CROSS-CULTURAL MUSIC EMOTION RECOGNITION

Ana Gabriela Pandrea
Music Technology Group, Universitat Pompeu Fabra

ABSTRACT

Music, like any other art, represents an expression of emotions and moods and while humans perceive them naturally and almost instantly, implementing emotions and perception into a machine becomes challenging. The Music Emotion Recognition (MER) field has been exploited for quite a while, but it still faces difficulties because emotion is a very subjective aspect and many datasets are not reliable enough. To address these issues, we propose a set of experiments for music emotion classification with music of different cultures, in order to discover more sensitive patterns in emotion detection. Our goals were to investigate the relevant features for training, the machine learning algorithms appropriate for the task and finally, the results from training and testing the selected algorithms in a cross-cultural manner with music in English, Chinese and Turkish.

Keywords: Music Emotion Recognition; Classification; Cross-Culture

1. INTRODUCTION

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.

Current studies on emotion consider one of the two main taxonomies: categorical and dimensional [7]. 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 [8].

In the dimensional approach, emotions are identified on the basis of their location in a space with a small number of emotional dimensions. It has the advantage of being able to detect more details and shades of emotions, while maintaining a universal framework. The most famous and accepted model is Russell's Valence and Arousal (VA) plane [13] (Figure 1). It classifies emotion in one of the four quadrants which would suggest, in part, some sort of emotions as following: Q1: high arousal and positive valence – happiness, enthusiasm; Q2: high arousal and negative valence – anger, stress; Q3: low arousal and negative valence – sadness, nostalgia; Q4: low arousal and positive valence – calm, serenity.

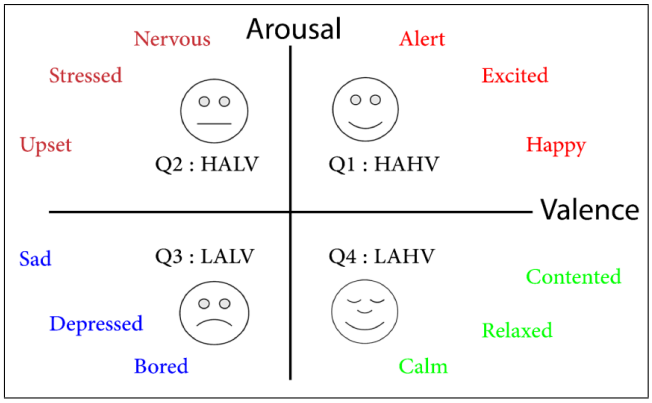


Figure 1. Arousal-Valence Plane for Emotions With Some Basic Emotions on it. The horizontal and vertical axes describe valence and arousal respectively [14].

After collecting the annotated data in one of these manners, the next step in a typical emotion recognition workflow is feature extraction. While [16] presents several purely sound-based and statistical low-level features like Mel-frequency Cepstrum Coefficients (MFCCs) or Statistical Spectrum Descriptors (SSDs) as the most commonly used descriptive features of MER, [1] reasons for more perceptual and meaningful-to-human features like Melodiousness, Articulation or Rhythmic Complexity. [11] also shows some effectiveness when considering texture and expressive techniques like vibratos and tremolos.

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 showed good results in some papers [12] [11] [16], to deep learning techniques like Convolutional Neural Networks (CNNs) with various architectures [15] [17], that proved to be especially good for speech and other sound related tasks [10].

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. [4] shows how the perceived consonance of chords is predicted by their relative similarity to voiced speech sounds. This means that because of the different word and sound dis-

tributions in each language, we expect that the extracted emotions will be different from one culture to the other. Other studies have shown that the bigger differences can be found in the valence plane that is a happy song in one language might not sound as happy to a different language speaker [9].

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. [9] 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 annotations. 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 particularly important for cross-dataset generalizability of models on arousal prediction.

Our aim with this paper is to further investigate this issue of cross-cultural music emotion recognition and observe the validity of the assumption that culture and language influence emotion perception. If this proves to be true, then we can conclude that music emotion recognizers can be improved using language-specific considerations. In this sense, we propose a set of experiments on three different datasets, with music in English, Chinese and Turkish, respectively, such that we can investigate the particularities and behaviours of each type of music when predicting emotions within-dataset and cross-dataset.

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.

The rest of the paper is structured as following: Section 2 describes the materials, i.e. datasets, and methods for feature analysis and classification we investigated. Section 3 describes the results of our individual and cross-cultural classification experiments, while Section 4 discusses these results and their implications. Our conclusions and further work are presented in Section 5.

2. MATERIALS & METHODS

The experiments that were conducted for this study can be divided into two categories. The first one is represented by individual investigations for achieving good classification results with each dataset on its own, along with feature analysis and extraction for each. Then, a cross-cultural analysis was done by training a classifier with music from one dataset, belonging to one culture - we will call this source dataset, and evaluating it with music from another - target dataset. The purpose of this was to see whether results with music from the target language are similar to the initial results from the same source language. The choice for this last classifier was made such that the classifier was the most optimal for the source dataset. To conclude our research, we also trained a classifier on all the three datasets and tested, on turn, for which cultural dataset it performs best, so that we can note how songs annotated in

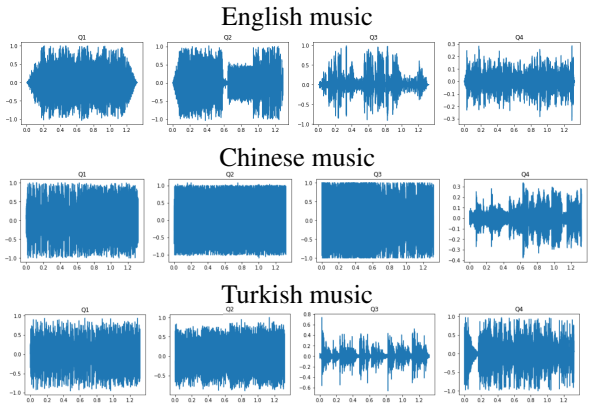


Figure 2. Sample waveforms representing a clip from each of the VA emotion quadrants.

different backgrounds behave, when given 'equal chances'. The number of training samples was equal for each culture.

2.1 Data

For the purpose of this study, we used three datasets with Western (English), Chinese (Mandarin) and respectively Turkish music. The Western one, mostly based on popularly consumed English music is called 4Q-Emotion and was derived through the AllMusic API by intersecting the original AllMusic tags with the Warriner's list [12] [11]. It contains 30-second excerpts from 900 different songs, equally split in one of the four quadrants of the Valence-Arousal plane (225 excerpts for each category).

The Chinese dataset, CH-818 [9], is made of 818 30-second excerpts, initially annotated with continuous values in the Valence-Arousal plane and then mapped to single quadrant output. Each clip was annotated by three music experts who were born and raised in Mainland China and thus were with a Chinese cultural background. This dataset was very unbalanced, therefore after balancing, we got only 89 songs of each category (356 in total).

The Turkish tr-music-dataset [5] is made of 400 song excerpts, 100 of each category, verbal and non-verbal music of different genres of Turkish music. It was originally split in the categorical taxonomy into four classes: happy, angry, sad, relaxed, which we mapped to the four quadrants: Q1, Q2, Q3 and respectively Q4. As a first impression about the three datasets, we plotted waveforms for each emotion quadrant for each of them (Figure 2).

In order to enrich the training of our classifiers, especially for the Chinese and Turkish datasets which are very small, we split the 30-second excerpts into four segments of 7-8 seconds each. Although some other researchers consider 30-second clips optimal [16], we listened to several clips and observed that they sound quite similar from the beginning to the end, in terms of timbre, rhythm and our subjective perception of emotion, thus we decided to apply segmentation as described. This resulted into 3600 equally distributed segments of English music, 1424 segments of Chinese and 1600 segments of Turkish music.

At this stage of data pre-processing, after balancing all

data, we split the individual datasets into training and testing sets with 20% of the data kept for testing. We performed this split before segmentation, in order to keep excerpts from the same songs in the same set and avoid a biased classification.

2.2 Features

Feature extraction was done with the Essentia music extractor [2] by retrieving 84 low-level descriptors, including loudness, silence rates, spectral features, mfccs and their statistic measures like mean or standard deviation. After extracting the features for each segment, we applied feature normalization in order to bring all features in the same scale, and then feature selection. Selection was performed with scikit-learn [6] by selecting the best k features for each dataset, where k was chosen experimentally as 50.

As expected, each dataset has its own set of best features, of which we investigated the first ten that obtained the highest scores in the selection process. If for the Western and Turkish music, we identified five common top-10 features (barkbands-spread-mean, spectral-centroid-mean, melbands-spread-mean, spectral-skewness-stdev, zerocrossingrate-mean), the Chinese dataset presented considerably different features. It only has silence-rate-30dB-stdev and silence-rate-60dB-stdev in common with Turkish and spectral-energyband-high-mean in common with English. Unique to English were features related to spectral complexity, rolloff and dissonance, to Chinese features related to hfc and pitch salience and to Turkish kurtosis statistics.

Top three features for the English dataset were barkbands-spread-mean, spectral-complexity-mean, spectral-centroid-mean, for Chinese silence-rate-20dB-stdev, silence-rate-60dB-stdev, hfc-stdev and for Turkish silence-rate-30dB-stdev, spectral-centroid-mean, spectral-skewness-stdev. In addition, we plotted the training data for each dataset onto some two-dimensional feature spaces to see how discriminative the first two features mapped together (Figure 3). While for Turkish the four classes can be discriminated quite well in this representation and for English only partly (only Q2 and Q4 are visually separated), the Chinese results are a bit underrepresented in space with many samples overlapping at x = 0; some boundaries could indeed be presumed, however.

The best ten features selected when we trained the classifier on combined music data, are as following: erbbands-crest-mean, spectral-entropy-stdev, dissonance-stdev, spectral-complexity-stdev, spectral-rolloff-mean, spectral-centroid-mean, zerocrossingrate-mean, spectral-complexity-mean, silence-rate-30dB-stdev and barkbands-spread-mean. Generally, they are similar to the previous individual results, but with an inclination towards Western music, with an overlap of seven out of ten features. This might indicate that

2.3 Classifiers

For this study, we considered several classifiers when training individual datasets, in order to select the one that

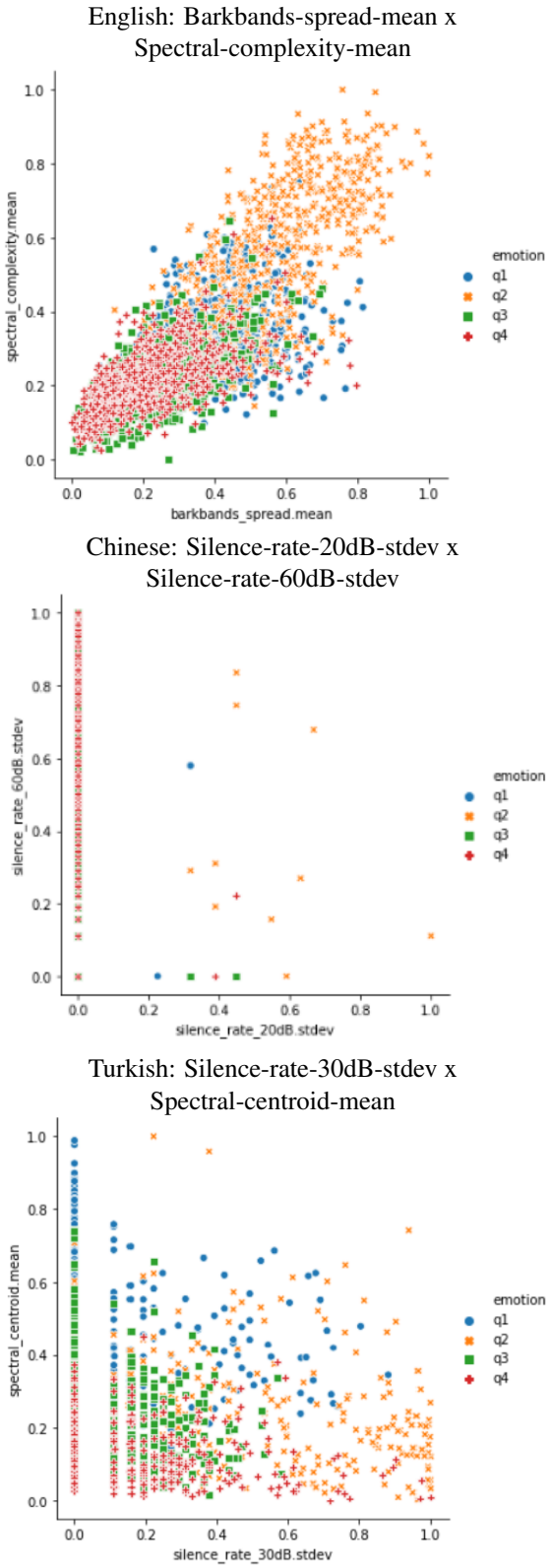
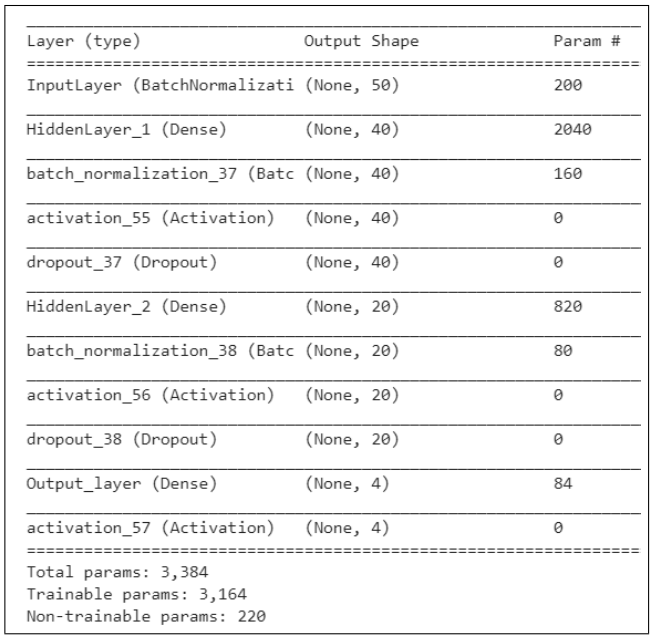


Figure 3. Training data plotted in the 2D feature space of the best 2 selected features.



Layer (type)	Output Shape	Param #
InputLayer (BatchNormalizati	(None, 50)	200
HiddenLayer_1 (Dense)	(None, 40)	2040
batch_normalization_37 (Batc	(None, 40)	160
activation_55 (Activation)	(None, 40)	0
dropout_37 (Dropout)	(None, 40)	0
HiddenLayer_2 (Dense)	(None, 20)	820
batch_normalization_38 (Batc	(None, 20)	80
activation_56 (Activation)	(None, 20)	0
dropout_38 (Dropout)	(None, 20)	0
Output_layer (Dense)	(None, 4)	84
activation_57 (Activation)	(None, 4)	0

Total params: 3,384
Trainable params: 3,164
Non-trainable params: 220

Figure 4. Summary of the Deep Neural Network considered in our experiments.

gives the best performance with our task of music emotion recognition, but also with each language. The more traditional machine learning algorithms, built with scikit-learn [6] that were compared for each dataset, are K-Nearest Neighbors (KNN) with n = 3, 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.

In addition, we also used a deep learning architecture, built with Keras [3], and trained a Deep Neural Network (DNN) using the same feature sets described in the previous subsection. We did keep the selected features because our datasets, especially the Chinese and Turkish ones, are not very big and there would not be enough samples to tune the model properly. The model is made of two hidden layers, each of them followed by batch normalization, activation and dropout. The Adam optimizer was used, along with binary cross-entropy loss, 300 epochs with a batch size of 50 segments and validation split of 20%. The training data was shuffled in order to ensure a random train-validation split and the output vectors were reshaped using One-Hot Encoding to allow for the four-neuron percentage output of the network. The summary of the network can be seen in (Figure 4).

As a first individual experiment, we ran a classifier comparison function on the set of seven classifiers from scikit-learn, 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 test set and computed its accuracy, precision, recall, f1-score and support, along with its confusion matrix, since we balanced our data. Finally, we compared this to the results of the presented deep neural network, for which we reported the same measurements, along with the learning curves for accuracy and loss. Results for each dataset will be provided in the next section.

In terms of cross-cultural analysis, a subset of the same scikit-learn classifiers was used, along with the same deep learning model and the same evaluation metrics. For the first cross-cultural experiment, we used the selected classifiers for each dataset, i.e. Random Forest (with depth=15), Multi-layer Perceptron and K-Nearest Neighbours (with k=7), but we also compared these to using the deep learning approach. Then, for the second experiment, where we trained with a multi-cultural dataset and tested for each culture on turns, we used the Random Forest Classifier as it gave persistent and reasonable results with all the datasets during individual investigations.

For all experiments, we made sure that the training sets are balanced and also, for the cross-cultural experiments, we had to ensure that the test sets contain the right set of features considered in the training set. All code for this research was written in Python and run using Google Colab, in the form of Jupyter notebooks and can be accesses at https://github.com/ana-pandrea/Cross-CulturalMER.

3. RESULTS

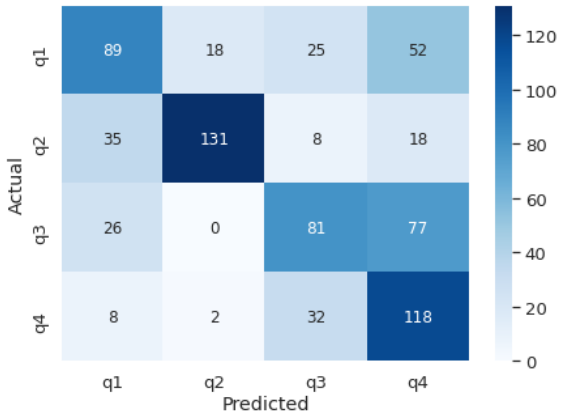
Our results in regard to music emotion classification in general do not have very high accuracies as we mostly used baseline classifiers and some of our datasets were quite small to be addressed with complex machine learning, like deep neural networks. However, we managed to compare three datasets of music in different languages, from various perspectives in the Music Emotion Recognition process, through several independent and cross-cultural experiments.

3.1 Individual experiments

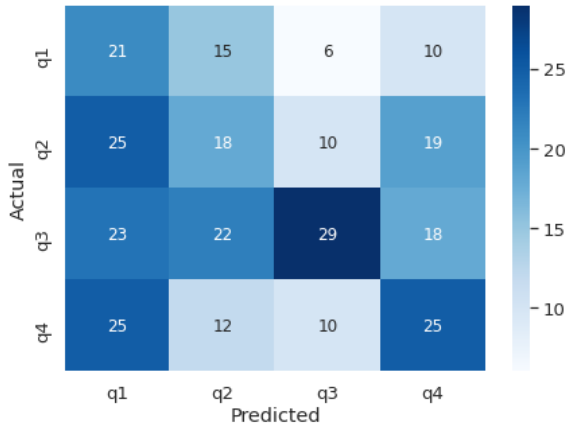
After running the classifier comparison on each of the three datasets with music in English, Chinese and Turkish, we selected the best model for each and reported its evaluation results. Therefore, for English, we picked the Random Forest Classifier with a 58% accuracy, for Chinese, results were sightly better, although not very satisfactory, with K-Nearest Neighbours with seven neighbours (32% accuracy), while Turkish music classification was optimised by the Multi-layer Perceptron (alpha = 1) with the highest accuracy among the three, at 69%. Their classification reports and the confusion matrices can be seen in Figure 5.

We can see that the best classified quadrants for English music were Q2 and Q4, with high confusion for Q3 tagged as Q4, suggesting that the valence of low arousal Western music could be quite difficult to identify. On the other hand, in the Chinese culture, Q3 seems to be the best classified emotion quadrant, while the most biases can be observed with wrong tags for Q1 across all four quadrants. This could imply that in Chinese culture, happiness and excitement might not be that strongly emphasized and confusions are easier to be made. In terms of Turkish music, Q2 provides the best classification accuracy, although

English: Random Forest Classifier, max_depth = 15



Chinese: K-Nearest Neighbours, k = 7



Turkish: Multi-layer Perceptron, alpha = 1

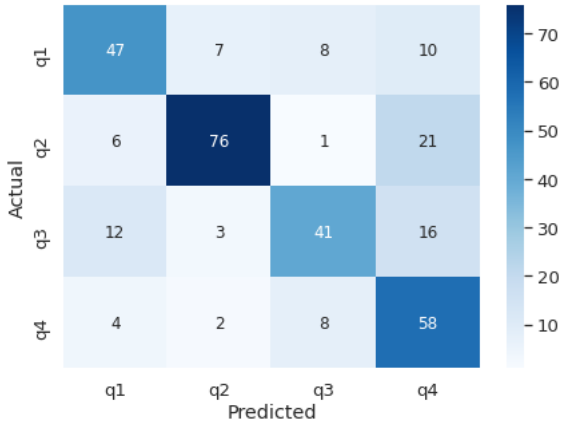


Figure 5. Confusion matrices for the best performing ML algorithms for each dataset.

Score	Precision	Recall	F1-score	Support	Total accuracy
EN	Q2	Q4	Q2	Q2	58%
CH	Q3	Q1	Q3	Q3	32%
TR	Q2	Q4	Q2	Q2	69%

Figure 6. Summary of the best results in the individual classification reports.

sometimes it is interestingly mistaken for Q4. In Figure 6 we selected the best scores from each classification report, along with their associated quadrants.

A common trend that we observed is that, among all the seven proposed models, the Random Forest Classifier, the Multi-layer Perceptron and the K-Nearest Neighbours models performed better than the other classifiers with all datasets. An exception is the Gaussian Process that was also particularly good with Turkish music. This generalisation suggests that the relevance of the best performing classifier stands behind the problem definition, MER in this case, and not that much in the training dataset.

Further on, we also experimented with the deep learning model, which performed slightly worse than the previous selected algorithms, with a 56% accuracy for English, 24% for Chinese and 54% for Turkish. With this model, similar patterns were observed: the mistaken classification of Q3 as Q4, but also better recall for the rest of the classes in the first dataset, the good recall of Q3, but also the really low f1-scores of Q2 and Q4, and finally, the general good results of Turkish music classification, but with an emphasized mistaken tagging of Q2 with Q4 and a 100% precision for Q2.

3.2 Cross-Cultural experiments

General results for the first cross-cultural experiments we conducted are summarized in Figure 7, where we trained the three selected classifiers (RFC, KNN, MLP) on their language sets accordingly and evaluated them on each of the three total test sets. While testing out the English trained classifier with another culture's music, we obtained significantly lower results, reasoning for some cultural particularities in terms of emotion discrimination, especially against Chinese. Similarly, results with evaluating the Turkish classifier with other sets were much lower than within-dataset testing. However, for the Chinese dataset, the cross-dataset accuracy scores were only slightly lower than the score on Chinese, all of them being very low. The reason for this could be the quality of the Chinese dataset in terms of both annotation agreement and quantity, but also the insufficient classification method.

Similarly, we also investigated the deep neural network with cross-dataset evaluations, such that we can note the

Train // Test	English	Chinese	Turkish
English	58%	28%	46%
Chinese	30%	32%	23%
Turkish	51%	29%	69%

Figure 7. Cross-Cultural evaluation of the best classifiers for the source (training) dataset.

Train // Test	English	Chinese	Turkish
English	57%	26%	38%
Chinese	32%	29%	24%
Turkish	48%	30%	55%

Figure 8. Cross-Cultural evaluation of the Deep Learning model.

testing behaviours on the same classifier, that had average performance for all datasets. We also investigated this model because of the various positive results of DL in MER literature. The accuracy scores maintain similar patterns as in the first experiment, with several lower scores, especially involving Turkish data. (Figure 8).

The final experiment that involved training the Random Forest Classifier on a mixed dataset and evaluating it for each different culture, provided the results in terms of confusion matrices in Figure 9. While the accuracies for English and Turkish are 54% and 62%, respectively, the accuracy of Chinese is very low, at only 24%. This suggests that the Chinese dataset is considerably different to the other two and samples from these only produce more noise. The model performed worse for all the considered datasets when trained collectively, but the variation is below 10%.

4. DISCUSSION

The first thing we investigated in this cross-cultural study were the selected features for training the machine learning classifiers. The fact that we obtained several different sets for the different training datasets considered suggests that music - emotion correlations are indeed different from one culture to another. While English and Turkish seem to distinct emotions by means of various spectral features and barkbands, Chinese uses more statistics of the silence rates and spectral energybands. This could mean that emotions in their culture are rather perceived more in terms of how dramatic the silent moments are, than in terms of what chords and harmonies are used, for example.

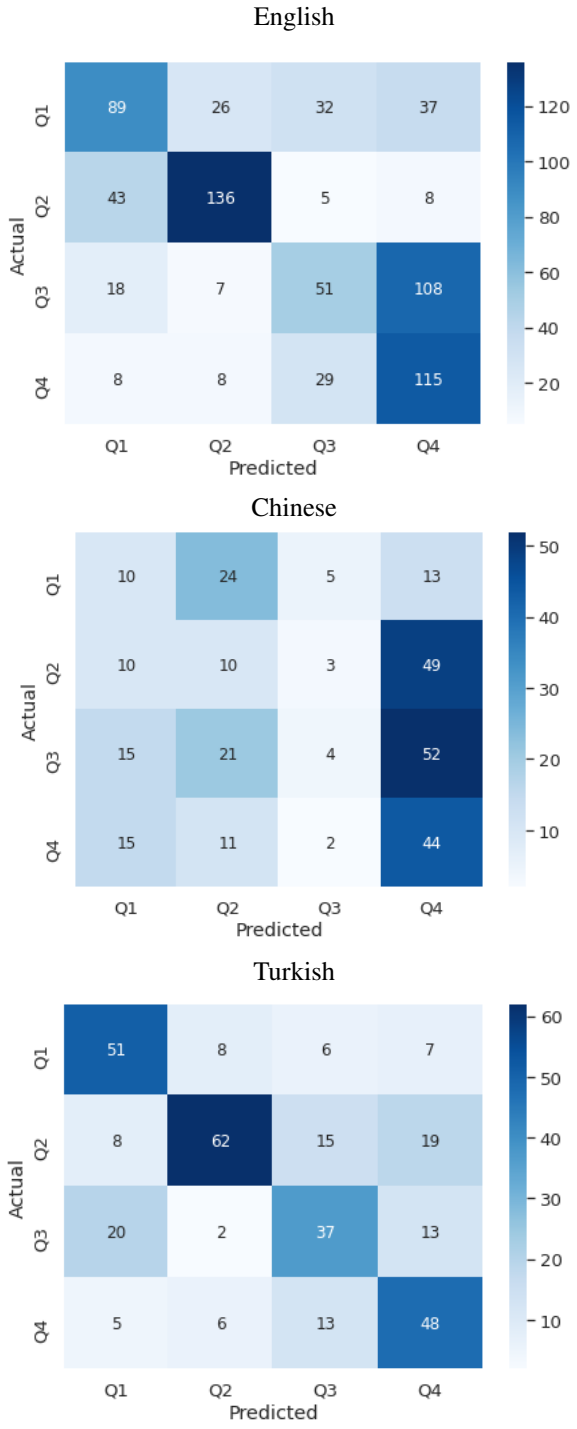


Figure 9. Confusion matrices for cross-cultural testing of the Random Forest Classifier with mixed training data.

In terms of machine learning methods, there were several algorithm configurations that seemed to perfom better than the others, independently of the dataset. This means that a good music emotion recognition solution should start with 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 Random Forest Classifier, the Multi-layer Perceptron and K-Nearest Neighbours. The deep learnig model that was proposed also provided some average-to-good results, but since our data was quite limited, this model did not perform at its full capacity.

In terms of prediction accuracy, the Turkish data generally scored best, with about 70% of data correctly classified. This was observed despite the fact that the English dataset is considerably bigger and supposed to allow for better learning. That is why we assumed that the emotionally discriminative features in this culture are more obvious and rigorously described or at least easier to learn computationally.

A particular observation that was made for the most experiments with the English dataset is the common confusion of Q3 with Q4, suggesting that, for this particular language, we should consider more discriminative features for sad and relaxed music. An option for this could be to consider whether the key is major or minor for the case of tonal music.

The Chinese dataset was the most problematic throughout experiments, obtaining really low classification scores. While this could suggest a problem with the dataset itself, it could also be the case that the selected features are not descriptive enough for this culture. If for English and Turkish, Figure 6 emphasizes similar performances in terms of the best classified quadrants, for Chinese music, results seem somehow opposed, denoting as well a possible cultural discrepancy.

Our hypotheses get confirmed with the cross-cultural experiments, where cross-dataset testing proves significantly less accurate than within-dataset. An exception for this is the increased accuracy of English music evaluation on the Chinese trained model with DL. As the algorithm in this case was KNN, the model did not rely too much on the training dataset, only on its best selected features and preoptimisation of k, thus these results are understandable.

In both Figure 7 and Figure 8 we can observe that the lowest results were again 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 Chinese language is completely different from the other two languages, first of all by the fact that it is tonal. Therefore, it is reasonable to believe that this aspect influences emotion perception in music, as well.

When training on the equally distributed mixed dataset, results show similar patterns like those discovered through individual experiments, like the very good classification of Q2 in the English and Turkish test sets. For the Chinese dataset, results were biased into classifying a lot of instances as Q4 while Q3 presented a very low recall and f1-score. In fact Q3 seems to be the weakest class among all the datasets, in this experiment, suggesting that Q3 could be a more subjective class depending on cultural background and language.

This last setup also strengthens the possible similarity of emotion perception in English and Turkish, by providing more accurate results, but also similar ratios within f1-scores. However, the Chinese results were not as good from the very beginning so we can not draw too many conclusions.

Nevertheless, if emotional response and features were similar enough within the three datasets, the evaluation results should increase as we feed the model with more data. But this was not the case, with even worse results, thus, we can infer that culture is an influential factor in emotion prediction and culture specific Music Emotion Classifiers are highly desirable.

5. CONCLUSION & FUTURE WORK

We believe that we fulfilled our scopes with this paper and showed that music in various languages and belonging to various cultures behaves differently in several circumstances within the MER cycle.

While Deep Learning is a promising approach for MER, we concluded that some traditional ML approaches could be better in some cases, especially with small datasets. Finding the very best algorithm for MER was beyond the scopes of this research, but we did investigate and find the best baseline algorithm for each of the three datasets with music in different languages.

Our main conclusion is that classifiers tested with the same language music perform better than with music in other languages. By investigating the best feature spaces and the results from our cross-cultural experiments, we observed that the CH-818 dataset seems the most uncorrelated to the others, suggesting that due to its different origins and characteristics, the Chinese language impacts emotion perception in music.

Further work includes analysing other features than the low-level ones and taking a more detailed look at the most relevant ones for each dataset. In addition, data augmentation or collecting more annotated files would also be a good idea, since we used a limited amount of files in this work. Finally, better ML algorithms should be considered, heading towards more complex and task specific DL architectures.

6. REFERENCES

[1] Anna Aljanaki and Mohammad Soleymani. A datadriven approach to mid-level perceptual musical feature modeling. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018, pages 615–621*, 2018.

[2] Wack N. Gomez E. Gulati S. Herrera P. Mayor O. et al. Bogdanov, D. Essentia: an audio analysis library for

music information retrieval. *International Society for Music Information Retrieval Conference (ISMIR'13). 493-498*, 2013.

[3] François Chollet et al. Keras. 2015.

[4] Dale Purvesb Daniel L. Bowlinga and Kamraan Z. Gill. Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences of the United States of America, 115(1) 216-221*, 2018.

[5] Mehmet Bilal Er and Ibrahim Berkan Aydilek. Music emotion recognition by using chroma spectrogram and deep visual features. *International Journal of Computational Intelligence Systems*, 12:1622–1634, 2019.

[6] Pedregosa et al. Scikit-learn: Machine learning in python. *JMLR 12, pp. 2825-2830*, 2011.

[7] Jacek Grekow. From content-based music emotion recognition to emotion maps of musical pieces. *Studies in Computational Intelligence, Volume 747, Springer International Publishing AG*, 2018.

[8] Downie J.S. Laurier C. Bay M. Ehmann A.F. Hu, X. The 2007 mirex audio mood classification task: lessons learned. *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, 14–18 September 2008, pp.462–467*, 2008.

[9] X. Hu and Y. Yang. Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs. *IEEE Transactions on Affective Computing*, 8(02):228–240, 2017.

[10] Yoshua Bengio Mirco Ravanelli. Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*.

[11] Paiva R. P. Panda R., Malheiro R. Musical texture and expressivity features for music emotion recognition. *19th International Society for Music Information Retrieval Conference – ISMIR 2018, Paris, France*.

[12] Paiva R. P. Panda R., Malheiro R. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing (IEEE early access)*, 2018.

[13] J.A. Russell. A circumplex model of affect. j. personal. *Soc. Psychol. 39(6), 1161–1178*, 1980.

[14] Paulina; Ferri Cèsar Vallejo-Huanga, Diego; Morillo. Icmla 2014/2015/2016/2017 accepted papers data set, mendeley data, v2 http://dx.doi.org/10.17632/wj5vb6h9jy.2file-131f313d-2347-4f23-a0a7-8cd5c81eca49.

[15] Shreyan Chowdhury Andreu Vall Verena Haunschmid Gerhard Widmer. Towards explainable music emotion recognition:the route via mid-level features.

[16] Juan Li Xinyu Yang, Yizhuo Dong. Review of data features-based music emotion recognition methods, springer-verlag gmbh germany. 2017.

[17] Xi Zhao Yizhuo Dong, Xinyu Yang and Juan Li. Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for music emotion recognition. *IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 21, NO. 12, DECEMBER 2019*.