ScripTONES: Sentiment-Conditioned Music Generation for Movie Scripts

Vishruth Veerendranath[1,2], Vibha Masti[1,2], Utkarsh Gupta[1], Hrishit Chaudhuri[1], Gowri Srinivasa[1]

[1]*PES Centre for Pattern Recognition, PES University*
[2]*Carnegie Mellon University*

Demo: Paper:

Problem Statement

- Film score essential for cinematic experience
- Automated system for generating emotion-aligned symbolic music
- Two components:
  - Movie script (text) encoder
  - Music generator decoder
- Quantifying emotion: valence-arousal [1]
  - Text: NRC VAD lexicon [2]
  - Music: EMOPIA [3]
    - Piano midi snippets tagged with quadrant of VA [4]

Attribute Vector Arithmetic

- Attribute vector arithmetic in VAEs extended to MusicVAE
- Four attribute vectors: high valence (z_{vh}), low valence (z_{vl}), high arousal (z_{ah}),and low arousal (z_{al})
- Averaging out latent vectors of EMOPIA samples encoded with MusicVAE.

$$z_{ec} = \begin{cases} |V| * z_{vh} + |A| * z_{ah} & (V \geq 0, A \geq \alpha) \\ |V| * z_{vh} + |A| * z_{al} & (V \geq 0, A < \alpha) \\ |V| * z_{vl} + |A| * z_{ah} & (V < 0, A \geq \alpha) \\ |V| * z_{vl} + |A| * z_{al} & (V < 0, A < \alpha) \end{cases}$$

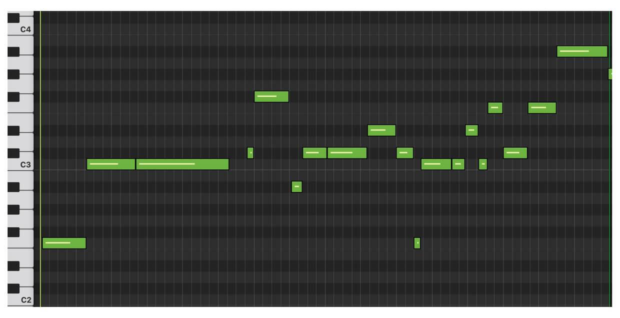
Fig. 4: Original music

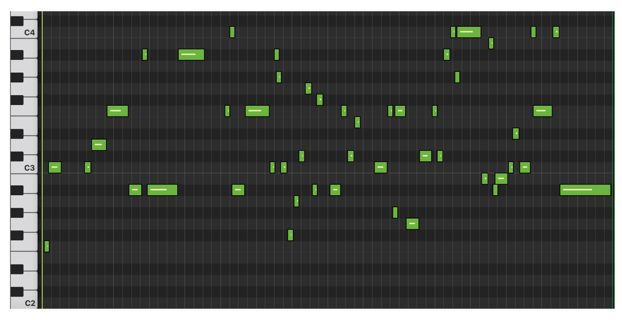
Fig. 5: Modified with increased valence

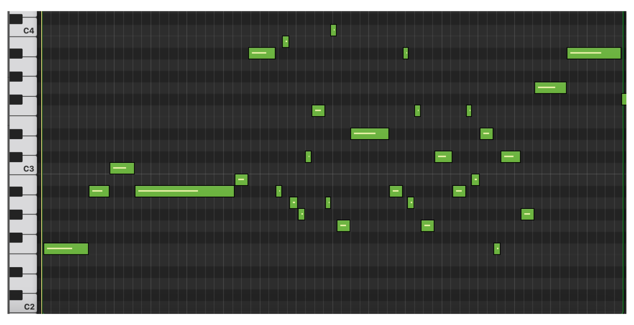
Fig. 6: Modified with increased arousal

Discrete Regularization

- Improving sentiment conditioning in VAEs
- Regularizing 2 latent dimensions to encode valence & arousal
- Finetune pretrained FIGARO [5] VAE on EMOPIA data as per loss below

$$L_{reg_{disc}} = BCE(v_{pred}, v_{gt}) + BCE(a_{pred}, a_{gt})$$

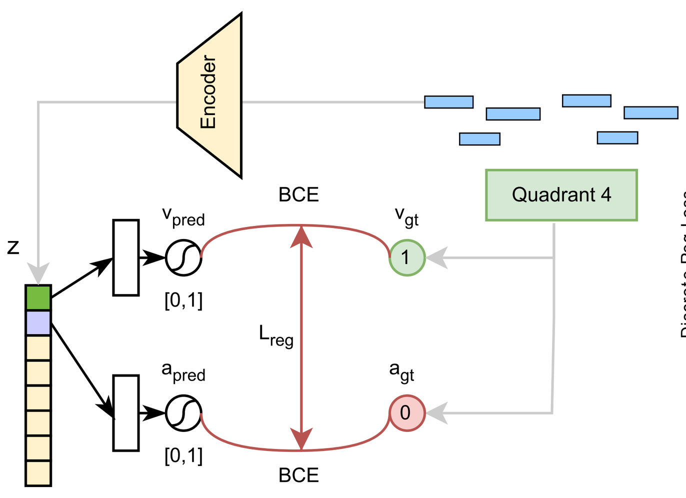

Fig. 7: Discrete Regularization

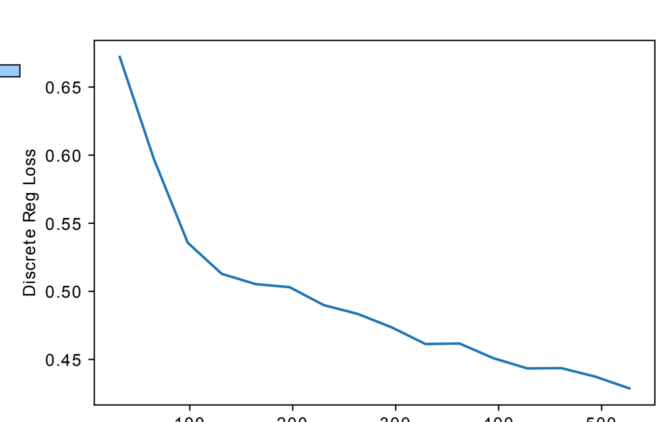

Fig. 8: Loss Plot of FIGARO on EMOPIA

Methodology

- Two-phase pipeline for sentiment-conditioned music generation
- Sentiment analysis of movie script
  - Unweighted average
- Conditional music generator – experimented with two different techniques
  - EMOPIA-CWT (Transformer-based) (Fig. 1 2(b))
  - MusicVAE with attribute vector arithmetic (Fig. 1 2(a))

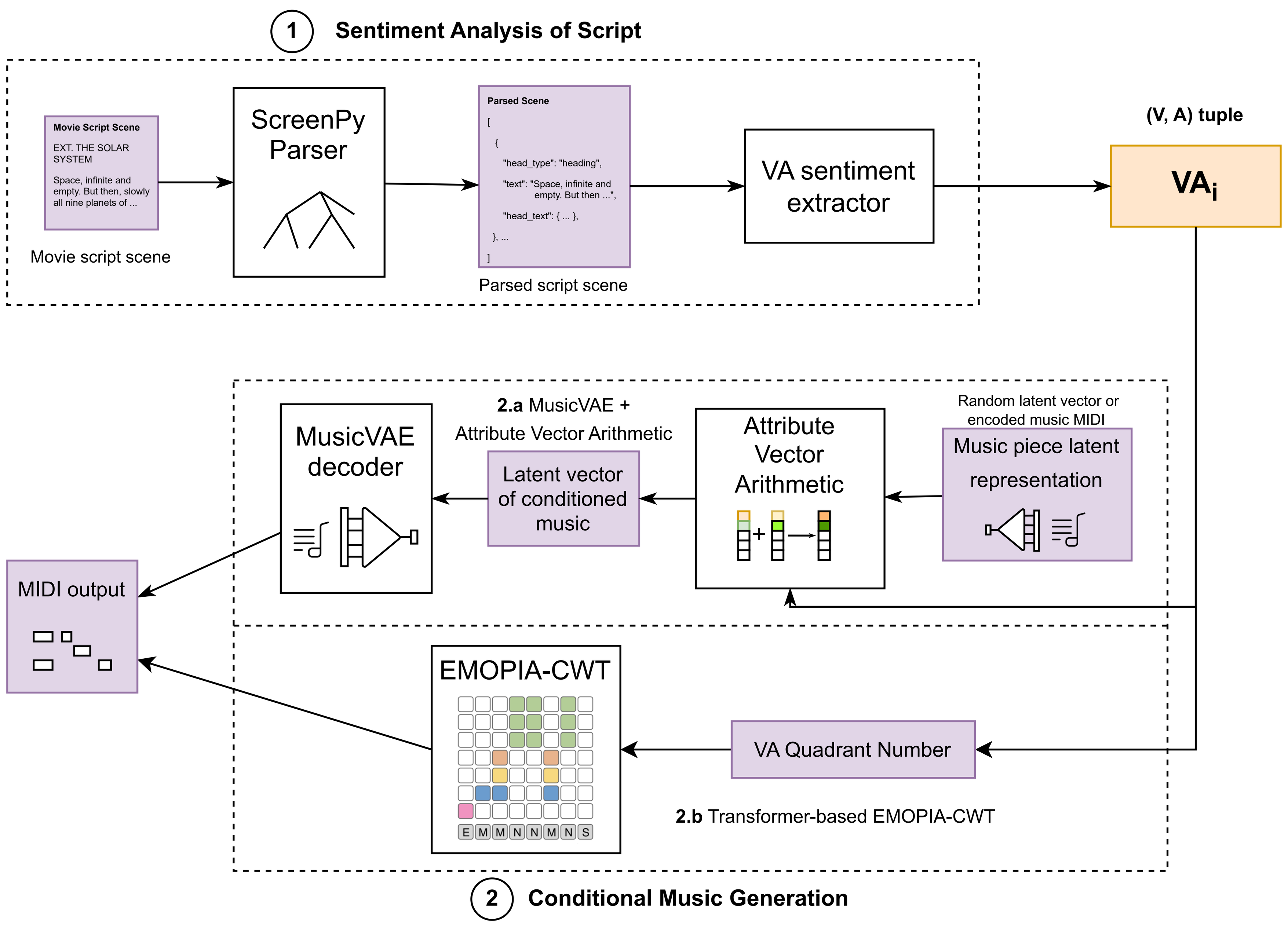

Fig. 1: ScripTONES pipeline

User Study Evaluation

- Survey on 31 users with varying musical knowledge
- 3 different movie scene demos, two different music pieces (EMOPIA-CWT and MusicVAE)
- For each piece, rate on a scale of 1-4:
  1. Valence/positivity **2.** Arousal/excitement **3.** Overall mood fit

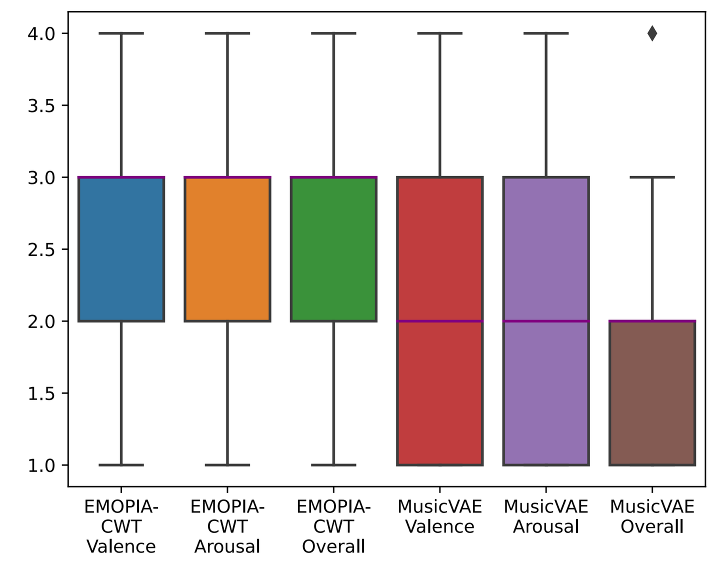

Fig. 2: Box-plot of user ratings for E-CWT & MVAE models

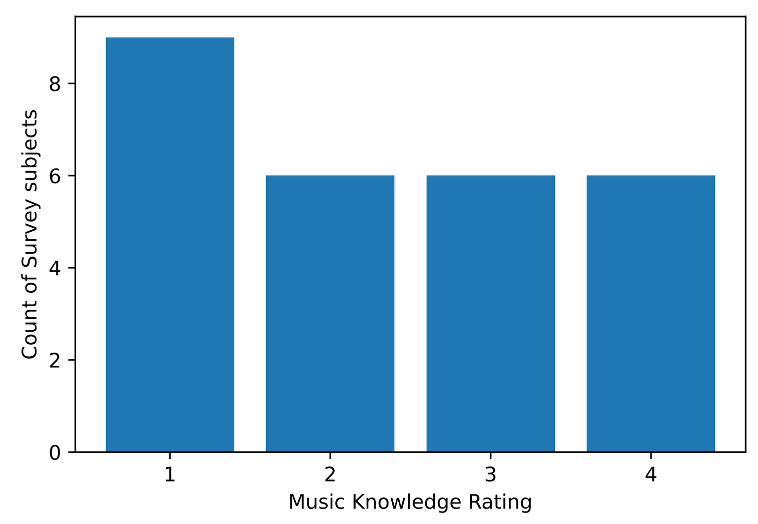

Fig. 3: Music Knowledge rating of survey subjects

Table 1: Average ratings of match between generated music and film scene

Attribute Rated	E-CWT	MVAE
Valence	**2.62**	1.96
Arousal	**2.44**	1.92
Overall Mood Fit	**2.48**	1.86

Table 2: User evaluated scene-wise overall mood fit ratings on a scale of 1-4

Scene Number	E-CWT	MVAE
Scene 1	**2.52**	1.58
Scene 2	1.87	**2.17**
Scene 3	**3.06**	1.84

References

1. Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology, 17(3):715–734, 2005.
2. Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018.
3. Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. arXiv preprint arXiv:2108.01374, 2021
4. Yu, Liang-Chih, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. "Building Chinese affective resources in valence-arousal dimensions." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 540-545. 2016.
5. Rütte, Dimitri von, Luca Biggio, Yannic Kilcher and Thomas Hofmann. "FIGARO: Controllable Music Generation using Learned and Expert Features." International Conference on Learning Representations (2023).