Music Speaks in Emotions (Extended Abstract)

Hortense Fong, Vineet Kumar
Yale School of Management
New Haven CT 06511, USA
{hortense.fong,vineet.kumar}@yale.edu

Music evokes emotion in listeners and emotions impact our state of mind. Being able to identify emotion in music provides information about the experienced emotion of the listener. This information is useful in a wide array of settings, ranging from music therapy to advertising. In this paper, we develop a deep neural network emotion classifier that uses different audio transformations (spectrograms) designed to capture specific music concepts and find that different emotions are best captured by different transformations. We also compare time and frequency filters with traditional black box square filters in a convolutional neural network to understand what the square filters may be capturing.

It is challenging to identify how music conveys emotion, which has generated vast amounts of psychology research on music and emotion (Johnson-Laird and Oatley 2016; Juslin and Laukka 2003; Juslin and Zentner 2001). The research suggests that different emotions are associated with different settings of music concepts, such as pitch and tempo (Johnson-Laird and Oatley 2016). For example, happy music typically has a wide range of pitches and a medium tempo while sad music typically has a small range of pitches and a slow tempo. Based on research about how humans perceive sound, engineers have developed a number of sound wave transformations to capture different music concepts. For example, the Mel spectrogram highlights the frequencies of perceptual relevance to a human listener while the short-time Fourier transform (STFT) spectrogram reflects linear frequencies.

Following the trend in many fields, deep neural networks (DNNs) have demonstrated significant performance gains in Music Emotion Recognition (MER) (Liu et al. 2017; Malik et al. 2017). While DNNs flexibly incorporate features from the data, with high dimensional data such as audio and video, conceptually developed input transformations using domain knowledge may improve performance. We examine *which* input transformations are most useful for emotion classification in music. DNNs tradeoff interpretability for performance, making it difficult to understand what features are captured (Lakkaraju, Bach, and Leskovec 2016). We therefore seek to answer the following two research questions: 1) Can insights from psychology about how humans perceive emotion in music and insights from physiology about how humans perceive sound improve the classification performance of a DNN emotion classifier? 2) Can ideas from music theory improve the interpretability of a DNN emotion classifier?

To answer the first research question, we take various two-dimensional visual representations of acoustic sound waves as inputs to a CNN to predict emotion. Since different types of visual representations better capture different concepts and different emotions can be mapped to these concepts, we explore which transformation is the most predictive for each emotion as measured by precision, recall, F1, and AUC. We consider the following five transformations, which are common to the MER literature and were developed to capture specific music concepts: short-time Fourier transform (STFT) spectrograms, Mel spectrograms, constant-Q transform (CQT) spectrograms, chromagrams (chroma), and Mel-frequency cepstral coefficients (MFCCs) (Choi et al. 2017). The horizontal dimension for these inputs is time and the vertical dimension is a transformation of frequency. For example, CQT maps to music notes and chroma maps to pitch class. The magnitude of each frequency is captured by color in a spectrogram. We also combine the various transformations using an ensemble classifier to take advantage of any diversity in information that may exist among the transformations (Dietterich 2000).

To answer the second research question, we build off the musically-motivated filters introduced by Pons, Lidy, and Serra (2016). Short and wide filters can be thought of as time filters while tall and skinny filters can be thought of as frequency filters. Time filters are designed to learn features that capture temporal variation (e.g., tempo). Frequency filters are designed to learn frequency-dependent features (e.g., pitch range). We compare time and frequency filters with more traditional black box square filters to shed light on what the black box filters can and cannot capture. In image recognition, square filters capture spatial relationships across both horizontal and vertical dimensions. However, the relationship to musical concepts is not direct here.

We test the various transformations and filters with the CAL500exp dataset, which contains 18 emotion tags for 3,223 acoustically homogeneous segments of music coming from 500 Western popular songs (Wang et al. 2014). We use

ten-fold cross-validation for model training and testing and split the data at the song-level to prevent data leakage. We find that the Mel spectrogram outperforms the STFT spectrogram in terms of F1 and AUC, indicating that incorporating domain knowledge into a DNN can aid music emotion recognition. However, transformations that most reduce the frequency resolution (i.e., CQT, chroma, MFCC) do worse than the less reduced visualizations (i.e., STFT, Mel). Regarding the musically-motivated filters, the classifiers generally learn more from the frequency filter than the time filter. The square filter outperforms both time and frequency filters for the less reduced visualizations (STFT, Mel), suggesting the square filter captures local dependencies across both frequency and time that impact emotion recognition the other two filters cannot capture. However, for CQT, chroma, and MFCC, square filters do worse than frequency filters, suggesting these transformations remove local dependencies useful in emotion recognition. When determining classifier design, it is important to consider the interaction between the input and the filter. We find that different emotions are best captured by different transformation-filter combinations. Out of the set of transformations and filters we analyze, Mel spectrograms with square filters on average yield the best classification performance as measured by F1.

Our Mel square DNN classifier outperforms the SVM classifier developed by the creators of the CAL500exp dataset (Wang et al. 2014) in terms of F1. The ensemble classifier, which takes the majority vote of the classifiers for each of the transformations, performs slightly better than the Mel square combination. A classifier that uses the best transformation-filter combination for each emotion outperforms the ensemble classifier.

Our empirical results suggest that despite the flexibility of DNNs, domain knowledge can enable the development of better performing emotion classifiers. In addition, filters designed to capture specific concepts helps us understand what black box filters are and are not capturing at a high level (time vs. frequency). Further research should delve deeper into what musical concepts, such as tempo or chord consonance, are distilled out and what additional concepts the square filters incorporate. Whereas we focus on emotion recognition in music, the same methods and ideas could be used more broadly with voice data.

References

Choi, K.; Fazekas, G.; Cho, K.; and Sandler, M. 2017. A tutorial on deep learning for music information retrieval. *arXiv preprint arXiv:1709.04396* .

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer.

Johnson-Laird, P. N.; and Oatley, K. 2016. Emotions in music, literature, and film. *Handbook of emotions* 82–97.

Juslin, P. N.; and Laukka, P. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin* 129(5): 770.

Juslin, P. N.; and Zentner, M. R. 2001. Current trends in the study of music and emotion: Overture. *Musicae scientiae* 5(1_suppl): 3–21.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–1684.

Liu, X.; Chen, Q.; Wu, X.; Liu, Y.; and Liu, Y. 2017. CNN based music emotion classification. *arXiv preprint arXiv:1704.05665* .

Malik, M.; Adavanne, S.; Drossos, K.; Virtanen, T.; Ticha, D.; and Jarina, R. 2017. Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv preprint arXiv:1706.02292* .

Pons, J.; Lidy, T.; and Serra, X. 2016. Experimenting with musically motivated convolutional neural networks. In *2016 14th international workshop on content-based multimedia indexing (CBMI)*, 1–6. IEEE.

Wang, S.-Y.; Wang, J.-C.; Yang, Y.-H.; and Wang, H.-M. 2014. Towards time-varying music auto-tagging based on CAL500 expansion. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.