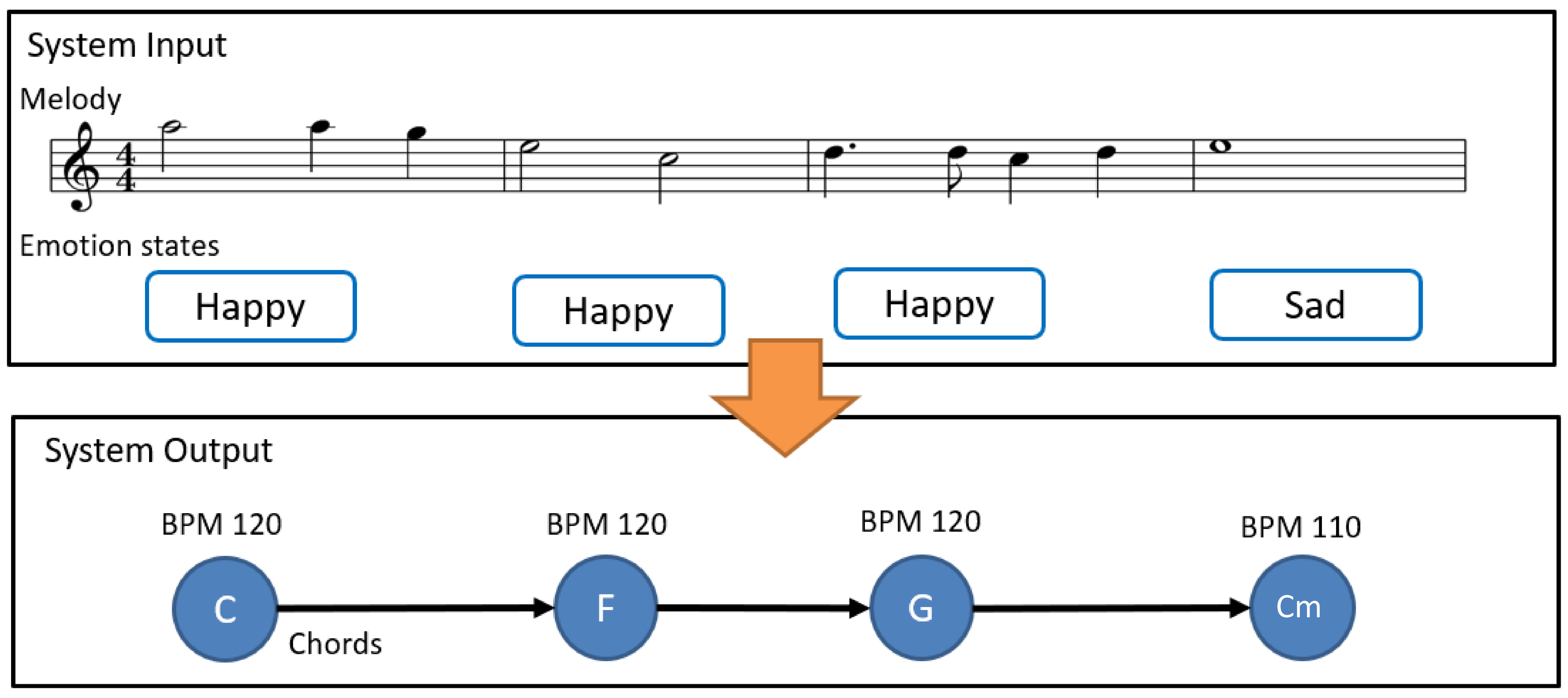
Emotion-driven Harmonisation and Tempo Arrangement of Melodies Using Transfer Learning

Takuya Takahashi Mathieu Barthet

Centre for Digital Music, Queen Mary University of London

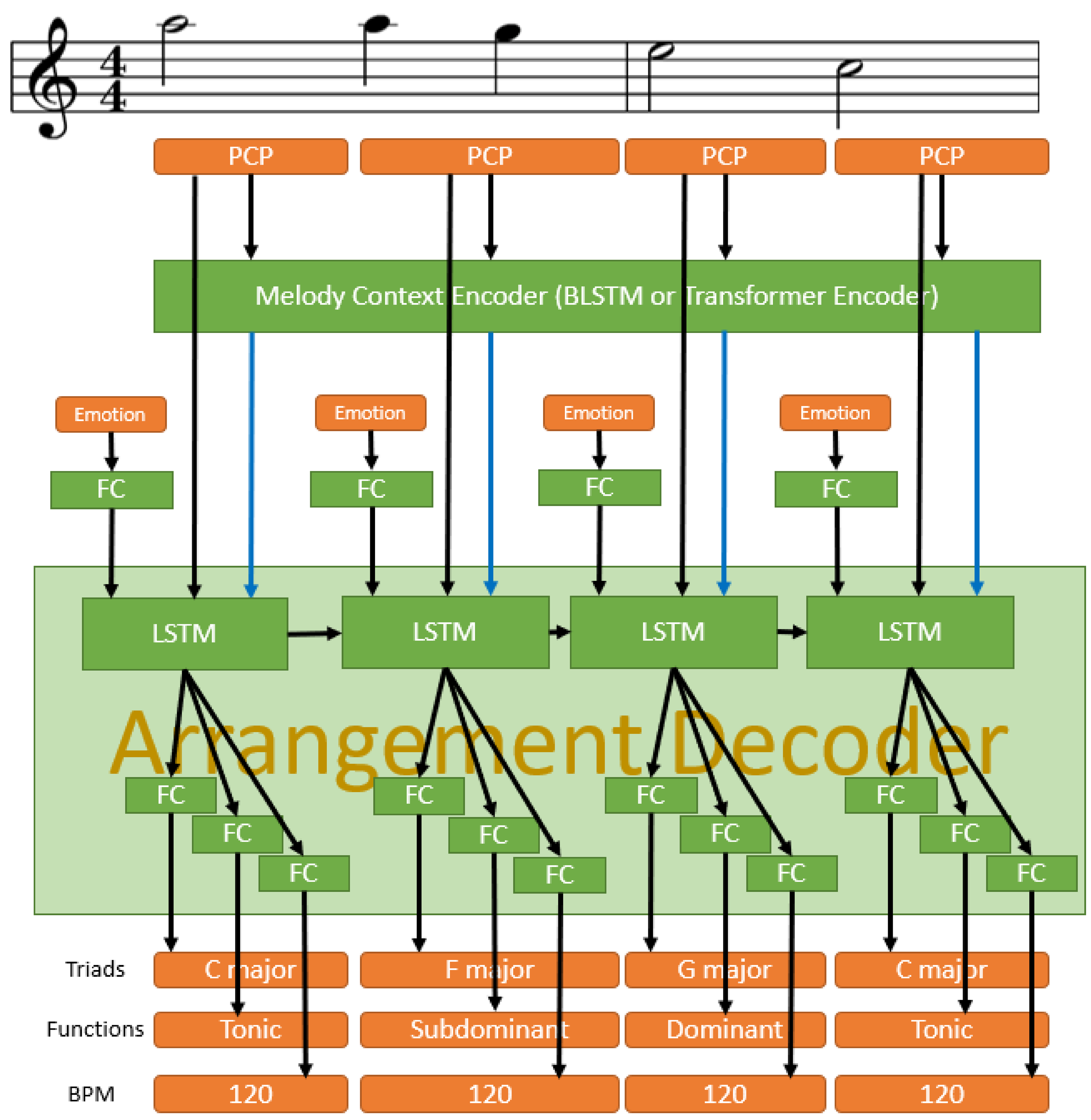
Overview

- **Objective**
  - Building a system that automatically arranges the melody to express the specific emotion
    - Harmonize melody
    - Change tempo



- **Contributions**
  - Proposed architecture for learning the relationships between symbolic melodies, chord progressions, tempo and expressed emotions
    - Can generate emotion-driven arrangements faster than ever before
  - A dataset of 4000 symbolic scores and emotion labels was gathered by expanding the HTPD3 dataset with mood tags from last.fm and allmusic.com
  - Evaluation experiments prove the effectiveness of the transfer learning and show the impact of the methods of quantifying emotions

DL Architecture



Melody Context Encoder

- A representation of melodies as input and outputs a 128-dimensional embedding at every time unit
- Use Bi-LSTM or Transformer encoder as Melody Context Encoder

Arrangement Decoder

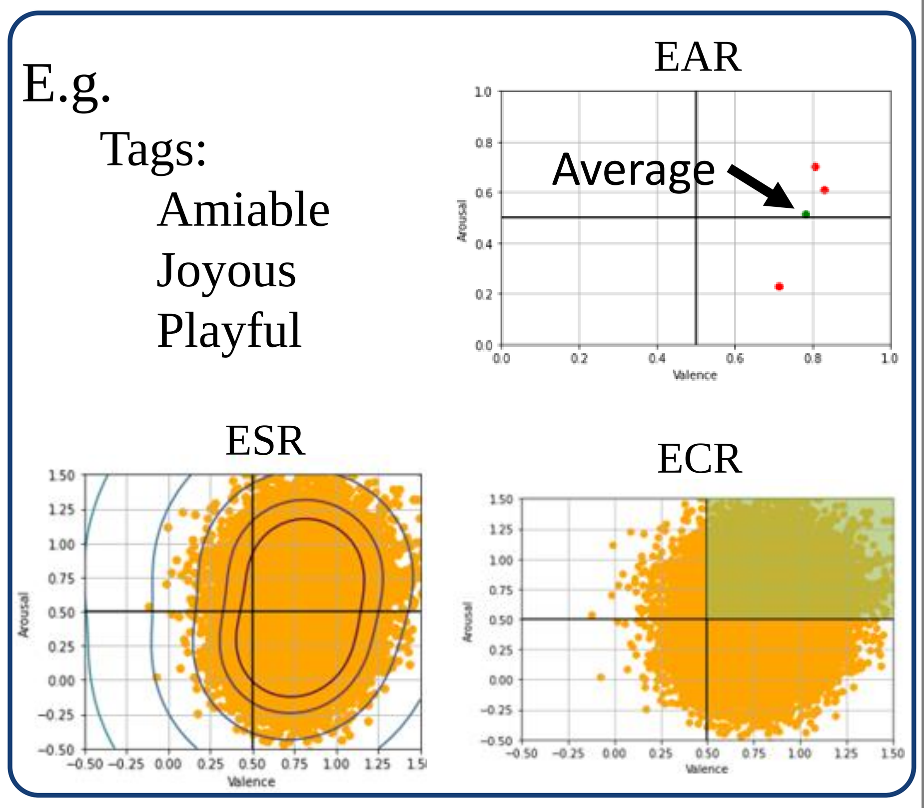
- Output arrangement information such as chords and tempi based on melody embedding and emotion
- Forward propagation only to reduce computational costs and to allow inference based only on historical information for near real-time applications

Transfer Learning Strategy

- Encoders are pre-trained using music examples without emotion labels
- The pre-trained encoders (weights are fixed) and randomly initialized decoders are concatenated and retrained only for the subset of tracks with emotion labels

Music Emotion Quantification

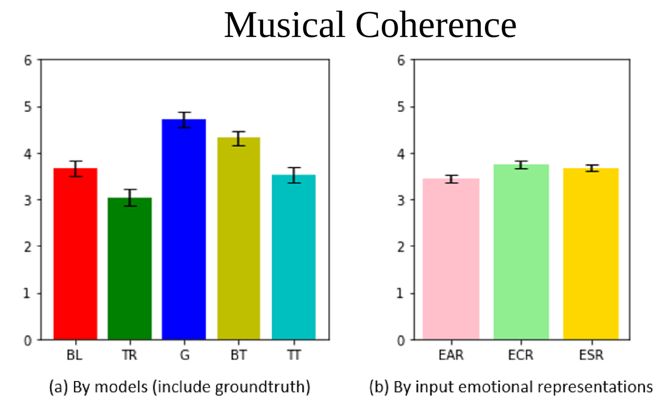
- **Convert emotional tags into numerical expression**
  - Use statistics (Warriner et al. 2013) on arousal-valence scores for emotional words
    - Russell's circumplex model (Russell 1980)
      - Arousal: How excited (aroused)?
      - Valence: How positive or negative (valence)
- **Handling of multiple emotion tags**
  - Emotion Average Representation (EAR)
    - Using the average of the arousal-valence scores of all tags
  - Emotion Surface Representation (ESR)
    - Using GMMs to represent surfaces on arousal-valence spaces
    - Random sampling of 10000 points from the mean and standard deviation for all emotional words and perform GMM
  - Emotion Category Representation (ECR)
    - The AV space quadrant with the highest AV annotations determines the emotional category of the music



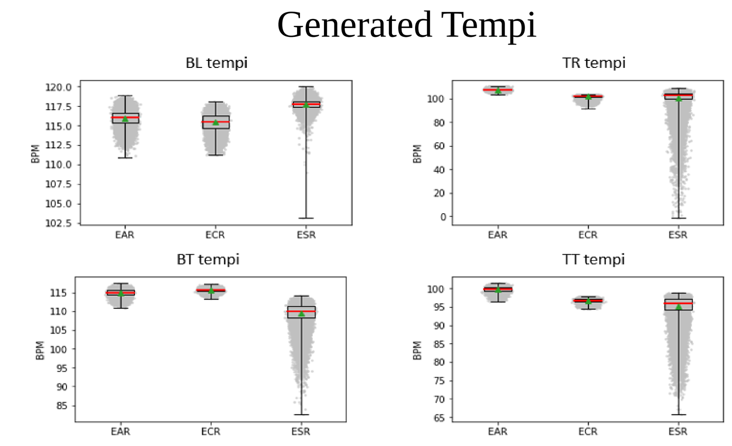
Experimental Evaluation

- **Experimental Conditions**
  - **Dataset (HED)**
    - Symbolic lead sheet data + emotional information dataset (4000 tracks)
      - expanding the HTPD3 with mood tags from last.fm and allmusic.com
  - **Comparisons**
    - Different melody context encoders
      - BLSTM without transfer learning or with transfer learning
      - Transformer encoder without transfer learning or with transfer learning
    - Groundtruth labelled by humans
  - **Procedures**
    - Participants listened to all comparisons generated from the 15 emotion presets and responded to the following:
      - Musical coherence of melody and chords
      - How exciting (arousal) do you perceive the music to be?
      - How negative or positive (valence) do you perceived the music to be?
  - **Participants:** 20 Japanese (Women: 7, Men: 7, Average age 30.15)
- **Results**

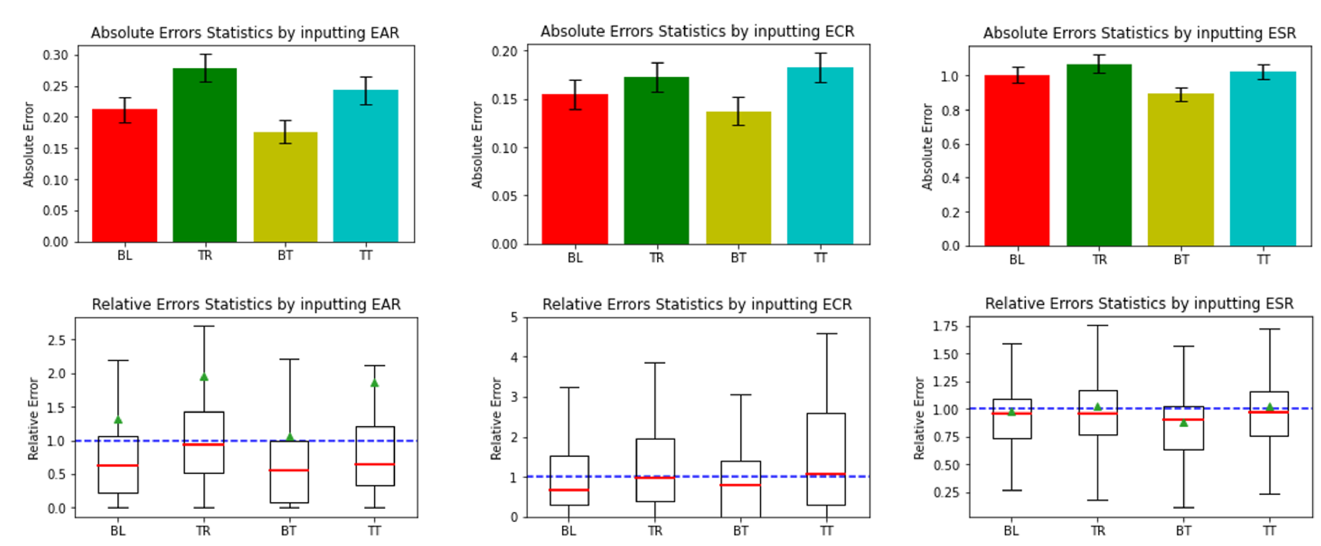
Musical Coherence



Generated Tempi



Emotion Errors



- **Absolute emotion error**
  - EAR: Euclidean distance between input and evaluation emotions
  - ESR: Negative log likelihood calculated from the GMM at the point of the evaluated emotion
  - ECR: Euclidean shortest distance between the quadrant and the evaluated emotion
- **Relative emotion error**

$$\frac{\text{Absolute emotion error of AI−generated arrangements}}{\text{Absolute emotion error between evaluated groundtruth emotion and network input emotion}}$$

Discussion

- The highest perceived musical coherence + The lowest absolute and relative emotion errors ➔ **BLSTM with transfer learning**
- Less than approximately 2.5 seconds per 16 bars for proposed models (**faster than previous study**: 50 seconds per 16 bars for the method proposed by Makris et al.)
- The generated tempo was most varied when **ESR** was used as input
- Possibility of overfitting or insufficient training data