chen, Y.-W., Yang, y.-h. & Chen, H. Cross-cultural music emotion recognition by adversarial discriminative domain adaptation. 467–472 (2018).

[49] Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J. & Moussallam, M. Music mood detection based on audio and lyrics with deep neural net (2018).

[50] Zhou, J., Xiaoou, C. & Yang, D. *Multimodel Music Emotion Recognition Using Unsupervised Deep Neural Networks*, 27–39 (2019).

[51] Umamaheswari, J. & Akila, A. An enhanced human speech emotion recognition using hybrid of prnn and knn. 177–183 (2019).

[52] Trigeorgis, G. *et al.* Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network (2016).

[53] Zeng, H. *et al.* Eeg emotion classification using an improved sincnet-based deep learning model. *Brain Sciences* **9**, 326 (2019).

[54] Panda, R., Malheiro, R. & Paiva, P., Rui. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing* **PP**, 1–1 (2018).

[55] Bogdanov, D. *et al.* Essentia: an audio analysis library for music information retrieval (2013).

[56] Porter, A., Bogdanov, D. & Serra, X. Mining metadata from the web for acousticbrainz. 53–56 (2016).

[57] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[58] Jagiello, R., Pomper, U., Yoneya, M. & et al. Rapid brain responses to familiar vs. unfamiliar music – an eeg and pupillometry study. *Scientific Reports* **9** (2019).

[59] Cowen, A., Fang, X., Sauter, D. & Keltner, D. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences* **117**, 201910704 (2020).

[60] Yang, y.-h. & Chen, H. Ranking-based emotion recognition for music organization and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on* **19**, 762 – 774 (2011).

[61] Hult, S., Kreiberg, L., Brandt, S. & Jónsson, B. Analysis of the effect of dataset construction methodology on transferability of music emotion recognition models. 316–320 (2020).

[62] Lasiman, J. & Lestari, D. Speech emotion recognition for indonesian language using long short-term memory. 40–43 (2018).

Appendix A

Further Results

We provide here some extra comparisons, scores and confusion matrices that support our study. Note that some of the confusion matrices reported in this chapter come as triplets, where the first one belongs to the English test set, the middle one to the Chinese set and the rightmost to the Turkish set, unless otherwise stated.

English	Chinese	Turkish
barkbands_spread.mean	silence_rate_20dB.stdev	silence_rate_30dB.stdev
spectral_complexity.mean	silence_rate_60dB.stdev	spectral_centroid.mean
spectral_centroid.mean	hfc.stdev	spectral_skewness.stdev
spectral_energyband_high.mean	spectral_energyband_high.stdev	melbands_spread.mean
spectral_complexity.stdev	silence_rate_60dB.mean	spectral_kurtosis.stdev
spectral_rolloff.mean	silence_rate_30dB.stdev	silence_rate_60dB.stdev
melbands_spread.mean	spectral_energyband_high.mean	barkbands_spread.mean
spectral_skewness.stdev	hfc.mean	zerocrossingrate.mean
zerocrossingrate.mean	pitch_salience.stdev	barkbands_flatness_db.mean
Dissonance.stdev	spectral_energyband_low.mean	spectral_kurtosis.mean

Figure 17: Best 10 features selected from the Essentia feature set. Highlighted with the same colour are similar features found relevant with different datasets.

English	Chinese	Turkish
logHNR_sma_stddev	silence_rate_20dB.stdev	logHNR_sma_stddev
barkbands_spread.mean	silence_rate_60dB.stdev	spectral_centroid.mean
spectral_complexity.mean	logHNR_sma_de_stddev	spectral_skewness.stdev
spectral_centroid.mean	silence_rate_30dB.stdev	spectral_kurtosis.stdev
spectral_energyband_high.mean	silence_rate_60dB.mean	melbands_spread.mean
pcm_fftMag_spectralRollOff90.0_sma_amean	spectral_kurtosis.stdev	barkbands_spread.mean
spectral_complexity.stdev	hfc.stdev	silence_rate_30dB.stdev
pcm_fftMag_spectralVariance_sma_amean	spectral_energyband_high.stdev	pcm_fftMag_spectralCentroid_sma_amean
spectral_rolloff.mean	hfc.mean	barkbands_flatness_db.mean
Melbands_spread.mean	pitch_salience.stdev	silence_rate_60dB.stdev

Figure 18: Best 10 features selected from the combination of Essentia with IS13 ComParE feature sets. Highlighted are the features belonging to IS13 ComParE.

	Essentia			IS13 ComParE		
English	1. MLP 68%	2. RF 68%	3. KNN 66%	**1. MLP 71%**	2. RF 70%	3. KNN 66%
Chinese	1. **KNN 49%**	2. SVM 48%	3. RF 47%	1. KNN 45%	2. RF 43%	3. MLP 42%
Turkish	1. **GP 89%**	2. MLP 89%	3. RF 84%	1. MLP 88%	2. RF 86%	3. KNN 85%

Figure 19: Best 3 algorithms for each feature set and each language, in terms of accuracy, when the 10-fold cross-validation was employed. In bold are the best scores and underlined is the model used as baseline.

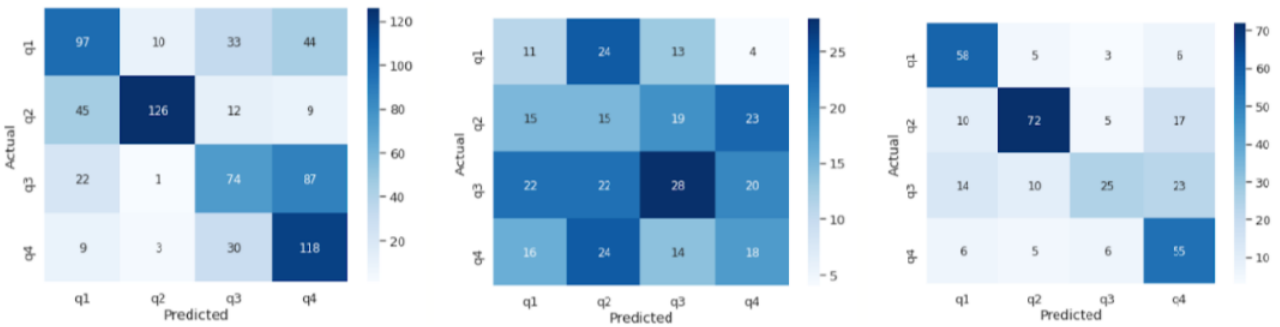

Figure 20: Confusion matrices for best Essentia set-up: the Random Forest Classifier.

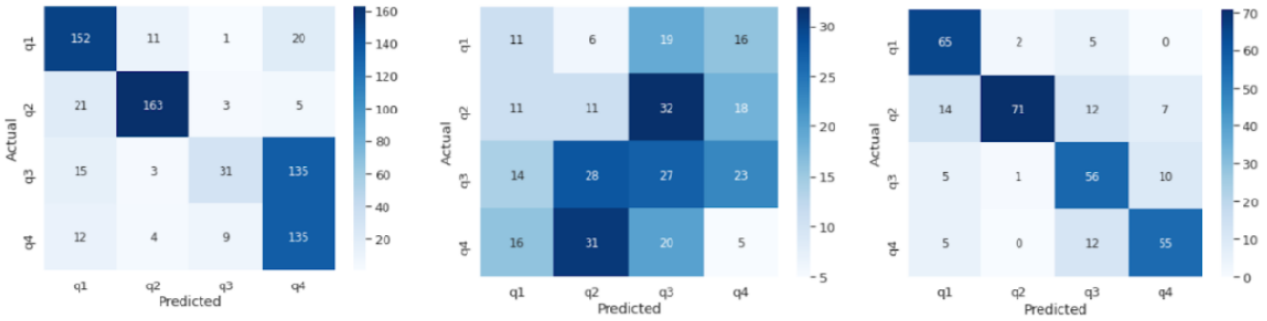

Figure 21: Confusion matrices for MLP on combined feature sets.

Features	Essentia				IS13 ComParE				Combined			
Score	P	R	F	A	P	R	F	A	P	R	F	A
4Q-EMOTION	58	56	55	56	66	64	63	64	72	67	64	67
CH-818	26	24	23	24	26	31	24	31	18	19	18	19
TR-MUSIC	71	70	70	70	74	71	72	71	80	77	77	77

Figure 22: Results of within-dataset with the 3 feature sets considered. Highlighted are the best results for each of the three datasets.

Dataset	4Q-EMOTION				CH-818				TR-MUSIC			
Model \ Score	P	R	F	A	P	R	F	A	P	R	F	A
Baseline MLP	65%	63%	62%	63%	23%	30%	23%	30%	74%	71%	71%	71%
SincNet 200ms	59%	57%	52%	57%	11%	27%	16%	27%	68%	63%	58%	63%
SincNet 500ms	58%	58%	56%	58%	3%	18%	6%	18%	68%	66%	66%	66%

Figure 23: Comparison of results with SincNet with 500ms frame sizes instead of original 200ms (batch size 32 instead of 128). The baseline achives best results (higlighted in pink). In yellow are the highest scores between the 2 SincNet set-ups

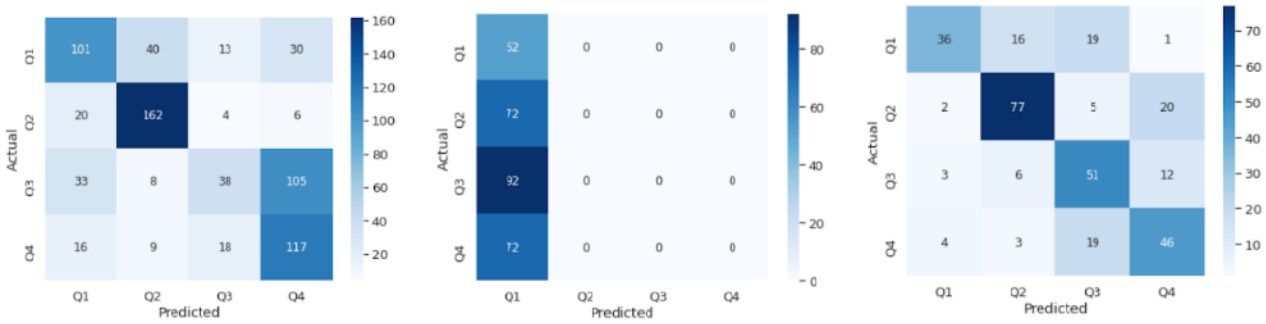

Figure 24: Confusion matrices for SincNet with 500ms frame sizes.

Train\Test	English				Chinese				Turkish			
Score(%)	P	R	F	A	P	R	F	A	P	R	F	A
English	**60**	**55**	**54**	**55**	**28**	28	**23**	28	55	47	45	48
Chinese	31	37	29	37	22	20	20	20	23	22	21	22
Turkish	55	52	50	52	19	**29**	22	**29**	**72**	**69**	**70**	**69**

Figure 25: Cross-dataset results with Essentia and MLP. In bold are the highest scores obtained for each test set.

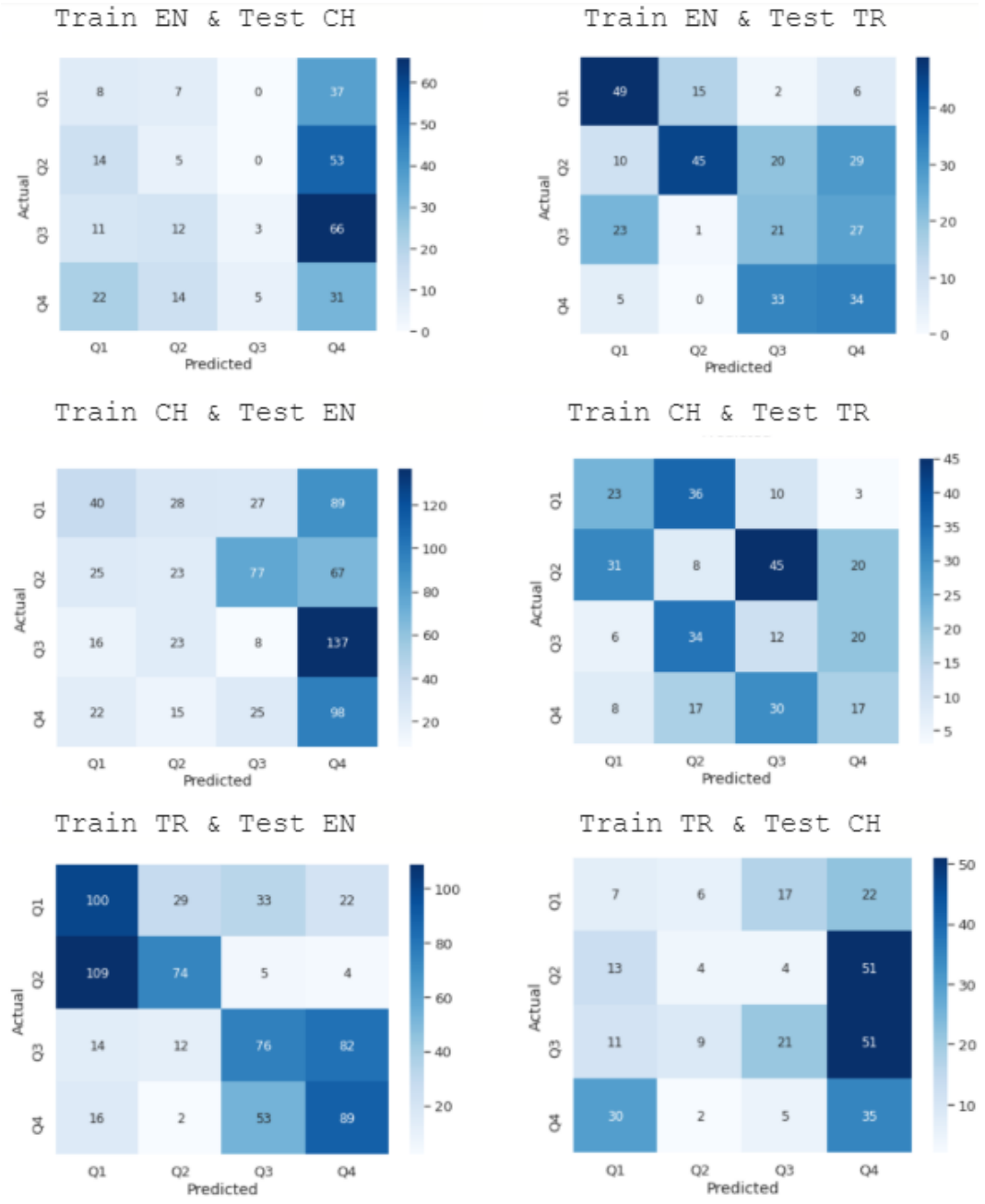

Figure 26: Confusion matrices for cross-dataset experiments with the baseline model.

Essentia	IS13ComParE
spectral_complexity.stdev	logHNR_sma_stddev
spectral_complexity.mean	pcm_fftMag_spectralRollOff90.0_sma_amean
spectral_centroid.mean	audspec_lengthL1norm_sma_de_stddev
barkbands_spread.mean	pcm_fftMag_spectralFlux_sma_amean
spectral_energyband_high.mean	pcm_fftMag_spectralVariance_sma_amean
spectral_skewness.stdev	pcm_fftMag_spectralCentroid_sma_amean
melbands_spread.mean	audspec_lengthL1norm_sma_amean
zerocrossingrate.mean	mfcc_sma[1]_amean
barkbands_flatness_db.mean	F0final_sma_stddev
silence_rate_30dB.stdev	F0final_sma_amean

Figure 27: Best 10 features for the mixed training set-up.

Test\Features	Essentia				IS13ComParE			
Score(%)	P	R	F	A	P	R	F	A
English	59	**57**	**57**	**57**	**60**	**57**	56	**57**
Chinese	**25**	**28**	**23**	**28**	22	23	22	23
Turkish	**70**	**66**	**67**	**66**	67	64	64	64

Figure 28: Comparison of mixed training results with the 2 feature sets. In bold are the higher scores for each culture.

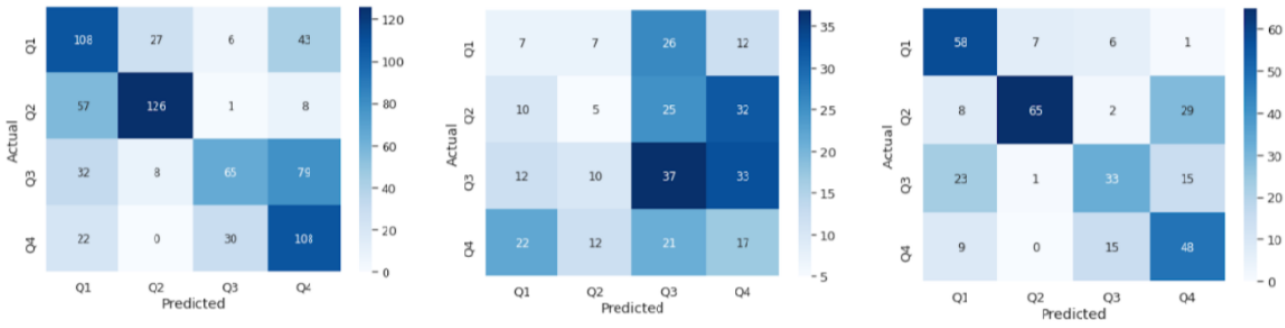

Figure 29: Confusion matrices for mixed training and individual testing with the baseline model.

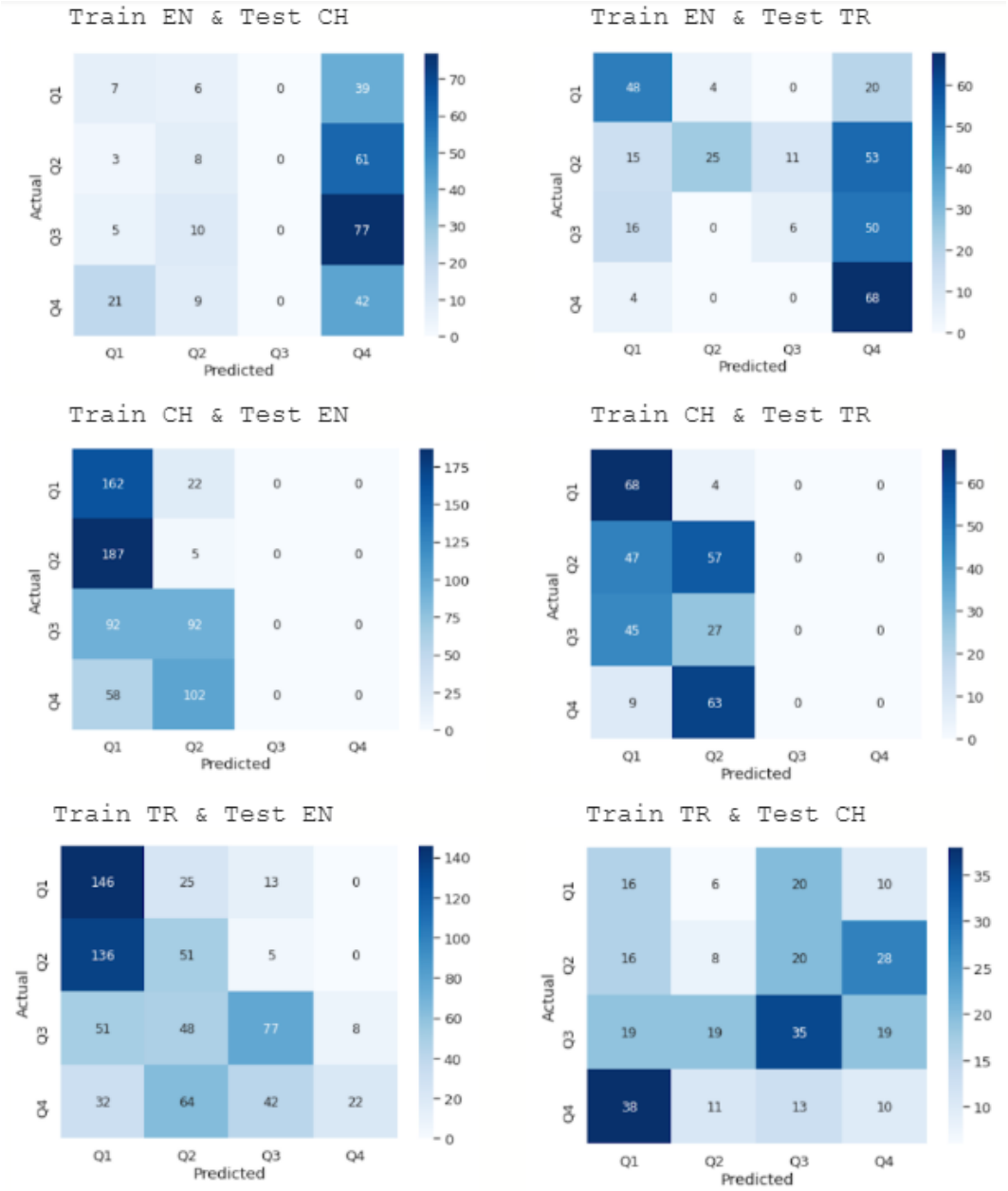

Figure 30: Confusion matrices for cross-dataset experiments with the end-to-end model.

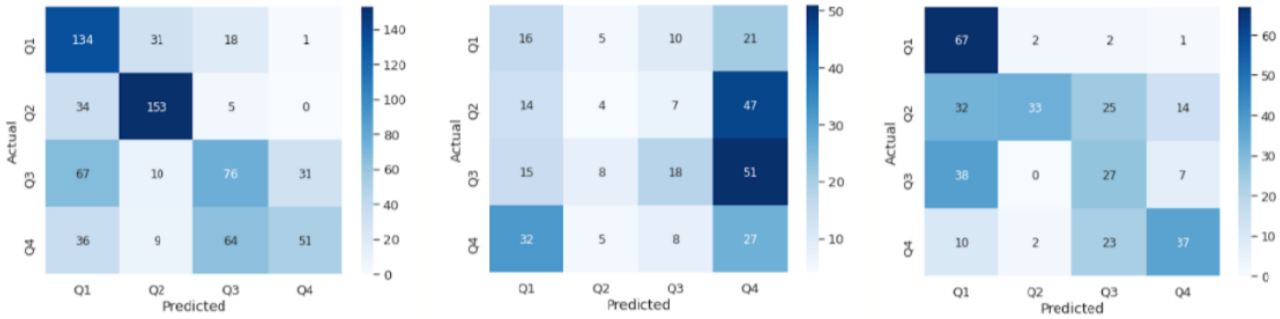

Figure 31: Confusion matrices for mixed training and individual testing with the end-to-end model.

Model \ Test	English (4Q)				Chinese (CH)				Turkish (TR)			
Score (%)	P	R	F	A	P	R	F	A	P	R	F	A
4Q-CH-TR	52	51	50	51	21	19	16	19	75	75	75	75
4Q-TR-CH	9	14	11	14	10	24	12	24	19	33	20	33
CH-4Q-TR	52	50	49	50	24	20	17	20	72	71	71	71
CH-TR-4Q	57	56	51	56	15	20	15	20	47	45	36	45
TR-4Q-CH	26	29	24	29	28	26	17	26	37	35	28	35
TR-CH-4Q	61	60	57	60	14	19	14	19	51	46	40	46

Figure 32: Transfer learning evaluations for all test sets. In pink we observe the results where the fine-tuning was made on music in the same language as the test set.

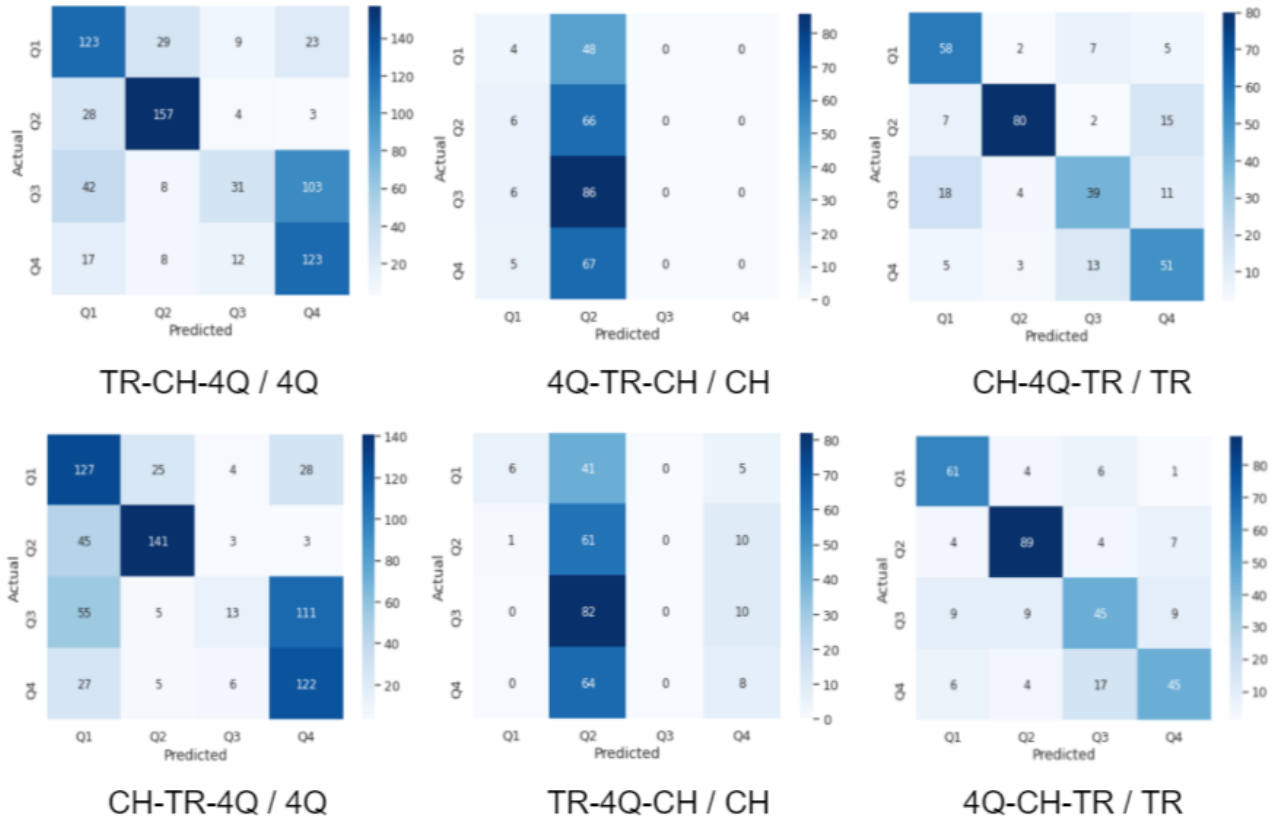

Figure 33: Confusion matrices for transfer learning evaluated on the target sets.

Model \ Score	P	R	F	A
4Q / 4Q	59%	57%	52%	57%
CH-TR-4Q / 4Q	57%	56%	51%	56%
TR-CH-4Q / 4Q	**61%**	**60%**	**57%**	**60%**
TR / 4Q	59%	58%	56%	58%
TR / TR	68%	63%	58%	63%
4Q-CH-TR / TR	**75%**	**75%**	**75%**	**75%**
CH-4Q-TR / TR	72%	71%	71%	71%
4Q / TR	73%	73%	73%	73%

Figure 34: Transfer learning comparisons on SincNet with only English and Turkish. Best results remain the previous ones, but the two new set-ups come close behind.

Appendix B

Précis for AES ML Symposium 2020

Music Emotion Recognition has grown to be an important part of the Music Information Retrieval field, with applications in music recommendation and search, playlist creation, but also in therapy and marketing. Because the topic of emotion is very ambiguous and subjective, there are many challenges to be defeated, including the good quality of annotations, the confusion between felt and perceived emotions, but also taking advantage of the extra-musical information, for example culture. Various pre-extracted feature sets were proposed in order to leverage these issues, as well as various deep learning architectures that directly retrieve information from spectrograms.

Music emotion recognition has not been approached yet with the end-to-end learning strategy that encodes and learns from the raw waveform input, therefore we propose it here as one of our novel contributions. In this way, we classify music segments under a mixture of dimensional and categorical taxonomy, in one of the four quadrants of the Valence-Arousal plane. Moreover, we aim to investigate the relevance of cultural and dataset specific characteristics by conducting experiments with music in three different languages, English (4Q-EMOTION dataset), Mandarin (CH-818 dataset) and Turkish (TR-MUSIC dataset). Our results show that the three datasets indeed behave differently throughout cross-dataset experiments, suggesting a need for context-based models.

At first, we examined several traditional machine learning algorithms and feature sets in order to establish a proper baseline model for the end-to-end proposal. We found that the most accurate results were generally obtained with the Multi-Layer Perceptron, built with Scikit-Learn, using the IS13 ComParE feature set, a well-evolved feature set for automatic recognition of audio emotion.

The architecture proposed for the end-to-end learning is called SincNet and was originally designed for speaker recognition. We hypothesize that the speaker cues detected by this architecture are prone to play a role in Music Emotion Recognition with language considerations. The outstanding feature of this architecture is its first convolutional layer based on sinc functions that implement rectangular band-pass filters. Preliminary results on the three datasets do not outperform the baseline model, but the end-to-end approach remains promising with further fine-tuning.

Moreover, we conducted two cross-dataset experiments with both models, our baseline and SincNet, where similar outcomes were observed. In the first set-up, models are trained and tested with different datasets and, as expected, features learned from music with lyrics in one language are not completely transferable to another. The second set-up aims to provide a big multicultural training dataset under the assumption that more data should improve the MER algorithm performance. Preliminary results show that when testing with individual datasets, classification is noisier when different language datasets are combined, therefore suggesting that cultural considerations might be relevant. However, it seems that by pre-training SincNet with other languages and tuning the model on the desired language dataset, the obtained results are better than the baseline.

In conclusion, we show promising ideas and evidence that MER could be improved with end-to-end models, providing that enough training data is available, as well as with cultural considerations and context-based models.

Appendix C

Extended Abstract & Poster for ISMIR 2020

Cross-Dataset Music Emotion Recognition: an End-to-End Approach

Ana Gabriela Pandrea **Juan Sebastián Gómez-Cañón** **Perfecto Herrera**
Music Technology Group, UPF, Barcelona, Spain

ABSTRACT

The topic of Music Emotion Recognition (MER) evolved as music is a fascinating expression of emotions, yet it faces challenges given its subjectivity. Because each language has its particularities in terms of sound and intonation, and implicitly associations made upon them, we hypothesize perceived emotions might vary in different cultures. To address this issue, we test a novel approach towards emotion detection and propose a language sensitive end-to-end model that learns to tag emotions from music with lyrics in English, Mandarin and Turkish.

1. INTRODUCTION

Music Emotion Recognition has become an important part of the Music Information Retrieval field, with applications in music recommendation and search, playlist creation, but also in therapy and marketing. Because the topic of emotion is very ambiguous and subjective, there are many challenges to overcome: improving the quality of annotations, the confusion between felt and perceived emotions, and taking advantage of the extra-musical information, e.g., culture. Various pre-extracted feature sets were proposed in order to leverage these issues [1], as well as various deep learning architectures that directly retrieve information from spectrograms [2].

2. METHODS

Music emotion recognition has started to be explored with the end-to-end learning strategy that encodes and learns from the raw waveform input, therefore we propose it here as a novel contribution. In this way, we classify excerpts under a mixture of dimensional and categorical taxonomy, in one of the four quadrants of the Russell's Valence-Arousal plane [3]. Moreover, we aim to investigate the relevance of cultural and dataset specific characteristics by conducting experiments with music in three different languages, English - 4Q-EMOTION dataset [4], Mandarin - CH-818 dataset [5] and Turkish - TR-MUSIC dataset [6].

At first, we examined several traditional machine learning algorithms and feature sets in order to establish a proper baseline model for the end-to-end proposal. From several classifiers built with Scikit-Learn [1] (K-Nearest-Neighbors, Support Vector Machine with linear kernel, Support Vector Machine with Radial Basis Function, Gaussian Process, Multi-Layer Perceptron, Gaussian Naive Bayes, Random Forest), we found that the most accurate results were generally obtained with the Multi-Layer Perceptron (MLP), having one hidden layer made of 100 neurons. We also compared the use of low-level descriptors extracted with Essentia [7] to the IS13 ComParE feature set [8], a well-evolved feature set for automatic recognition of audio emotion. The latter performed better thus it was used for the baseline model.

The end-to-end architecture is called SincNet [9], a Convolutional Neural Network originally designed for speaker recognition. We hypothesize that the speaker cues detected by this architecture are prone to play a role in Music Emotion Recognition with language considerations. The outstanding feature of this architecture is its first convolutional layer based on Sinc functions that implement rectangular band-pass filters [2].

3. RESULTS

We conducted 3 main experiments with both the baseline model and the end-to-end model, followed by a fourth only on SincNet. The first is a within-dataset evaluation, i.e., the models were trained and tested with music in the same language independently for each of the 3 sets, confirming research by [5]. While none of our models manages to learn from the Mandarin set, for English and Turkish several similarities were observed with the baseline model in terms of the best selected features and the common confusions between quadrants. We also observed that SincNet under this configuration does not outperform neither the baseline nor the state-of-the-art MER models.

Secondly, cross-dataset evaluations were performed, i.e., the model was trained with music from one culture and tested with music from another. As expected, results cross-dataset are worse than within-dataset with both MLP and SincNet. Thirdly, a mixed dataset configuration was employed under the assumption that training with more data should aim at better results if the learning is not language specific. Results do not improve under this consideration and training a single model with music from different languages only worsens the performance metrics of the MER system. Finally, with SincNet we also considered a trans-

[1] https://scikit-learn.org/

[2] We refer the reader to [9] for more information on the architecture.

			Baseline: MLP				End-to-end: SincNet			
Configuration	**Dataset (train / test)**	**S**	**A**	**P**	**R**	**F**	**A**	**P**	**R**	**F**
Within dataset (English)	4Q / 4Q *	720	0.63	0.65	0.63	0.63	0.57	0.59	0.57	0.52
TL 1 - finetune 4Q	CH-TR-4Q / 4Q	720	-	-	-	-	0.56	0.57	0.56	0.51
TL 2 - finetune 4Q	TR-CH-4Q / 4Q **	720	-	-	-	-	0.60	0.61	0.60	0.57
Mixed dataset	ALL / 4Q	720	0.57	0.60	0.57	0.56	0.57	0.58	0.57	0.56
Within dataset (Chinese)	CH / CH *	288	0.30	0.23	0.30	0.23	0.27	0.11	0.27	0.16
TL 1 - finetune CH	4Q-TR-CH / CH	288	-	-	-	-	0.24	0.10	0.24	0.12
TL 2 - finetune CH	TR-4Q-CH / CH	288	-	-	-	-	0.26	0.28	0.26	0.17
Mixed dataset	ALL / CH **	288	0.23	0.22	0.23	0.22	0.23	0.26	0.23	0.21
Within dataset (Turkish)	TR / TR	320	0.71	0.74	0.71	0.71	0.63	0.68	0.63	0.58
TL 1 - finetune TR	4Q-CH-TR / TR	320	-	-	-	-	0.75	0.75	0.75	0.75
TL 2 - finetune TR	CH-4Q-TR / TR * **	320	-	-	-	-	0.71	0.72	0.71	0.71
Mixed dataset	ALL / TR	320	0.64	0.67	0.64	0.64	0.51	0.61	0.51	0.50

Table 1. Summary of results from baseline and end-to-end models. S stands for samples, A for accuracy, P for precision, R for recall, and F for F-score. * stands for best overall results for each language, ** stands for best SincNet results.

fer learning approach, from one source culture to the second and fine-tuning on the third. This seemed to be the most promising set-up for this architecture, especially with the Turkish data, where one of the transfer learning set-ups gave the best results from all our experiments.

Furthermore, we extended this first set of experiments with 3 more set-ups that were decided based on these preliminary results:

- We combined both feature sets that were proposed and extracted the best 50 features from the whole set, among which there were slightly more from Essentia. This worked even better than IS13 ComParE, with increasing scores for both English (increase of 1 percent points in F-score) and Turkish music (increase of 5 percent points in F-score).
- As English and Turkish appeared to be more similar in terms of relevant features, we considered the transfer learning experiments with only these two sets. Although differences are small, it appears that when considering the Chinese set as an intermediary tuning set, scores improve.
- Since the initial SincNet configuration considered frames of only 200 milliseconds and this amount of time might be questionable for emotion depiction, we wanted to use a longer fragment. According to the available computing power, the highest window we tested was 500 milliseconds, with a smaller batch size of 32 samples. Results are comparable, with an increase of 4 percent points in F-score for English and 8 percent point for Turkish, suggesting that similar set-ups could be further explored.

4. DISCUSSION

We addressed the problem of automatic music emotion recognition with the end-to-end SincNet architecture. We found that our approach is limited with respect to the tuning of the network and the size and structure of the datasets as none of our results is sensitive enough. Our findings suggest that traditional methods remain the best choice w.r.t. both within- and mixed-dataset set-ups. The end-to-end architecture might be promising provided that it is better adapted to the task. Transfer learning and fine-tuning SincNet on the target language gives better results than SincNet within-dataset, suggesting that the architecture could be more suitable for with this approach and that more training data improves performance in this set-up. Future work in this area should consider larger fragments and adaptations of SincNet, and the extension of similar studies to other cultures and datasets.

5. REFERENCES

[1] Yang, Dong, and Li, "Review of data features-based MER methods," *Multimedia Systems*, 2017.

[2] Dong, Yang, Zhao, and Li, "Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for MER," *IEEE Trans. on multimedia*, 2019.

[3] Russell, "A circumplex model of affect," *J. Personal. Soc. Psychol.*, 1980.

[4] Panda, Malheiro, and Paiva, "Novel audio features for MER," *IEEE Trans. on Affective Computing*, 2018.

[5] Hu and Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," *IEEE Trans. on Affective Computing*, 2017.

[6] B. Er and B. Aydilek, "MER by using chroma spectrogram and deep visual features," *Intern. J. of Computational Intelligence Systems*, 2019.

[7] Bogdanov and et al., "Essentia: an audio analysis library for mir," *Intern. Soc. for MIR Conf.*, 2013.

[8] Schuller and et al., "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech Language*, 2018.

[9] Ravanelli and Bengio, "Speaker recognition from raw waveform with sincnet," *IEEE Spoken Language Technology Workshop*, 2018.

Cross-Dataset Music Emotion Recognition: an End-to-End Approach

Ana Gabriela Pandrea, Juan S. Gómez Cañón, Perfecto Herrera

ISMIR MTL2020

In this work, we address **music emotion recognition with a context-based end-to-end model**. Two main research questions are addressed:

- Are there differences/correlations between the emotions perceived in music by listeners raised with different mother tongues?
- Can an end-to-end model trained on raw waveforms, that was efficient for speech tasks, be employed for language-oriented MER?

Our hypothesis is that perceived emotion depends on cultural characteristics and models should be trained in the target test language in order to obtain sensitive results. The end-to-end architecture should outperform state-of-the-art models since it allows for more information to be processed.

Problem definition: music emotion recognition

- **Valence-Arousal plane** - 4 quadrants classification [1]
- **Raw waveform** vs. handcrafted features
- End-to-end deep learning: **SincNet** [2]
- Musical emotion complexity relates to **cultural-specific associations**

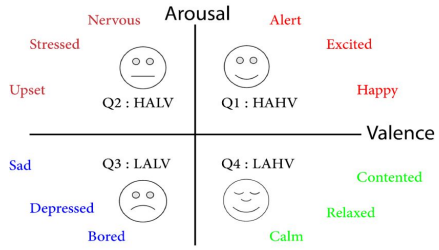

Fig 1. Valence-Arousal plane, taken from [3]. L is low, H is high, A is arousal, V is valence .

Methodology: data & models

- 3 languages: **English** [4], **Mandarin** [5], **Turkish** [6]
- Baseline feature sets: Essentia [7] & IS13 ComParE [8]
- Baseline Multi-Layer Perceptron (MLP) vs. End-to-end SincNet
- SincNet: CNN with **sinc filters** in the feature extractor
- Set-ups: **Within-dataset**, **Cross-dataset**, **Mixed training**, **Transfer Learning**
- Evaluation based on: weighted averages of precision (P), recall (R), f-score (F), accuracy (A) and confusion matrices

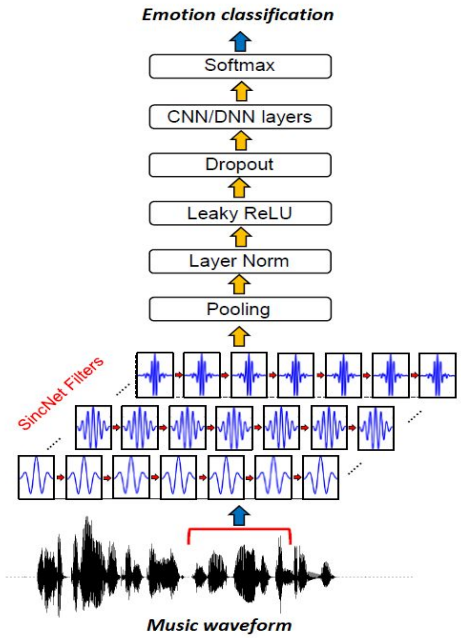

Fig 2. The SincNet architecture, adapted from [2]

Results: Baseline, SincNet, Cross-Dataset

- Baseline model: **MLP**, best from default traditional ML in Scikit-Learn [9]
- Baseline features: **IS13 ComParE**, with 65 low-level descriptors, statistics
- Best features were found to be more similar for English and Turkish
- Combined with Essentia => better scores for English, Turkish
- SincNet: 200ms frame size, 128 batch size, 100 epochs
- **Bigger frame size to better detect emotion**: 500ms frame size, 32 batch size, 100 epochs => better f-score and accuracy for English, Turkish
- Cross-dataset: train & test with different sets => general poor results
- Transfer learning 1* & 2*: **tune SincNet with the 3 sets in specific orders**: CH-TR-EN & TR-CH-EN for English, EN-TR-CH & TR-EN-CH for Mandarin, EN-CH-TR & CH-EN-TR for Turkish
- Transfer learning 3*: train with EN / TR & test with TR / EN => scores don't necessarily increase, thus CH also contributes to transfer learning
- Mixed dataset: train on combined datasets & test separately => **within-dataset is better, thus mixing brings noise**

Dataset		English (4Q-EMOTION)				Mandarin (CH-818)				Turkish (TR-MUSIC)			
Set-up	**Model**	**P**	**R**	**F**	**A**	**P**	**R**	**F**	**A**	**P**	**R**	**F**	**A**
Within dataset	MLP Essentia	58	56	55	56	26	24	**23**	24	71	70	70	70
Within dataset	MLP IS13 ComParE	65	63	63	63	23	**30**	**23**	**30**	74	71	71	71
Within dataset	MLP combined	**72**	**67**	**64**	**67**	18	19	18	19	**80**	**77**	**77**	**77**
Within dataset	SincNet 200ms	59	57	52	57	11	***27***	16	***27***	68	63	58	63
Within dataset	SincNet 500ms	58	58	56	58	3	18	6	18	68	66	66	66
Transfer learning 1*	SincNet 200ms	57	56	51	56	10	24	12	24	***75***	***75***	***75***	***75***
Transfer learning 2*	SincNet 200ms	***61***	***60***	***57***	***61***	***28***	26	17	26	72	71	71	71
Transfer learning 3*	SincNet 200ms	59	58	56	58	-	-	-	-	73	73	73	73
Mixed dataset	MLP IS13 ComParE	60	57	56	57	22	23	22	23	67	64	64	64
Mixed dataset	SincNet 200ms	58	57	56	57	26	23	***21***	23	61	51	50	51

Fig 3. Summary of results. Underlined are best scores for each dataset from all training set-ups, in italics are best scores with SincNet. *Transfer learning set-ups as described above.

- **Conclusions:** end-to-end models need more data and better fine-tuning & general cues are learned from different cultures, but **sensitivity is achieved with context-based fine-tuning**
- **Future work**: continuous valence and arousal values, better data curation, data augmentation, bigger chunk and batch sizes, fine-tuning SincNet, exploring other cultures

References

[1] Russell, "A circumplex model of affect," J. Personal. Soc. Psychol., 1980.
[2] Ravanelli and Bengio, "Speaker recognition from raw waveform with sincnet," IEEE Spoken Language Technology Workshop, 2018
[3] Soroush et al., "Emotion Classification through Nonlinear EEG Analysis Using Machine Learning Methods", Intern. Clinical Neuroscience Journal, 2018.
[4] Panda, Malheiro and Paiva, "Novel audio features for MER," IEEE Trans. on Affective Computing, 2018.
[5] Hu and Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," IEEE Trans. on Affective Computing, 2017.
[6] B. Er and B. Aydilek, "MER by using chroma spectrogram and deep visual features," Intern. J. of Computational Intelligence Systems, 2019.
[7] Bogdanov and et al., "Essentia: an audio analysis library for mir," Intern. Soc. for MIR Conf., 2013.
[8] Schuller and et al., "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," Computer Speech Language, 2018.
[9] Pedregosa et al., "Scikit-learn: Machine Learning in Python, JMLR 12, 2011.

Acknowledgments

This project was developed as part of a Master's Thesis at UPF and can be found and reproduced from: https://github.com/ana-pandrea/SincNet-MER

BCN Music Technology Group

upf. Universitat Pompeu Fabra Barcelona

MTG Music Technology Group