Transfer learning from speech to music: towards language-sensitive emotion recognition models

Juan Sebastián Gómez Cañón*, Estefanía Cano†, Perfecto Herrera*, Emilia Gómez*‡
*Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
†Social and Cognitive Computing Department, A*STAR, Singapore, Singapore
‡Joint Research Centre, European Commission, Seville, Spain
Email: juansebastian.gomez@upf.edu

***Abstract*—In this study, we address emotion recognition using unsupervised feature learning from speech data, and test its transferability to music. Our approach is to pre-train models using speech in English and Mandarin, and then fine-tune them with excerpts of music labeled with categories of emotion. Our initial hypothesis is that features automatically learned from speech should be transferable to music. Namely, we expect the intra-linguistic setting (e.g., pre-training on speech in English and fine-tuning on music in English) should result in improved performance over the cross-linguistic setting (e.g., pre-training on speech in English and fine-tuning on music in Mandarin). Our results confirm previous research on cross-domain transferability, and encourage research towards language-sensitive Music Emotion Recognition (MER) models.**

***Index Terms*—Sparse convolutional autoencoder, speech emotion recognition, music emotion recognition, unsupervised learning, transfer learning, multi-task learning.**

I. Introduction

There is a strong relationship between the recognition of perceived emotions in both speech and music: shared taxonomies of emotion [1], [2], shared biological and evolutionary processes in the brain [3], [4], similar acoustic cues that are related in both domains [5], [6], and models that attempt to recognize perceived emotions across domains [7], [8]. Researchers have evaluated how individual differences influence emotions perceived in music and speech prosody by English speakers [9]. They found that the ratings on speech are unaffected by factors such as personality, musical training, emotional intelligence and gender, but may be influenced by age. On the other hand, ratings on music are unaffected by gender and emotional intelligence, and minimally influenced by personality, musical expertise and age. These findings are congruent with previous hypotheses of shared affective processing of stimuli in the brain for both domains [3], but also suggest that individual differences may be more important in the domain of music than for speech. Research in music psychology has pointed out that emotion perception is also influenced by additional factors such as empathy, cultural background, generation, sex, and personality [10].

To deal with the inherent subjectivity in music emotions, a reasonable approach in the literature has been to build MER models tailored to different users (personalized MER) and user groups (groupwise MER). Personalized MER refers to using annotations from a specific user, and training a personalized model. Groupwise MER gathers users according to individual factors (e.g., demographics, musical expertise, and personality), and averages the annotated data as common "ground truth" [11]. Studies have shown that personalized models result in heavy cognitive burden to the users during the annotation process, and that groupwise models have not significantly outperformed general models in terms of accuracy [12]. The scope of our work is to study the effect on classification accuracy of groupwise language-sensitive MER models. We address two research questions in this paper: (1) Can inductive transfer, as used in the field of deep learning, be used to create music emotion recognition models that are sensitive to language? (2) Should language be considered as a personal difference to be used when designing and improving the performance of MER models?

The rest of the paper is structured as follows: in Section II we discuss related work. In Section III we detail the methodology of our study, including the selected data sets and network architectures. Section IV describes our results, which are later discussed in Section V.

II. Related work

The majority of work regarding speech emotion recognition (SER) and music emotion recognition (MER) has focused on the extraction of emotionally-relevant features, and the implementation of classification or regression models that predict emotion categories or arousal-valence (AV), respectively[1]. However, very few studies are directed to cross-domain research, even though there is ample evidence of shared cognitive processes in the humans' parsing of speech and music [16]. Researchers have analyzed audio to find correlations in emotion-related features across speech, music and sound [6]. The authors obtained a set of low-level acoustic descriptors for the recognition of AV across the three domains, and proposed a cross-domain correlation coefficient as a method for selection of features that generalize for cross-domain AV prediction. This resulted in the InterSPEECH2013 Computational Paralinguistic Challenge (IS13 ComParE) feature set of cross-domain emotionally-relevant descriptors.

[1]For a detailed review on SER refer to [13], [14] and on MER refer to [11], [15].

Other studies have also compared perceived emotions by music and speech prosody using time-continuous evaluations. In [7] for example, the authors attempted to produce a model of psychoacoustic cues of emotion communication common to both domains. More recently, these authors have also explored shared acoustic codes between speech and music using deep learning [8]. They predicted perceived emotion in music using models trained on emotional speech and vice-versa by using transfer learning techniques [8]. In this case, the authors pre-trained a denoising autoencoder (DAE) using IS13 features from speech in British English, and performed transfer learning to predict time-continuous AV in instrumental classical music. Their results show that transferring from speech to music was more successful than in the opposite direction.

Other works have explored cross-cultural and cross-data set MER [17], [18]. Authors used acoustic features related to loudness, pitch, rhythm, timbre, and harmony, which have been thoroughly studied for Western music [5], [19]. Their study tested the generalization of models trained with these features on non-Western music. Researchers applied their models on three data sets developed for emotion prediction, and selected subsets in order to homogenize inter-rater agreement (as defined by Krippendorff's α)[2]. Finally, they trained support vector regressors, and evaluated their performance across the different data sets. Their results suggest that the most important factor for cross-data set generalizability is inter-rater agreement and that it is largely supported, mainly for the arousal dimension.

The goal of our work is to build upon the work of [7], [8] to develop emotion *classifiers*, and use language both as a *source of data* (in the case of speech) and as a *personal difference* (in the case of lyrics of music). We aim to develop language-sensitive MER models, that are customized to different user groups and evaluate systematically different architectures. To achieve this, the following steps were conducted: (1) We collected users ratings to understand music emotion perception and agreement among people with different mother tongues, (2) We implemented a benchmark model based on [8] for music emotion classification, and (3) We proposed and extended a model with language-sensitive characteristics using a multi-task learning approach [21]. Our contribution is to take language of speech and music into account to develop our models, while [7], [8] use speech in English and French indistinctly since IS13 features do not focus on linguistic aspects of speech. Additionally, we use transfer learning to exploit speech data and improve performance, differing from [17], [18].

III. Methodology

A. Agreement analysis

To better understand the influence of language in emotion perception, we conducted online surveys in four languages (English, Spanish, German and Mandarin) to test for differences and similarities of the emotions perceived in music by listeners with different native languages (see [22] for details). We analyzed emotion annotations using inter-rater reliability statistics of musical fragments from different styles (mainly pop and rock in English) that belong to the 4Q-Emotion data set (see Section III-B2). We used 22 musical fragments related to 11 categories of emotion by querying the emotion from the metadata, and asked participants to rate them on a 5-point Likert response format. Additionally, we gathered information on the participants' music sophistication, preference, familiarity, and lyrics comprehension (LC) for each fragment. We had unbalanced participation for the surveys: English ($n = 26$), Spanish ($n = 56$), German ($n = 17$), and Mandarin ($n = 27$). Hence, we initially analyzed the resulting 23562 ratings from all participants ($n = 126$).

[2]Agreement is the proportion of the observed to the expected above-chance agreement amongst different raters [20].

TABLE I
Summary of speech and music data sets: AV refers to arousal-valence and size in (n h) is the amount of data used for training.

	LibriSp.	**Aish.**	**4Q-Emo.**	**CH-818**
Type	Speech	Speech	Music	Music
Language	Eng.	Man.	Eng./Spa.	Man.
Annotation	-	-	Quadrant AV	Numeric AV
Size	100h (7.5h)	178h (3h)	900 clips (7.5h)	356 clips (2.96h)

B. Data sets

To train the models in this work, different speech and music data sets, both in English and in Mandarin, were used.

1) Speech Data: To train models on English speech, the Librispeech data set was used in this work [23]. Librispeech is a speech recognition data set containing more than 1000 hours of speech from public domain audio books belonging to the LibriVox project. To train the models with Mandarin speech, the AISHELL data set was used [24]. AISHELL was collected from 400 participants from different regions in China who read 500 sentences covering different domains: smart homes, autonomous driving, entertainment, science, and news. To train our models, we randomly selected a subset from each data set: 85% of the data was used to train, and 15% was used for validation during pre-training (see Table I).

2) Music Data: Labeled music data was used to train our MER models. To train our English models, the 4Q-emotion data set was used [25]. It contains mainly popularly consumed music, including pop, rock, and metal. The metadata was collected from the AllMusic API by selecting tags, intersecting them with emotional adjectives, and then mapping annotations to the four quadrants of AV space [1]. *Q1* corresponds to positive arousal-positive valence (e.g., happiness), *Q2* corresponds to positive arousal-negative valence (e.g., anger), *Q3* corresponds to negative arousal-negative valence (e.g., sadness), and *Q4* corresponds to negative arousal-positive valence (e.g., tenderness). Even though the 4Q-emotion data set contains

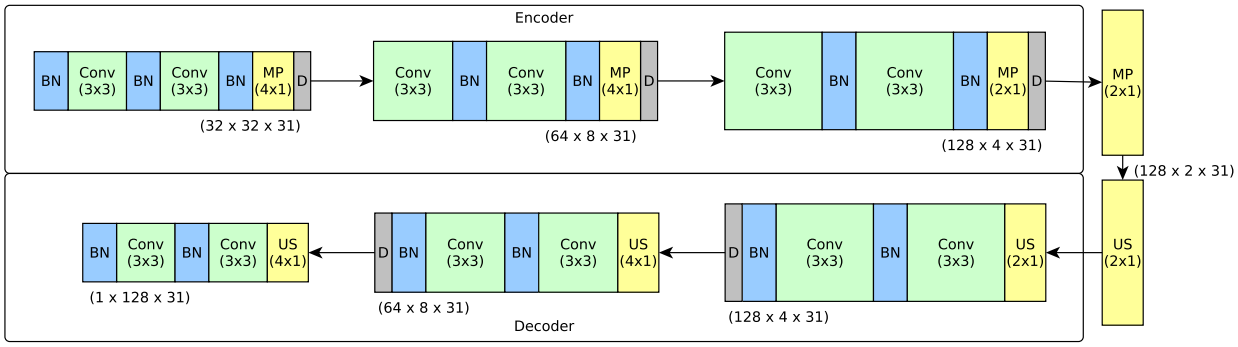

Fig. 1. Proposed network architecture where BN is batch normalization, D represents dropout, LS is the higher-dimensionality latent space, MP is MaxPooling2D, and US is UpSampling2D. Each double conv-layer increases number of filters linearly: 32, 64 and 128, respectively.

other languages than English, an analysis of the data set using polyglot reveals that at least 80% of the data is sung in English.To train our Mandarin models, the CH-818 data set was used [18]. It contains Chinese pop songs released in Taiwan, Hong Kong and Mainland China. Each clip was annotated by three musical experts from China with an interface consisting of two sliding bars of continuous real values between [-10,10] for AV space. To make the English and Mandarin music data sets comparable, numeric AV annotations in CH-818 were mapped to their corresponding quadrants. Both data sets were split considering the number of classes into the following: 70% for training (85% training, 15% validation), and 30% for testing (see Table I). It is important to note that the 4Q-Emotion and LibriSpeech data sets contain non-tonal languages (mainly English), while AISHELL and CH-818 contain tonal languages (only Mandarin). This distinction is important since tonal languages can convey different semantic meaning through speech prosody (i.e., different intonations of the same words can have different meanings). Given the scarcity of music emotion data sets, we balance equal amount of speech and music data (in hours) for each language, respectively (see Table I).

C. Models

Firstly, we reproduced the model presented in [8]: a denoising autoencoder (DAE) with Long Short-Term Memory (LSTM) latent space that inputs feature vectors injected with Gaussian noise. We extracted 260 emotionally-relevant features at a rate of 1 Hz, following [6], [8]. After pre-training, weights were kept fixed, and a new LSTM block with an output layer using sigmoid activations was added for transfer learning. We obtained comparable results for regression to the ones presented by the authors only on the transfer from speech to music. In order to implement classifiers, we substituted the sigmoid with softmax activations at the output layer and processed different time frames: (1) processing 5s per batch (hidden layer of 100 neurons - *DAE* - *Classifier* in Table II), and (2) using an over-complete latent space and processing 10s per batch (hidden layer of 800 neurons - *DAE* - *Sparsity*). DAEs have been found to improve their denoising performance with sparsity: a latent space with higher dimensionality than the input space [26].

Secondly, we designed a sparse convolutional autoencoder (SCAE) with rectified linear unit activations (ReLU), as seen in Figure 1. All data sets were processed with the librosa package to extract mel-spectrograms. The audio was converted to mono and downsampled to 16kHz. A Short-Time Fourier Transform (STFT) with a window size of 1024 ($\sim$46ms) and 512 hop size ($\sim$23ms) was used. The resulting mel-spectrograms had a dimensionality of 128 mel-bands by 31 time frames per second, extracted with an overlap of 50%. The dimensionality of an input mel-spectrogram feature (1 x 128 x 31) is increased to (128 x 2 x 31) in the latent space, by three double conv-layers augmenting the number of filters in the encoder: 32, 64, and 128, respectively. Dropout is set to 0.25 after every double conv-layer to prevent overfitting. Additionally, max-pooling and up-sampling are used to diminish and augment the dimensionality of the features with a variable pool size. Batch normalization is applied after each non-linearity to address internal covariate shift during training. We train each model four times and report performance using macro-weighted averages across experiments. We use a mean square error loss function for pre-training, learning rate of 0.001 (Adam optimization) and add random Gaussian noise (μ=0, σ=0.3). After pre-training, transfer learning is implemented by removing the decoder and adding a flattening layer, 3 fully connected layers each with 512 neurons, followed by a Dropout layer each. Since we perform multi-task learning (MTL), we add three blocks of 2 fully connected layers (512) followed by a Dropout layer each, and three output layers with softmax activation. We implement MTL, since optimizing losses in the auxiliary tasks, can help improve generalization upon a main task. Each block represents a classifier: (1) quadrant prediction (4 classes, one per quadrant), (2) arousal prediction (positive: Q1 and Q2, negative: Q3 and Q4), and (3) valence prediction (positive: Q1 and Q4, negative: Q2 and Q3).We then use categorical cross-entropy as the loss function for quadrant classification, and binary cross-entropy for classification of positive/negative arousal and positive/negative valence. Transfer learning is performed first by freezing the weights from the encoder, and fine-tuning the

TABLE II

OVERALL RESULTS OF PRECISION (P), RECALL (R), AND F-SCORE (F) FOR ALL EXPERIMENTS. WE REPORT ONLY MACRO-WEIGHTED AVERAGES TO ACCOUNT FOR CLASS IMBALANCE. INTRA-LINGUISTIC SETTINGS ARE REPORTED AS ENG2ENG (E.G., PRE-TRAIN ON LIBRISPEECH AND TRANSFER LEARN ON 4Q-EMOTION) AND CROSS-LINGUISTIC SETTING AS MAN2ENG (E.G., PRE-TRAIN ON AISHELL AND TRANSFER ON 4Q-EMOTION).

		Mandarin			English		
		P	R	F	P	R	F
Baseline CNN	Quadrants	0.29	0.41	0.34	0.23	0.48	0.31

		Man2Man			Man2Eng			Eng2Eng			Eng2Man		
		P	R	F	P	R	F	P	R	F	P	R	F
DAE - Classifier	Quadrants	0.46	0.48	0.46	0.65	0.65	0.65	**0.64**	**0.64**	**0.64**	**0.46**	**0.48**	**0.46**
DAE - Sparsity	Quadrants	0.46	0.48	0.46	0.64	0.64	0.64	**0.56**	**0.54**	**0.54**	**0.46**	**0.48**	**0.45**
SCAE - Feat. Ext.	Quadrants	**0.42**	**0.58**	**0.49**	**0.52**	**0.49**	**0.46**	0.52	0.48	0.45	0.42	0.58	0.49
	Arousal	**0.63**	**0.64**	**0.63**	**0.67**	**0.64**	**0.62**	0.65	0.63	0.62	0.63	0.64	0.63
	Valence	**0.77**	**0.78**	**0.77**	**0.78**	**0.74**	**0.74**	0.77	0.72	0.71	0.78	0.79	0.78
SCAE - Full	Quadrants	0.50	0.58	0.50	0.57	0.55	0.54	**0.61**	**0.58**	**0.58**	**0.29**	**0.51**	**0.36**
	Arousal	0.65	0.64	0.64	0.70	0.66	0.65	**0.69**	**0.67**	**0.66**	**0.42**	**0.60**	**0.49**
	Valence	0.80	0.80	0.80	0.82	0.81	0.81	**0.83**	**0.83**	**0.83**	**0.51**	**0.68**	**0.57**

network on the remaining layers at a learning rate of 0.0001 (*SCAE - Feat. Ext.* in Table II). On a second test, we unfreeze the weights of the whole network and continue training with a learning rate of 0.0005 (*SCAE - Full*), following [27]. We perform Bayesian optimization to select optimal learning rates and decays for Adam algorithm. We make the trained models available for testing[3].

IV. RESULTS AND DISCUSSION

A. Agreement Analysis

Inter-rater statistics show evidence that there are significant differences of emotional ratings by listeners raised in different mother tongues. In general, participants showed different distributions of ratings in the majority of cases. Interestingly, only the distributions of ratings of *joy* and *peace* appeared to have similar distribution across languages. Our results have also confirmed that basic emotions will have higher universal agreement, while complex ones will show the opposite. We found overall low agreement for emotions such as *bitterness*, *fear*, *power*, *surprise*, and *transcendence*. Finally our findings suggest that preference, familiarity, and lyrics comprehension (LC) improve agreement for emotions in quadrants Q1 and Q3, and decreases it for quadrants Q2 and Q4. Namely, it relates to the type of emotions mapped to each of the quadrant, and subjectivity regarding valence. This has given us new understanding of the effect of LC and its impact on different emotions: in the case of Q1 (positive arousal and valence) and Q3 (negative arousal and valence) higher agreement is found, as opposed to Q2 (positive arousal/negative valence) and Q4 (negative arousal/positive valence) where dimensions have opposite signs. Thus we conclude that using less categories (i.e., quadrants) is more consistent when attempting cross-cultural emotion recognition, due to the difficulty of using equivalent emotion adjectives in all languages. This further motivates the need of attempting to create language-sensitive models, since improved agreement in annotation could potentially lead to higher performance of models.

[3]https://github.com/juansgomez87/quad-pred

B. Classifiers

To effectively test the feasibility of cross-domain transferability, and to verify that our models are indeed language-sensitive, we test three scenarios: (1) Models with the same architecture as our SCAE trained only on music (*Baseline CNN* on Table II), (2) One-step transfer learning (i.e., without unfreezing weights) yields a feature extractor trained only on speech and a classifier trained on music, which we hypothesize should retain emotion-related representations from speech in each language, and (3) Intra-linguistic configurations which we hypothesize should show improved performance over cross-linguistic configurations (i.e., a model pre-trained with Librispeech and fine-tuned with 4Q-Emotion should have a higher performance than a model fine-tuned on CH-818).

Classification results are summarized in Table II. The outcome of using a CNN trained only on music results in poor performance (F-score $\sim$ 0.32) for both music data sets. This suggests that although the architecture could be improved to obtain better performance, our SCAE models exploit the data learned during pre-training positively obtaining higher performance with the same architecture. Secondly, using the SCAE exclusively as a feature extractor shows average performance for all configurations (F-score $\sim$ 0.48). Nonetheless, it outperforms the baseline CNN, suggesting that the features learned during pre-training are generally transferable to music as well, confirming the findings from [8]. Further inspection using confusion matrices shows that in both intra-linguistic (eng2eng and man2man) and cross-linguistic (eng2man and man2eng) settings, the principal confusions are made between Q1 and Q2 (both with positive arousal) and Q3 and Q4 (both with negative arousal). This confirms research where arousal is more easily predicted, since it relates to features such as tempo and loudness, while valence is more subjective and cultural-specific [5]. Our model (SCAE-Full) improves quadrant, arousal, and valence prediction in most cases (man2man, man2eng, and eng2eng) w.r.t. SCAE-Feat. Ext., demonstrating benefits of fine-tuning with music. Although the SCAE-Full does not outperform the DAE-Classifier for man2eng,

eng2eng, and eng2man, our model is solely based on mel-spectrogram pattern recognition, while the DAE relies on previously extracting carefully hand-crafted features.

Finally, preliminary results suggest that the hypothesized improvement of intra-linguistic models over cross-linguistic models is feasible (highlighted in bold): eng2eng achieves an improvement up to $\sim$18% F-scores over eng2man in DAE-Classifier and in the SCAE-Full model, improving the prediction of quadrants, arousal, and valence. It must be noted that the amount of data for each speech dataset might differ fine-tuning results. In general, our DAE reproduction does not exhibit language-sensitive features (eng2eng outperforms eng2man, but not man2man over man2eng). Interestingly, man2man shows similar performance to man2eng in the full model (SCAE-Full). We argue that a possible reason is the existence of confounding acoustic features of excerpts belonging to Q2 (*angry*) and Q4 (*relaxed*) in CH-818. The CH-818 data set contains mainly pop music with high acoustic homogeneity. In contrast in 4Q-Emotion, Q2 contains fragments of rock and metal, which have very distinctive acoustic features (i.e., guitar distortion, screaming voice). With respect to languages, we find that our model shows language-sensitive features for Mandarin (tonal) in the SCAE-Feat. Ext. configuration, while showing it for English (non-tonal) in the SCAE-Full configuration.

V. Conclusions

In this work, we present preliminary results on MER language-sensitive models obtained by using transfer learning on different neural network architectures. We first reproduced the work of [8], and evaluated its performance on the classification task, while taking into account the language of speech and music. With respect to our research questions: (1) We proposed sparse convolutional autoencoders for automatic high-level feature learning of mel-spectrograms using a MTL approach. Our approach is based on feature learning as opposed to the existing model that uses hand-crafted features, and shows partial evidence of the plausibility of language-sensitive models. (2) We show that pre-training on speech can result beneficial for MER, and our surveys confirm previous research w.r.t. cultural differences in musical emotion perception, motivating this study. As future work, we intend to improve the unsupervised learning phase to extract better features from speech and perform transfer learning on more similar domains (i.e., speech and choir music).

VI. Acknowledgements

The research work conducted in the Music Technology Group at the Universitat Pompeu Fabra is partially supported by the European Commission under the TROMPA project (H2020 770376).

References

[1] J. A. Russell, "A Circumplex Model of Affect," *Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] K. Hevner, "Experimental studies of the elements of expression in music," *American Journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.

[3] A. D. Patel, *Music, Language and the Brain*. Oxford University, 2008.

[4] D. Purves, *Music as Biology*. London, England: Harvard University Press, 2017.

[5] P. N. Juslin, *Musical Emotions Explained*, 1st ed. Oxford: Oxford University Press, 2019.

[6] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, K. R. Scherer, and J. Krajewski, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, pp. 1–12, 2013.

[7] E. Coutinho and N. Dibben, "Psychoacoustic cues to emotion in speech prosody and music," *Cognition and Emotion*, vol. 27, no. 4, pp. 658–684, 2013.

[8] E. Coutinho and B. Schuller, "Shared acoustic codes underlie emotional communication in music and speech - evidence from deep transfer learning," *PLoS ONE*, vol. 12, no. 6, 2017.

[9] N. Dibben, E. Coutinho, J. A. Vilar, and G. Estévez-Pérez, "Do Individual Differences Influence Moment-by-Moment Reports of Emotion Perceived in Music and Speech Prosody?" *Frontiers in Behavioral Neuroscience*, vol. 12, pp. 1–13, 2018.

[10] J. Vuoskoski and T. Eerola, "The role of mood and personality in the perception of emotions represented by music," *CORTEX*, vol. 47, pp. 1099–1106, 2011.

[11] Y.-H. Yang and H. H. Chen, "Machine Recognition of Music Emotion: A Review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, 2012.

[12] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen, "Music Emotion Recognition: The Role of Individuality," National Taiwan University, Tech. Rep., 2007.

[13] B. W. Schuller, *Intelligent audio analysis*. Springer, 2013.

[14] ——, "Speech Emotion Recognition two decades in a Nutshell," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[15] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS 1*, pp. 1–22, 2017.

[16] D. L. Bowling, J. Sundararajan, S. Han, and D. Purves, "Expression of Emotion in Eastern and Western Music Mirrors Vocalization," *PLoS ONE*, vol. 7, no. 3, p. 31942, 2012.

[17] X. Hu, J. H. Lee, K. Choi, and J. S. Downie, "A cross-cultural study of mood in K-POP Songs," in *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 385–390.

[18] X. Hu and Y.-H. Yang, "Cross-Dataset and Cross-Cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs," *IEEE TRANS. ON AFFECTIVE COMPUTING*, vol. 8, no. 2, pp. 228–240, 2017.

[19] P. N. Juslin, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford: Oxford University Press, 2010.

[20] K. H. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. SAGE Publications, 2004.

[21] J. Kim, J. Urbano, C. C. S. Liem, and A. Hanjalic, "One deep music representation to rule them all? a comparative analysis of different representation learning strategies," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1067–1093, Feb 2020.

[22] J. Gómez-Cañón, P. Herrera, E. Gómez, and E. Cano, "The emotions that we perceive in music: the influence of language and lyrics comprehension on agreement," 2019. [Online]. Available: https://arxiv.org/abs/1909.05882v1

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[24] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proceedings of the 20th Conf. of the Oriental Chapter of the Int. Coord. Committee on Speech Databases and Speech I/O Sys. and Assessment*, Nov 2017, pp. 1–5.

[25] R. Panda, R. M. Rui, and P. Paiva, "Musical texture and expressivity features for music emotion recognition," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[26] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008, pp. 1096–1103.

[27] A. Diment and T. Virtanen, "Transfer learning of weakly labelled audio," in *IEEE Workshop on Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 6–10.