ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413721

LANGUAGE-SENSITIVE MUSIC EMOTION RECOGNITION MODELS: ARE WE REALLY THERE YET?

Juan Sebastián Gómez-Cañón[⋆] *Estefanía Cano*[†] *Ana Gabriela Pandrea*[⋆]
Perfecto Herrera[⋆] *Emilia Gómez*[‡⋆]

[⋆] Music Technology Group, Universitat Pompeu Fabra, Spain
[†] Songquito UG, Erlangen, Germany
[‡] European Commission, Joint Research Centre, Seville, Spain

ABSTRACT

Our previous research showed promising results when transferring features learned from speech to train emotion recognition models for music. In this context, we implemented a denoising autoencoder as a pretraining approach to extract features from speech in two languages (English and Mandarin). From that, we performed transfer and multi-task learning to predict classes from the arousal-valence space of music emotion. We tested and analyzed intra-linguistic and cross-linguistic settings, depending on the language of speech and lyrics of the music. This paper presents additional investigation on our approach, which reveals that: (1) performing pretraining with speech in a mixture of languages yields similar results than for specific languages - the pretraining phase appears not to exploit particular language features, (2) the music in Mandarin dataset consistently results in poor classification performance - we found low agreement in annotations, and (3) novel methodologies for representation learning (Contrastive Predictive Coding) may exploit features from both languages (i.e., pretraining on a mixture of languages) and improve classification of music emotions in both languages. From this study we conclude that more research is still needed to understand what is actually being transferred in these type of contexts.

Index Terms— Contrastive predictive coding, speech emotion recognition, music emotion recognition, representation learning, transfer learning, multi-task learning.

1. INTRODUCTION

The need of including context-based information to the Music Information Retrieval field, and particularly to Music Emotion Recognition (MER), has become critical [1]. In the case of music and emotions, the strong relationship between speech and music could be considered context [2], since our linguistic and cultural background reflect fundamental differences in our perception of sound. This theory is known as the *vocal similarity* hypothesis [3]. Furthermore, the perception of speech, music, and sound share acoustic features that are "emotionally-relevant" [4]. Therefore, it is likely to assume that the acoustic cues humans use to recognize emotions in sound might be common for both speech and music [5, 6] - which could explain associating a yelling person and black metal with the emotion *anger*. Recent cross-cultural studies demonstrate the lack of universality regarding the subjective perception of emotions [7, 8]. Given that culture-specific characteristics drive emotional perception, our research explores the *transferability* of acoustic features from speech in a *particular* language to music. While *vocal similarity* is still being studied and debated, our approach is based on training a neural network with speech in English and Mandarin, and then performing transfer learning to classify emotions from music with lyrics in each language. The present work is, additionally, an honest revisitation from our previous work [9] and contributes with further research using novel deep learning architectures for unsupervised representation learning.

The rest of the paper is structured as follows: in Section 2 we discuss related work. Section 3 details the methodology of our study, including the selected datasets and network architectures. Section 4 describes our results, which are later discussed in Section 5.

2. RELATED WORK

Research on the transfer of Speech Emotion Recognition (SER) models to the MER regression task [10], has shown successful transfer learning for recognition of emotions from speech in English to classical music.[1] In the case of SER, the work on linguistic research has gained more importance lately: using a bag-of-audio-words approach to exploit linguistic features [14], combining speech-based and linguistic classifiers [15], or using linguistic and acoustic cues for end-to-end models [16]. The linguistic approach to emotion recognition is mainly due to the fact that different emotion adjectives will tend to have diverse meanings across cultures [17] and that translation of words might result questionable [8]. However, this topic has just recently started to be explored in MER.

A logical approach to handle the inherent subjectivity of emotion annotations is to group annotations from similar annotators based on personal characteristics [18, 19]. In our previous work [9], we attempted to develop language-sensitive MER models by using language both as a *source of pretraining data* (in the case of speech) and as a *personal characteristic* (in the case of the lyrics of music). We trained convolutional denoising autoencoders using time-frequency representations of audio as input (i.e., mel-spectrograms), in order to obtain a feature extractor trained on speech. Our aim was to automatically extract features from speech in a particular language, which could be then used to train a classifier of music with lyrics in the same language.

We employed classifiers with four distinct classes related to the arousal-valence emotion model. Russell popularized this emotion taxonomy: *arousal* refers to the amount of energy from an emotion, while *valence* refers to its positiveness or pleasantness [20]. For example, an emotion such as *happiness* would have positive arousal and valence, while *anger* would have positive arousal and nega-

[1] For a detailed review of SER, we refer the reader to [11, 12] and on MER to [13].

tive valence. Since the exact position of an adjective in this space may be imprecise [21], we reduced the possibilities by defining four categories of quadrants following [22]: Q1 (positive arousal - positive valence), Q2 (positive arousal - negative valence), Q3 (negative arousal - negative valence), and Q4 (negative arousal - positive valence). By employing multi-task learning (MTL), our classifier was trained simultaneously for four-class classification (in the case of quadrants) and for binary classification (positive and negative arousal and valence) in order to improve generalization [23, 24]. Our results showed feasibility of our hypothesis: intra-linguistic settings (e.g., pretraining on speech in English and fine-tuning with music in English) yield better results than cross-linguistic settings (e.g., pretraining on speech in Mandarin and fine-tuning with music in English). In this study, we collect those results and further explore the impact of the pretraining stage, given that deep learning methodologies are prone to performance fluctuations due to pretraining variations.

The aim of the present study is twofold: (1) explore variations from the unsupervised learning step in order to validate our initial hypothesis, and (2) understand the impact of using different languages as pretraining data on transfer learning. Regarding (1), our previous approach might result too coarse for the extraction of meaningful features related to emotion. In the present work, we tested novel architectures of representation learning - Contrastive Predictive Coding (CPC) [25]. CPC is inspired on fundamental signal processing techniques: instead of using a loss function to remove the noise from the input representation, the architecture introduces a contrastive loss that predicts if the following observations are consecutive or not. We propose the use of this architecture given the temporal and anticipatory nature of both speech and music. With respect to (2), we performed a mixture of speech using both English and Mandarin datasets and modified the amount of pretraining data. The following section refers to the particular settings used for our study.

3. METHODOLOGY

3.1. Data

Since we used the same datasets as in our previous work [9], we briefly describe their contents, train-test distributions, and processing (see Table 1). Unlabeled speech data was used for pretraining our models. To train models on English speech, the Librispeech data set was used [26]. To train the models with Mandarin speech, the AISHELL data set was used [27]. We randomly selected a subset from each data set: 85% of the data was used to train, and 15% was used for validation during pretraining. Labeled music data was used to train our MER models. To train our English models, the 4Q-emotion data set was used [22]. To train our Mandarin models, the CH-818 data set was used [28]. Both data sets were split considering the number of classes into the following: 70% for training (85% training, 15% validation), and 30% for testing. Since the CH-818 contained only 3 hours of data, we balance equal amount of speech and music data (in hours) to the least amount available – we randomly sampled all other datasets to the same quantity (Experiment 1 - see section 4.2.1). We also evaluated the effect of the amount of pretraining data by randomly selecting 30 hours of speech data (Experiment 2 - see section 4.2.2). Finally, a mixture speech dataset was assembled by randomly selecting samples from the Librispeech (English) and AISHELL (Mandarin) datasets with equal distribution (mixture).

All datasets were processed with the libROSA library [29]: converted to mono, downsampled to 16kHz, performed a Short-Time Fourier Transform (window size: 1024 samples $\sim$ 46ms; hop size: 512 $\sim$ 23ms), and extracted a mel-scale spectrogram. The resulting mel-spectrograms had a dimensionality of 128 mel-bands by 31 time frames per second, extracted with a 50% overlap.

	Speech			Music	
Dataset	**L**	**A**	**M**	**4Q**	**CH**
Language	Eng.	Man.	Eng./ Man.	Eng./ Spa.	Man./ Can.
Annotation	-	-	-	Quad. AV	Num. AV
Size	100h	178h	278h	7.5h	2.96h
Exp. 1	3h	3h	3h	3h	3h
Exp. 2	30h	30h	30h	3h	3h

Table 1. Summary of speech and music data sets: L stands for Librispeech), A for AISHELL, M for mixture. AV refers to arousal-valence, where Quad. refers to quadrants and Num. to continuous values mapped to a quadrant. The datasets have languages in English, Mandarin, Spanish, and Cantonese.

3.2. Annotation analysis

We conducted a thorough agreement analysis on the annotations for a subset of the 4Q-emotion dataset in [19]. Regarding the CH-818 dataset, we found consistent low performance using several classification algorithms. Nonetheless, we continued to use it since it has incorporated better standards for the annotation procedure and very few datasets contain non-Western music annotated with emotions. We performed interviews with native speakers regarding the quality of these annotations [30]. The aim of these interviews was to better understand the relationship between the semantic of the lyrics and the emotion annotation. Results from the annotation analysis and these interviews are detailed in Section 4.1.

3.3. Models

We previously implemented a classifier in [9], which is a reproduction of the work by Coutinho and Schuller [10]. Since this architecture did not exhibit language-sensitive attributes, we proposed a sparse convolutional denoising auteoncoder (*SCAE*). 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 convlayers augmenting the number of filters in the encoder: 32, 64, and 128, respectively. Dropout is set to 0.25 after every double convlayer to prevent overfitting. [2]

For the present work, the Contrastive Predictive Coding architecture (*CPC*) was implemented using the same autoencoder design as our previous work (see Figure 1). While any design of autoencoder may be used with CPC, 2D convolutional neural networks are considered to extract meaningful features from audio [31]. The intuition behind CPC is the existence of underlying shared information between different parts of high-dimensional data. The autoencoder (AE) inputs a sequence of observations into a sequence of latent representations h. Since this sequence of representations contains high mutual information within neighboring samples – there is a 50% overlap in the mel-spectrograms – using an autoregressive

[2] We refer the reader to [9] for a description of the SCAE architecture.

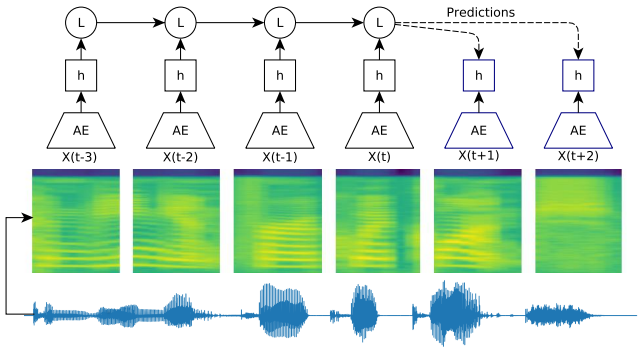

Fig. 1. CPC implementation, adapted from [25]. Mel-spectrograms have a dimensionality of 128 mel-bands x 31 frames per second.

model L allows to optimize a loss function based on the probability that future samples (x_{t+1} and x_{t+2}) are, in fact, consecutive. In our implementation, we present four consecutive mel-spectrograms to the network, and randomly assign two more that may follow or not.[3] We tested pretraining with 2 and 4 past samples ($t \leq 0$), and 2 and 4 future samples ($t > 0$). For our data, experimental results showed that using 4 past and 2 future samples improved pretraining performance. These results are consistent with methodologies such as Linear Predictive Coding, where predicting less samples results beneficial. This unsupervised learning approach proves to be faster to converge, since the optimization problem becomes a binary classification task – the presented samples are either consecutive or not. We employ a learning rate of 0.001 (4e-6 decay per epoch) with Adam optimizer of binary cross-entropy. These values resulted from Bayesian hyperparameter optimization.

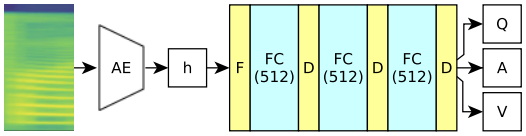

Fig. 2. Multi-task learning approach: F stands for Flatten, FC for fully connected, D for dropout.

After pretraining, transfer learning was implemented by extracting the encoder and adding a flattening layer, 3 fully connected layers each with 512 neurons, followed by a Dropout layer each (see Figure 2). To implement multi-task learning, we used three classification blocks (Q stands for Quadrants, A for Arousal, and V for Valence). Each block was made up from 2 fully connected layers (256 neurons), followed by a Dropout layer each, and a final output layer with softmax activations. Each block represents a task 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). Finally, we obtained two variations of our models: (1) fixing the weights from the encoder, and fine-tuning the network on the remaining layers at a learning rate of 0.0001 (*Feat. Ext.*), and (2) releasing the weights of the whole network and continue training with a learning rate of 0.0005 (*Full*). For the present work, we proposed two architectures (i.e.,*SCAE* and *CPC*), each with two configurations (i.e., *Feat. Ext.* and *Full*). We trained each model four times and report macro-weighted averages of metrics across experiments.

[3]https://github.com/juansgomez87/lang-sens-mer

4. RESULTS AND DISCUSSION

4.1. Annotation analysis

Our agreement study in [19] explored the impact of individual differences on emotion labeling for the 4Q-emotion dataset. Participants annotated music that was previously tagged with a particular emotion. Analysis shows overall low agreement for emotions such as *bitterness*, *fear*, *power*, *surprise*, and *transcendence*. The idea that reducing the number of categories for MER models results in better performance justifies using four quadrants in AV space. Additionally, language appears to have a central role when selecting annotators for MER experiments. We found that using annotations from participants who report understanding the lyrics consistently improve classification performance [19].

The CH-818 data set also appears to have low agreement on the annotations, which is expected for subjective annotations of emotion. Interviews conducted with native speakers suggest that: (1) there is disagreement between the semantic content of lyrics and the musical cues for emotion - when evaluating lyrics, listeners tend to use their comprehension to resolve valence, while the musical structure or acoustic features from the dataset appear homogeneous across quadrants, (2) there is disparity of annotations from fragments belonging to Q2 (high arousal and low valence) and Q4 (low arousal and high valence) - this reflects previous findings that valence evaluation is culture-specific and that agreement for these quadrants is low, and (3) interviewees described a high dependence on lyrics to assess the meaning of a song in Chinese pop culture, as opposed to the musical features - further research should be centered on the design of cross-cultural annotation methodologies.

4.2. Classifiers

4.2.1. Experiment 1

This experiment was centered on evaluating the effect of introducing a mixture of speech as pretraining data and balancing all datasets to the exact same amount of data (3 hours). In previous results, we reported evidence that intra-liguistic models (e.g. pretrained with English speech and transfer learning with music in English - *eng2eng*) result in better classification than cross-linguistic models (e.g. pretrained with English speech and transfer learning with music in Mandarin - *eng2man*). Given space constraints, we offer a general description of the results from this experiment.[4] Our results conflict with our initial hypothesis: (1) in the case testing with the CH-818 dataset, *man2man - SCAE-Feat. Ext.* performs similarly to *mix2man - SCAE-Full* – the feature extractor trained exclusively on speech is not necessarily extracting meaningful features for our classification purposes, (2) in the case of testing with the 4Q-Emotion dataset, *mix2eng - SCAE-Feat. Ext.* yields better results than other settings and all *SCAE-Full* models have similar results – we argue that "catastrophic forgetting" appears to predominate after releasing the weights resulting in uniform performance across models, and (3) both *CPC-Feat. Ext.* and *CPC-Full* models appear to show best results in cross-linguistic settings (*eng2man* and *man2eng*) and their classification scores are substantially lower than our *SCAE* models.

[4]We refer the reader to the supplementary material for results from Experiment 1.

		Test data: CH-818									Test data: 4Q-Emotion								
		man2man			eng2man			mix2man			**eng2eng**			man2eng			mix2eng		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SCAE-Feat Ext	Q	0.46	0.55	0.50	0.42	0.58	0.49	0.52	0.57	**0.51**	0.50	0.50	0.47	0.54	0.53	0.50	0.53	0.53	**0.51**
	A	0.61	0.61	0.61	0.64	0.65	0.62	0.64	0.64	**0.64**	0.67	0.66	0.65	0.68	0.67	0.66	0.68	0.67	**0.67**
	V	0.77	0.78	0.77	0.78	0.78	0.77	0.79	0.79	**0.79**	0.77	0.72	0.71	0.78	0.76	0.76	0.79	0.77	**0.77**
SCAE-Full	Q	0.53	0.58	**0.52**	0.48	0.60	**0.52**	0.54	0.58	**0.52**	0.57	0.57	**0.57**	0.57	0.57	0.56	0.56	0.56	0.55
	A	0.65	0.64	0.64	0.67	0.67	**0.67**	0.65	0.63	0.63	0.68	0.68	0.67	0.69	0.69	**0.68**	0.67	0.67	0.67
	V	0.81	0.80	**0.81**	0.81	0.80	**0.81**	0.82	0.79	0.80	0.81	0.81	**0.81**	0.80	0.80	0.80	0.80	0.80	0.80
CPC-Feat Ext	Q	0.36	0.50	0.36	0.36	0.49	0.35	0.38	0.52	**0.41**	0.34	0.29	0.19	0.30	0.29	0.20	0.42	0.35	**0.33**
	A	0.61	0.60	0.48	0.60	0.60	0.49	0.62	0.62	**0.55**	0.54	0.52	0.46	0.53	0.52	0.46	0.54	0.53	**0.51**
	V	0.70	0.67	0.58	0.68	0.65	0.55	0.71	0.70	**0.66**	0.69	0.56	0.47	0.66	0.56	0.49	0.72	0.64	**0.60**
CPC-Full	Q	0.44	0.59	**0.50**	0.44	0.59	**0.50**	0.44	0.59	**0.50**	0.47	0.48	0.47	0.45	0.45	0.45	0.51	0.51	**0.51**
	A	0.65	0.65	0.65	0.65	0.66	**0.66**	0.65	0.66	0.65	0.62	0.60	0.59	0.59	0.58	0.56	0.63	0.62	**0.61**
	V	0.79	0.79	**0.79**	0.78	0.78	0.78	0.79	0.79	**0.79**	0.78	0.78	0.78	0.79	0.79	**0.79**	0.79	0.79	**0.79**

Table 2. Summary of results for Experiment 2 and multi-task learning of tasks: Q stands for quadrants, A for arousal, and V for valence. P stands for precision, R for Recall, and F for F-score. F-scores are bold for the best scores for each classifier.

4.2.2. Experiment 2

This experiment was centered on evaluating the effect of the amount of pretraining data. We increased the amount of data from 3 to 30 hours of speech. Deep learning algorithms tend to excel in particular tasks, but usually require massive amounts of data. Classification results are summarized in Table 2. We observe two general trends: (1) both *SCAE-Full* and *CPC-Full* models exhibit similar results across all variations, (2) both *SCAE-Feat. Ext.* and *CPC-Feat. Ext.* appear to have better results using *mix2man* and *mix2eng* instances, and (3) *SCAE-Feat. Ext.* consistently shows better performance than *CPC-Feat. Ext.* ($\approx$ 10 percentage points in all scores).

With respect to (1), we argue to find a general trend of "catastrophic forgetting" mentioned in the previous section. While the feature extractor is trained only on speech, we hypothesized that it should retain emotion-related representations from speech in each language. In short, our argument is that if all tests converge to similar classification metrics, the weights of the neurons are probably converging to similar values – the pretraining step is probably not having an impact on the final classification model. In this sense, it could prove beneficial to use the supervised learning on the pretraining phase – pretraining on speech data labeled with emotion annotations could help improve the transfer learning approach. Regarding (2), we find of particular interest the fact that the mixture of speech (*mix2man* and *mix2eng*) consistently results in better performance for both *SCAE* and *CPC* models. Although this disproves our initial hypothesis, we believe this to be of particular importance, since it is consistent with results that greater diversity of the input data results in improved generalization. As to (3), given that CPC has originally been tested with raw audio and images [25], further research is needed regarding implementations using mel-spectrogram inputs.

5. CONCLUSIONS

In this work, we present in-depth results of our experiments into MER language-sensitive models. We evaluated transfer learning from speech to music, based on the experiments by [10, 9]. As proof of concept, we completed our preliminary results by extending the experiments to pretraining on a mixture of speech. Moreover, we altered the amount of pretraining data in order to test the effect of data quantity for pretraining. We evaluated the annotations from the CH-818 dataset, which consistently performs poorly for the classification task. We implemented a novel representation learning methodology (Contrastive Predictive Coding), that allows faster convergence and a more refined approach to learn a latent representation. Our findings reveal that: (1) training on a mixture of speech (*mix2man* and *mix2eng*) may improve the classification performance – we argue that the diversity of the data presented during pretraining allows the classifier to generalize better to music datasets, (2) our methodology does not appear to learn meaningful emotion-related features from speech that are transferred to emotion recognition – as future work, we recommend to use emotional speech as pretraining data, (3) further annotation analysis is needed for the CH-818 dataset – short interviews with native speakers reveal low agreement with the annotations, and (4) Contrastive Predictive Coding appears to yield poorer classification results than a denoising approach – nonetheless this architecture offers fast convergence and further investigation is advised.

Recent research has found that feature reuse in transfer learning is successful if low-level statistics from data are not fundamentally disturbed when shuffling input representations [32]. Our input representations are mel-spectrograms which are not being shuffled in this process, but perhaps the success to achieve language-sensitive MER model lies in using speech that also has emotional content. In this way, low-level statistics of "angry" sounds might still correlate from speech in a given language to music from the same culture – particularly when we overestimate our capacity to understand how others feel if they speak an unfamiliar language.

6. 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). We also thank Xiao Hu and Yi-Hsuan Yang for facilitating CH-818 for our experiments.

7. REFERENCES

[1] Markus Schedl, Arthur Flexer, and Julián Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, pp. 523–539, 2013.

[2] Aniruddh D. Patel, *Music, Language and the Brain*, Oxford University Press, 2008.

[3] Dale Purves, *Music as Biology*, Harvard University Press, London, England, 2017.

[4] Felix Weninger, Florian Eyben, Björn W. Schuller, Marcello Mortillaro, Klaus R Scherer, and Jarek 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.

[5] Shui'er Han, Janani Sundararajan, Daniel Liu Bowling, Jessica Lake, and Dale Purves, "Co-Variation of Tonality in the Music and Speech of Different Cultures," *PLoS ONE*, vol. 6, no. 5, pp. 20160, 2011.

[6] Daniel L Bowling, Janani Sundararajan, Shui'er Han, and Dale Purves, "Expression of Emotion in Eastern and Western Music Mirrors Vocalization," *PLoS ONE*, vol. 7, no. 3, pp. 31942, 2012.

[7] Catherine Stevens and Tim Byron, "Universals in Music Processing Entrainment, Acquiring Expectations, and Learning," in *The Oxford Handbook of Music Psychology*, pp. 19–31. Oxford University Press, 2016.

[8] Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 1522, no. December, pp. 1517–1522, 2019.

[9] Juan Sebastián Gómez-Cañón, Estefanía Cano, Perfecto Herrera, and Emilia Gómez, "Transfer learning from speech to music: towards language-sensitive emotion recognition models," in *Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020, pp. 136–140.

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

[11] Björn W. Schuller, *Intelligent audio analysis*, Springer, 2013.

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

[13] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS 1*, pp. 1–22, 2017.

[14] Maximilian Schmitt, Fabien Ringeval, and Björn W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proceedings of the 17th Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA, 2016, pp. 495–499.

[15] David Griol, José Manuel Molina, and Zoraida Callejas, "Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances," *Neurocomputing*, vol. 326-327, pp. 132–140, 2019.

[16] Swapnil Bhosale, Rupayan Chakraborty, and Sunil Kumar Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 7189–7193.

[17] Lisa Feldman Barrett, *How emotions are made: the secret life of the brain*, Houghton Mifflin Harcourt, 2017.

[18] Yi-Hsuan Yang, Ya-Fan Su, Yu-Ching Lin, and Homer H Chen, "Music Emotion Recognition: The Role of Individuality," Tech. Rep., National Taiwan University, 2007.

[19] Juan Sebastián Gómez-Cañón, Estefanía Cano, Perfecto Herrera, and Emilia Gómez, "Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 853–860.

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

[21] Patrik N. Juslin, *Musical Emotions Explained*, Oxford University Press, Oxford, 1 edition, 2019.

[22] Renato Panda, Ricardo Malheiro Rui, and Pedro 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.

[23] Reza Loftian and Carlos Buzo, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proceedings of the 19th Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India, 2018, pp. 951–955.

[24] Jaehun Kim, Juliá Urbano, Cynthia C S Liem, and Alan Hanjalic, "One deep music representation to rule them all? A comparative analysis of different representation learning strategies," *Neural Computing and Applications*, pp. 1–27, 2019.

[25] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proceedings of the 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[27] 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.

[28] Xiao Hu and Yi-Hsuan Yang, "Cross-Dataset and Cross-Cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 8, no. 2, pp. 228–240, 2017.

[29] Brian McFee et al., "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference (SCIPY)*, Austin, USA, 2015, pp. 18–25.

[30] Ana Gabriela Pandrea, Juan Sebastián Gómez-Cañón, and Perfecto Herrera, "Cross-dataset music emotion recognition: an end-to-end approach," in *Late breaking/Demo of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[31] Jordi Pons et al., "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 637–644.

[32] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang, "What is being transferred in transfer learning?," *CoRR*, vol. arXiv/2008.11687, 2020.