

UNTANGLING INDICES OF EMOTION IN MUSIC USING NEURAL NETWORKS

Dorien Herremans¹, Kin Wai Cheuk^{1,2}, Yin-Jyun Luo^{1,2}, and Kat Agres^{2,3}

¹Singapore University of Technology and Design

²Social & Cognitive Computing Department, Institute of High Performance Computing, A*STAR

³Yong Siew Toh Conservatory of Music, National University of Singapore



ABSTRACT: Music and emotions are intrinsically connected [8, 7]. We explore how recent advances in machine learning can be used to predict perceived emotion in music. Models of emotion in music have a large variety of applications in domains such as healthcare, entertainment, and musicology. First, we explored at the task of static emotion prediction, i.e., predicting one emotion (in terms of valence and arousal), from an entire song. This was achieved using Triplet Neural Networks (TNNs) [9], which learn a compact latent space representation of audio features that is optimized to cluster songs per emotion category. Second, dynamic emotion prediction, i.e. predicting how emotion changes throughout a song over time, was assessed using a variational autoencoder (VAE) [3]. The resulting models from both tasks are able to efficiently disentangle audio features to create effective new representations for emotion prediction.

Static emotion prediction

TNNs were initially introduced for classification, not for regression. We propose a mechanism that allows them to work in a regression context [1]. By using this method to define positive samples (songs with the same emotion) and negative samples (songs with a different emotion), we train the network (see Fig. 1) to learn a new, low dimensional feature representation that disentangles musical pieces based on the perceived emotion.

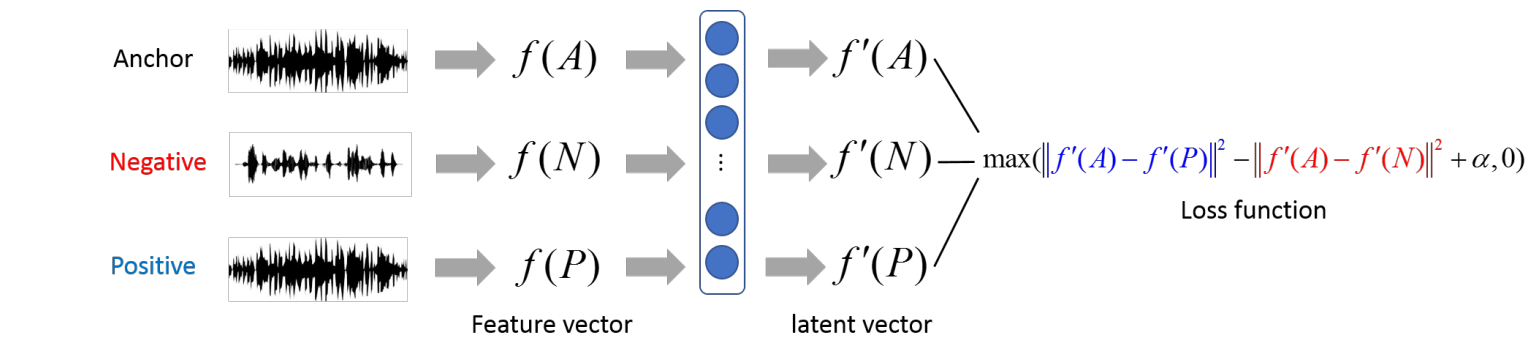


Fig. 1: TNNs learn embeddings which minimize the distance between positive samples, and maximize the distance between negative samples.

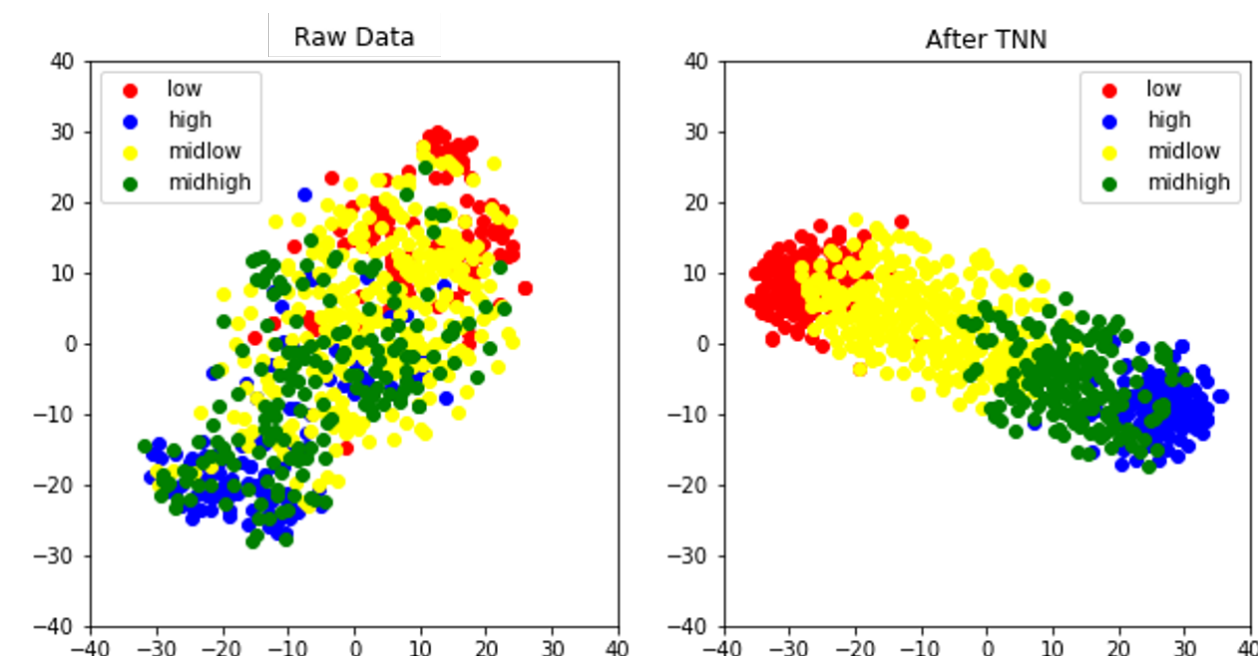


Fig. 2: Visualisation of original (left) and learned (right) features.

Figure 2 shows a t-SNE projection [6] of the original features of songs in the MediaEval Dataset [10], as well as the learned features by the TNN. It is apparent that the new representation disentangles the features such that they are able to better distinguish different levels of valence and arousal (marked in color).

This new representation can then be used as input to traditional classifiers such as support vector machines (SVM). We tested our approach on the MediaEval Dataset [10], which resulted in a 90% feature dimension reduction with a 9% improvement in valence prediction and 4% improvement in arousal prediction with respect to a baseline SVM model. When principal component analysis (PCA), a popular dimension reduction method, was combined with the same classifier (SVM), the results were far worse than for our TNN approach.

Model	Valence	Arousal
SVM (6,669 dim.)	0.347	0.614
PCA-SVM (600 dim.)	0.087	0.224
TNN-SVM (600 dim.)	0.378	0.638

Tab. 1: Results (R² score) for MediaEval 2013: 774 songs with 6,669 features each.

Dynamic emotion prediction

In dynamic emotion prediction, we continuously estimate the perceived emotion (every 0.5s). Since what is previously heard (context) may influence current ratings of perceived emotion, we opted to use a long-short term model (LSTM), combined with a variational autoencoder (VAE). The strength of VAEs is that they are able to learn disentangled latent representations [2], which have exhibited superior performance given baseline features for emotion classification in speech [5, 4].

The VAE was given Mel-spectrograms as input, and was pretrained on a collection of over 2,800 audio files from CAL500 (493), GTZAN (1,000), and DEAM 2015 (1,313). No emotion labels were needed for this pretraining, as it was unsupervised.

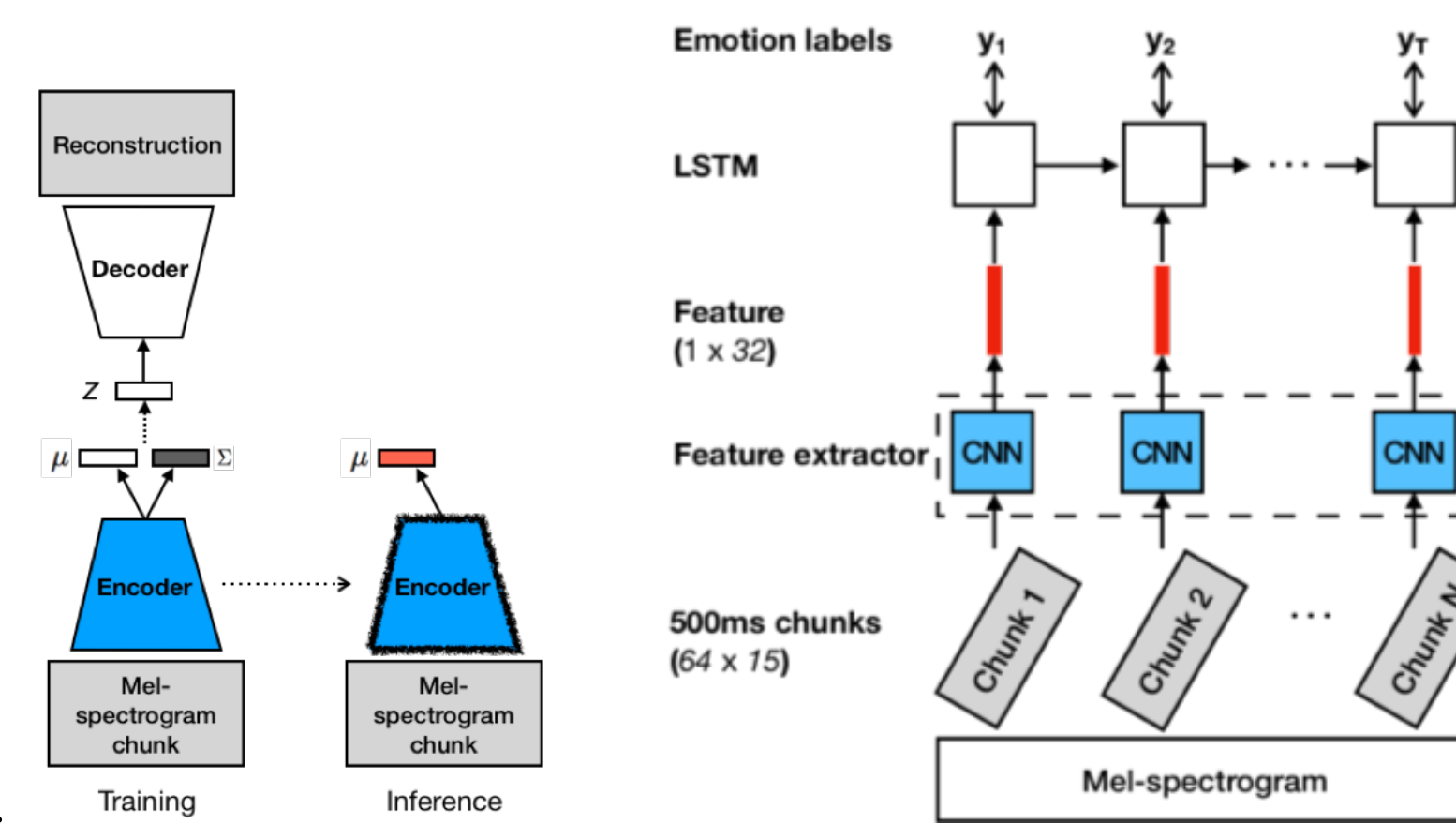


Fig. 3: The VAE (left) and LSTM (right)

The DEAM dataset was used to evaluate our proposed VAE-LSTM dynamic emotion prediction system. Our model was compared to a state-of-the-art model from Xu *et al.* [11], which uses extreme learning machines and recurrent neural networks. We also compare our system with LSTM-based approaches utilising different input processing: an autoencoder (AE-LSTM) and a convolutional neural network (CNN-LSTM).

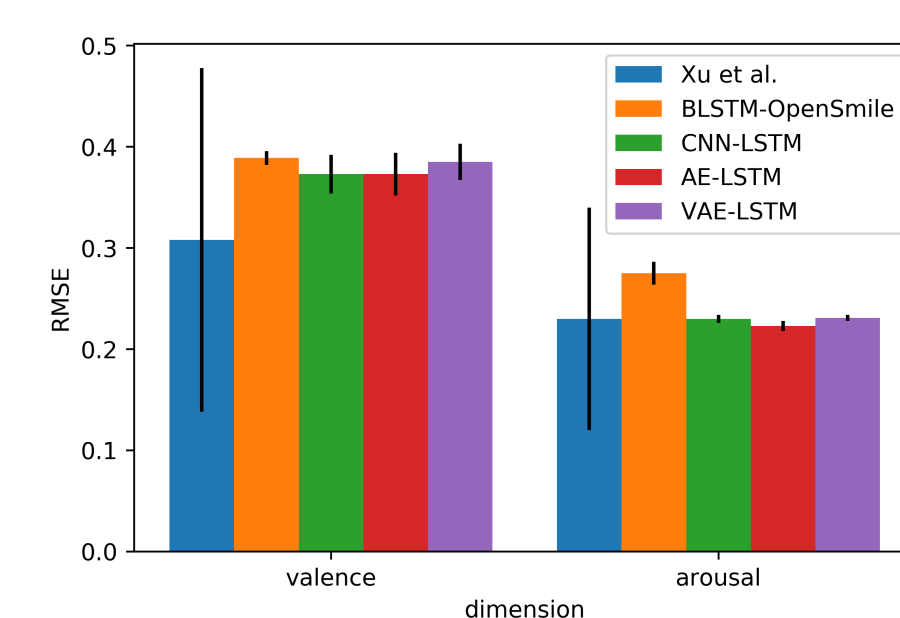


Fig. 4: Emotion prediction results in terms of root mean square error (RMSE).

For valence, with a latent representation of 32 dimensions, the AE-LSTM and VAE-LSTM achieve comparable performance with the BLSTM with 260-dimensional OpenSmile features. For arousal, the AE-LSTM performs best with a much lower variance and fewer features compared to Xu *et al.* (6,373 song-level and 130 segment-level features).

User Interface for Recommendation

Our models have been implemented in a web interface (see screenshots below), which allows the user to:

- pick a point in valence-arousal space to generate a list of songs that match this emotion (left);
- draw a valence and arousal profile and find songs that match the desired evolution of emotion as closely as possible (right).

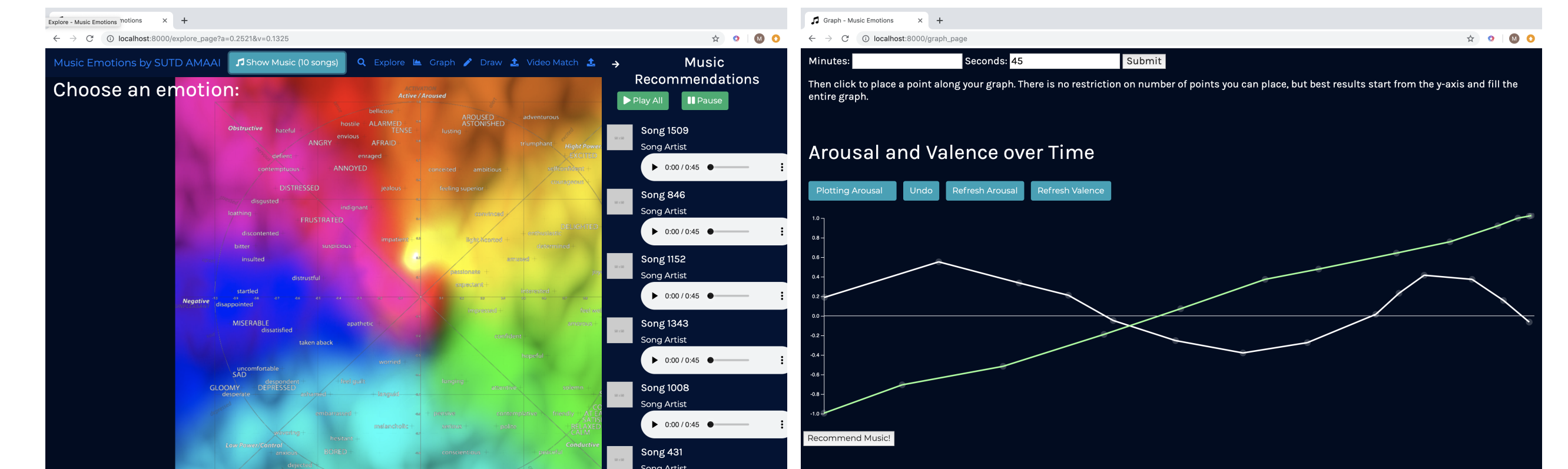


Fig. 5: Screenshot of the aiMuVi system.

Conclusions

Our proposed models are able to efficiently predict perceived emotions from music. In follow up research, we are creating similar models for video. This will result in a music recommendation system for videos based on perceived emotion. We are currently implementing self-attention networks to further increase the accuracy of our predictions. More info on our project can be found at dorienherremans.com/emotion.

References & acknowledgements

- [1] "Blacklisted speaker identification using triplet neural networks". In: *The 1st Multi-target speaker detection and identification Challenge Eval.* (2018).
- [2] I. Higgins et al. "Beta-vae: Learning basic visual concepts with a constrained variational framework". In: *Int. Conf. on Learning Representations.* 2017.
- [3] D. P. Kingma and M. Welling. "Auto-encoding variational bayes". In: *Int. Conf. on Learning Representations.* 2014.
- [4] S. Latif et al. "Variational autoencoders for learning latent representations of speech emotion: a preliminary study". In: *Proc. of Interspeech* (2018), pp. 3107-3111.
- [5] R. Lu et al. "Improving emotion classification through variational inference of latent variables". In: *2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- [6] L. v. d. Maaten and G. Hinton. "Visualizing data using t-SNE". In: *J. Mach. Learn. Res.* 9:Nov (2008), pp. 2579-2605.
- [7] L. B. Meyer. *Emotion and meaning in music*. University of chicago Press, 2008.
- [8] C. C. Pratt. "Music as the language of emotion." In: *The Library of Congress.* 1952.
- [9] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015, pp. 815-823.
- [10] M. Soleymani et al. "1000 Songs for Emotional Analysis of Music". In: *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia.* CrowdMM '13, Barcelona, Spain: ACM, 2013, pp. 1-6. ISBN: 978-1-4503-2396-3.
- [11] M. Xu et al. "Multi-Scale Approaches to the MediaEval 2015" Emotion in Music" Task." In: *MediaEval.* 2015.

This research has received funding from SMART-MIT under grant no. ING-000091 ICT.