HAJIM
SCHOOL OF ENGINEERING
& APPLIED SCIENCES
UNIVERSITY of ROCHESTER

Short-Time Emotion Tracker for Music

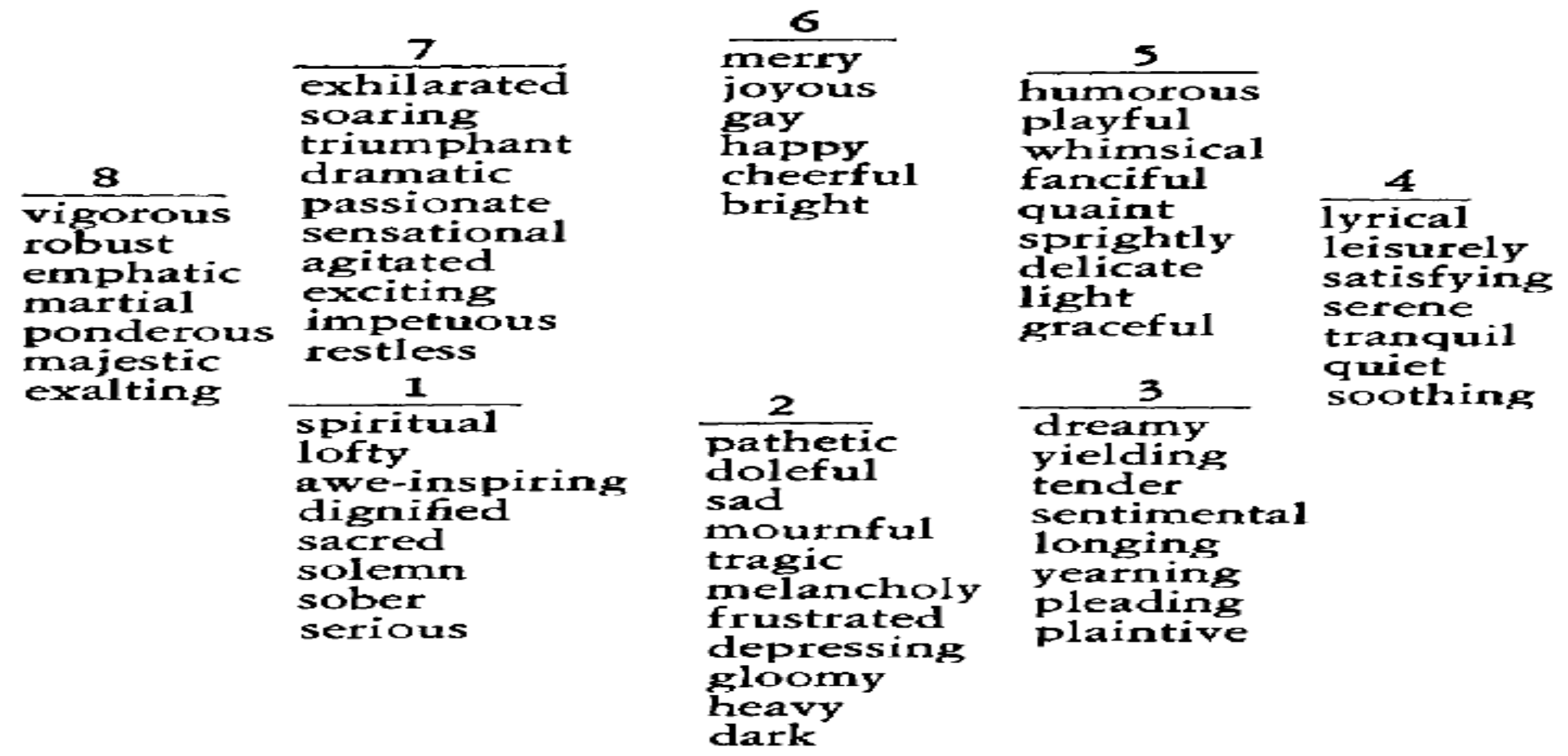
Yoon mo Yang
Department of Electrical and Computer Engineering

Abstract

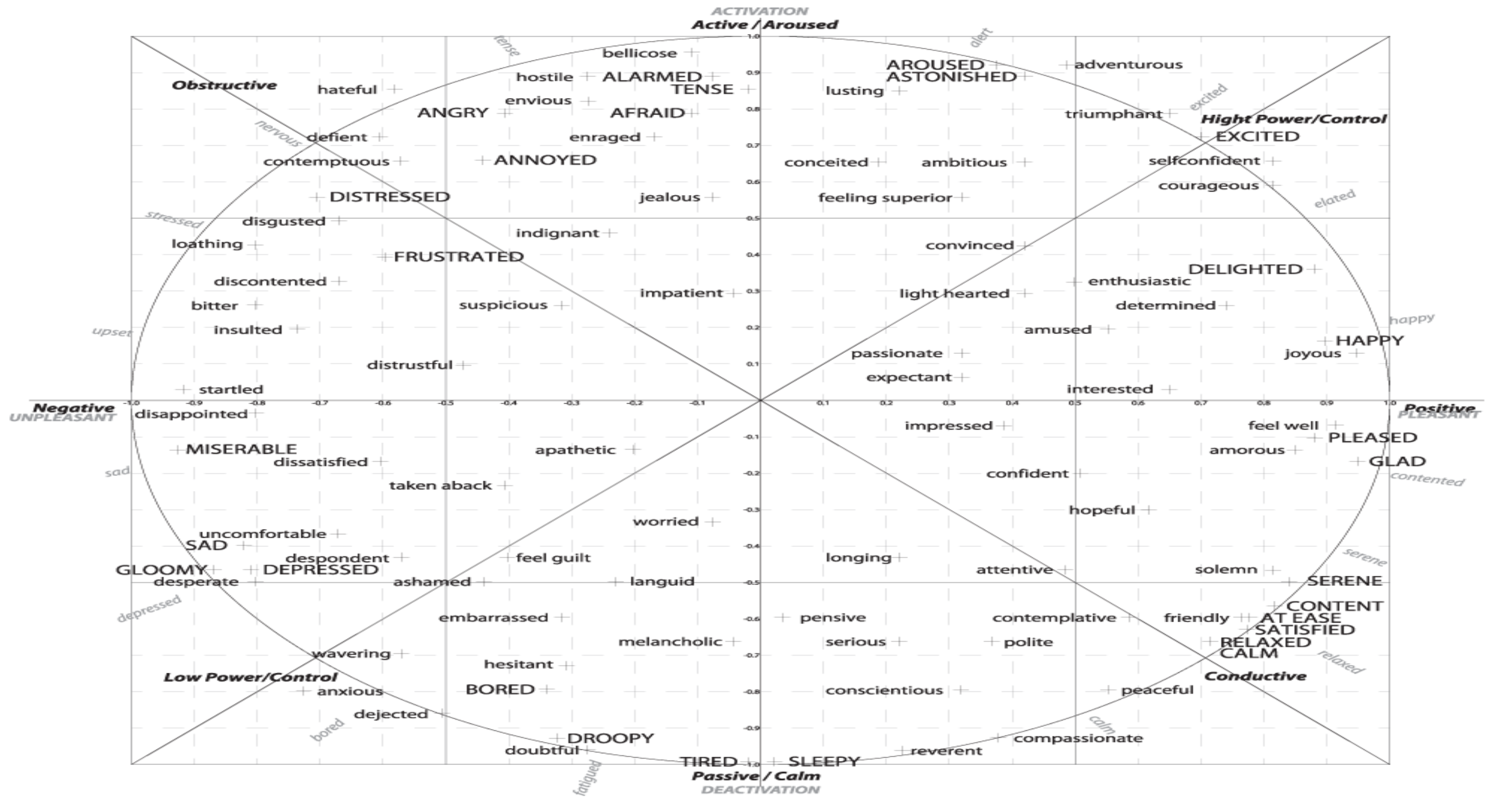
Music is not only meant for entertainment, but has the power to deliver different emotions to people. From the time machine learning and deep learning came under the spotlight, researchers have tried to use these techniques to retrieve a variety of information from music – genre, instrumentation and even emotion. In this paper, we propose a Convolutional Auto-encoder[1] that can extract arousal and valence values that represent the dynamics of emotions of a given song. The results of the proposed architecture are compared with a baseline model.

Emotion and Music

- **Two views on emotion and music**
- **Emotivists:** Music induces real emotional responses in the listener.
- **Cognitivists:** Music simply expresses an emotion.
- **Two representations of emotion**
- **Categorical psychometrics:** Utilize some set of emotional adjectives (tags) based on their relevance to given music. Ex. Hevner's 8 groups of emotions with 66 adjectives.



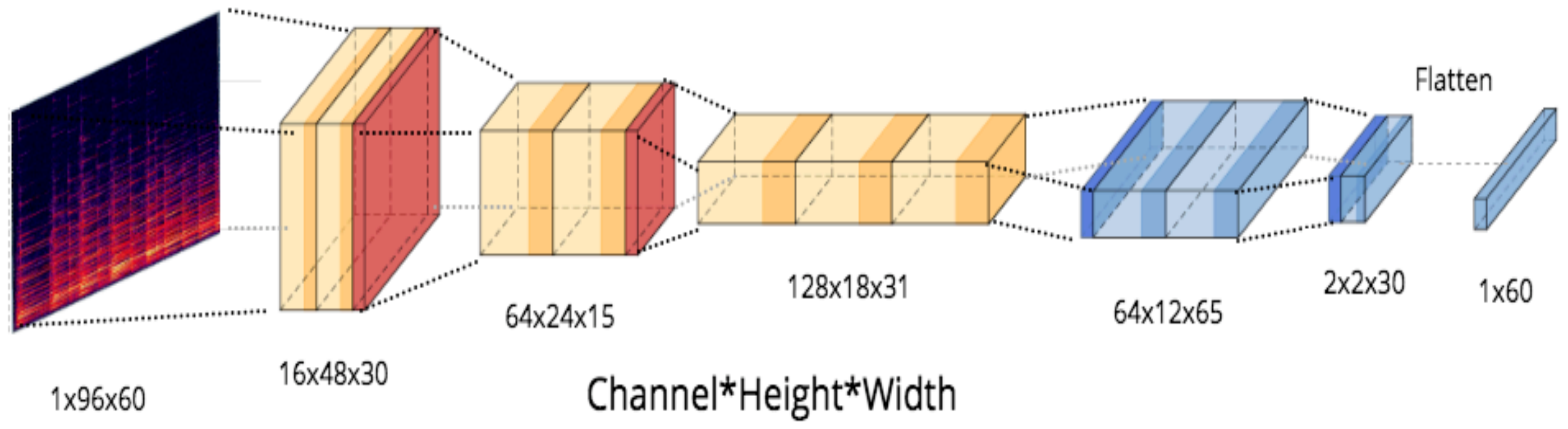
- **Dimensional psychometrics:** Mood is scaled and measured by simple multidimensional metrics. Ex. Russell and Thayer's two *Valence-Arousal* space[2].
- Arousal: Intensity, ranging high-to-low.
- Valence: An appraisal of polarity, ranging positive-to-negative.



Dataset

- Emotion in Music Database[3]: contains 744 songs and 619 of them are for development and 125 of them are for evaluation.
- Each song has 45 seconds. From 15 seconds, the continuous annotations of arousal and valence values are annotated by 300 crowdworkers on Amazon Mechanical Turk.
- Annotation interface: used a mouse to annotate arousal and valence continuously.
- The sampling rate of the annotations is 2 Hz so each song has a pair of 60 annotations for 30 seconds.

Proposed Method

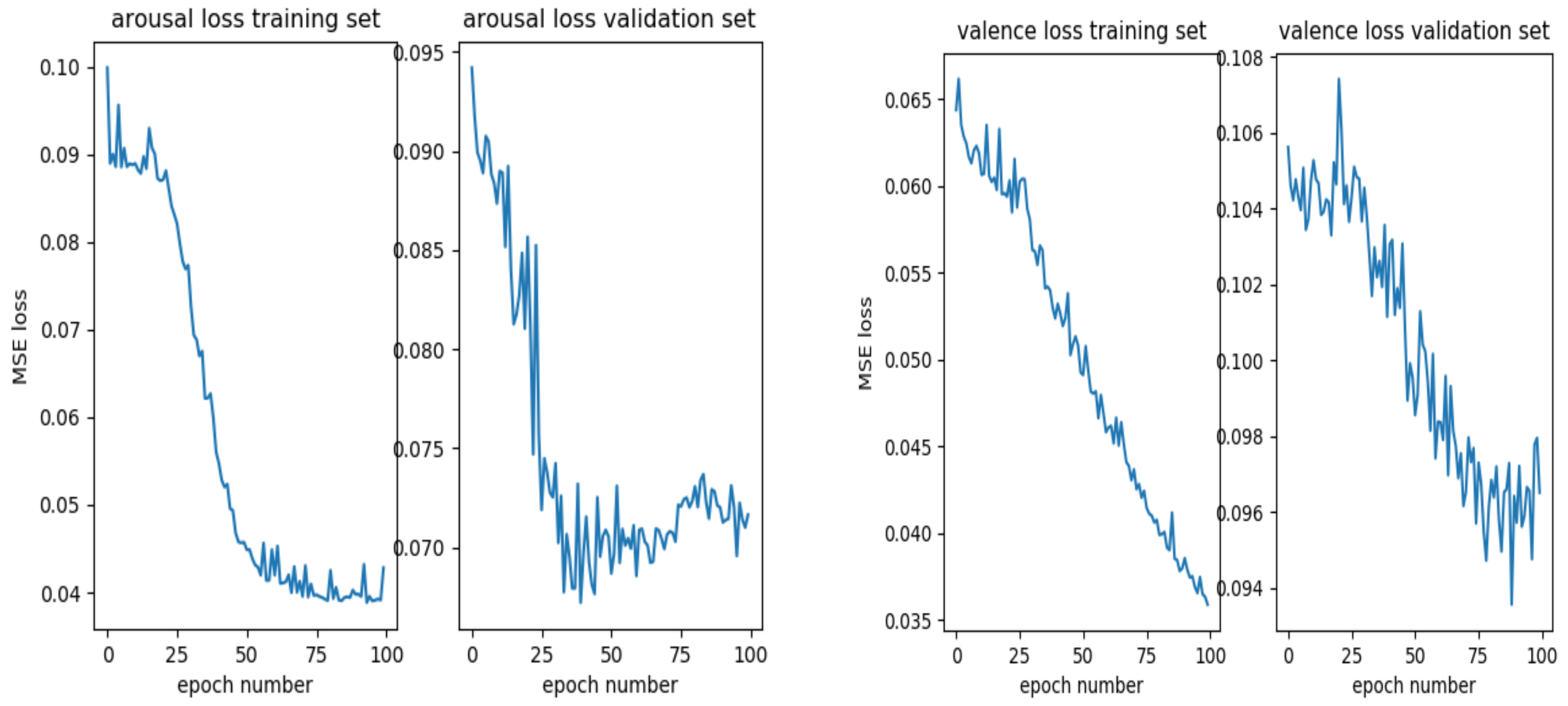


- The dataset was pre-processed to obtain faster training process.
- Mel-spectra as input features.
- Extracted mel-band features from 500ms with 60ms of window length and 30ms of hop size (≈ 50% overlap) by using the librosa python library.
- Took average over 17 time frames to get 60 time frames.
- Pytorch was utilized to implement the proposed model.
- Each convolutional layer is followed by a Leaky ReLU activation function.
- The output of the last convolutional layer gets downsampled by a max pooling layer which is followed by a tanh activation function.
- Dropout is used for each convolutional layer with 75% to avoid overfitting.
- Adam optimizer is utilized with learning rate 0.0001.

Layer	Kernel size, stride, padding
CNN1	[7,7], [2,2],[3,3]
CNN2	[5,5],[2,2],[2,2]
CNN3 (Transposed)	[5,5],[1,2],[5,1]
CNN4 (Transposed)	[7,7],[1,2],[6,1]
CNN5 (Transposed)	[7,1],[1,1],[1,6]
Max Pooling	[1,1],[14,2]

Results

- Results on 610 songs (**training set**) and 9 songs (**validation set**).



- Results on **test set**: the **baseline model** (two convolutional layers and three fully-connected layers).

	Arousal	Valence
MSE	0.1088	0.0908
RMSE	0.330	0.301

- Results on **test set**: the **proposed model**.

	Arousal	Valence
MSE	0.0965	0.0606
RMSE	0.310	0.246

Conclusion

- Proposed method outperforms our baseline model on the test set on both arousal and valance values.
- However, our method is sensitive to hyperparameters and easily gets overfitted.
- RNN layer is not yet explored.
- Data augmentation (ex. Adding Gaussian noise) is necessary to have robustness .

References

[1] Jonathan Masci, Ueli Meier, Dan Cire¸san, and J¨urgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In Proceedings of the 21th international conference on Artificial neural networks - Volume Part I, ICANN'11, pages 52–59, Berlin, Heidelberg, 2011. Springer-Verlag.
[2] R. E. Thayer. The Biopsychology of Mood and Arousal . Oxford University Press, New York, 1989.
[3] E. M. Schmidt C.-Y. Sha M. Soleymani, M. N. Caro and Y.-H. Yang. 1000 songs for emotional analysis of music. In Proceedings of the 2nd ACM InternationalWorkshop on Crowdsourcing for Multimedia, pages 1–6, 2013.