ENHANCING MUSIC DISCOVERY AND ANNOTATION WORKFLOWS WITH

AUTOMATED CHORUS DETECTION

Data Science Intensive Capstone Project

Author: Dennis Dang

Mentor: Andrea Constantinof PhD

CHALLENGES OF MUSIC DISCOVERY

Music

Spotify®

YouTube Music

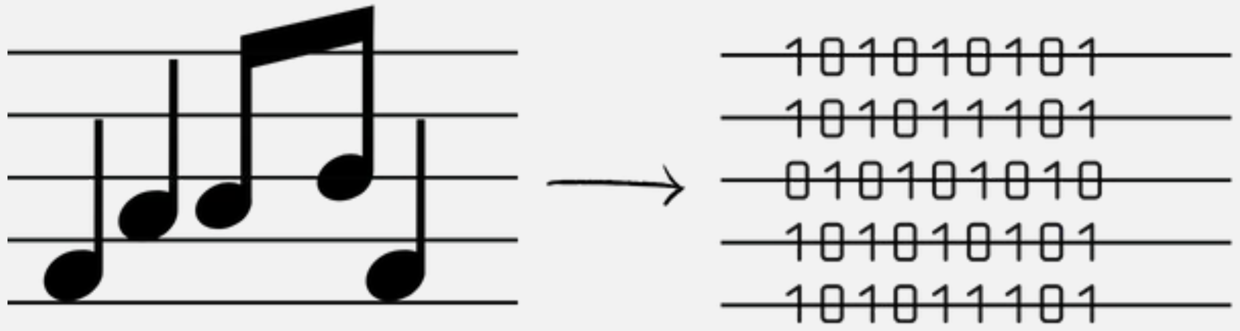
100M+ SONGS

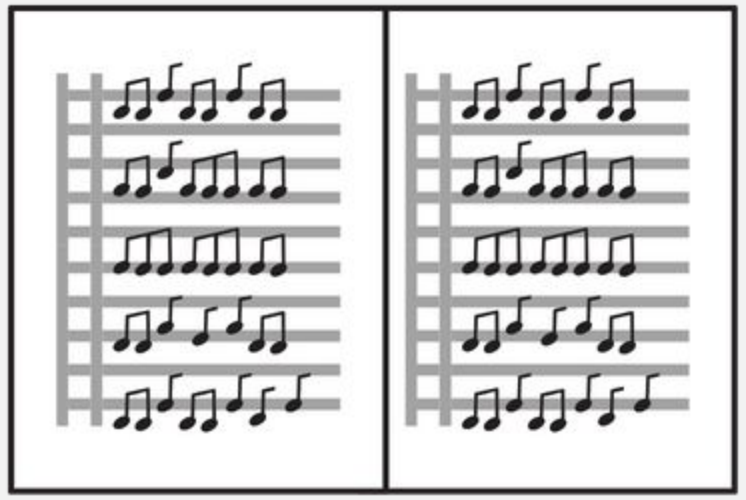
- User data-driven recommendation systems are biased towards popular artists
- Manual music annotation is costly
- Solution: Automated Music Annotation

AUTOMATED MUSIC ANNOTATION

- **Design intelligent systems to...**

Encode music into data → **Learn relevant patterns** → **Make predictions**





STAKEHOLDERS

MUSIC STREAMING PLATFORMS

Music

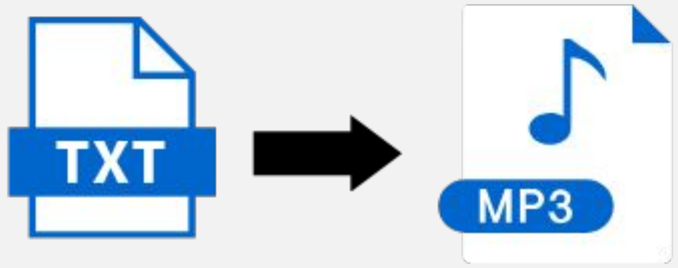
Spotify

YouTube Music

TIDAL

SOUNDCLOUD

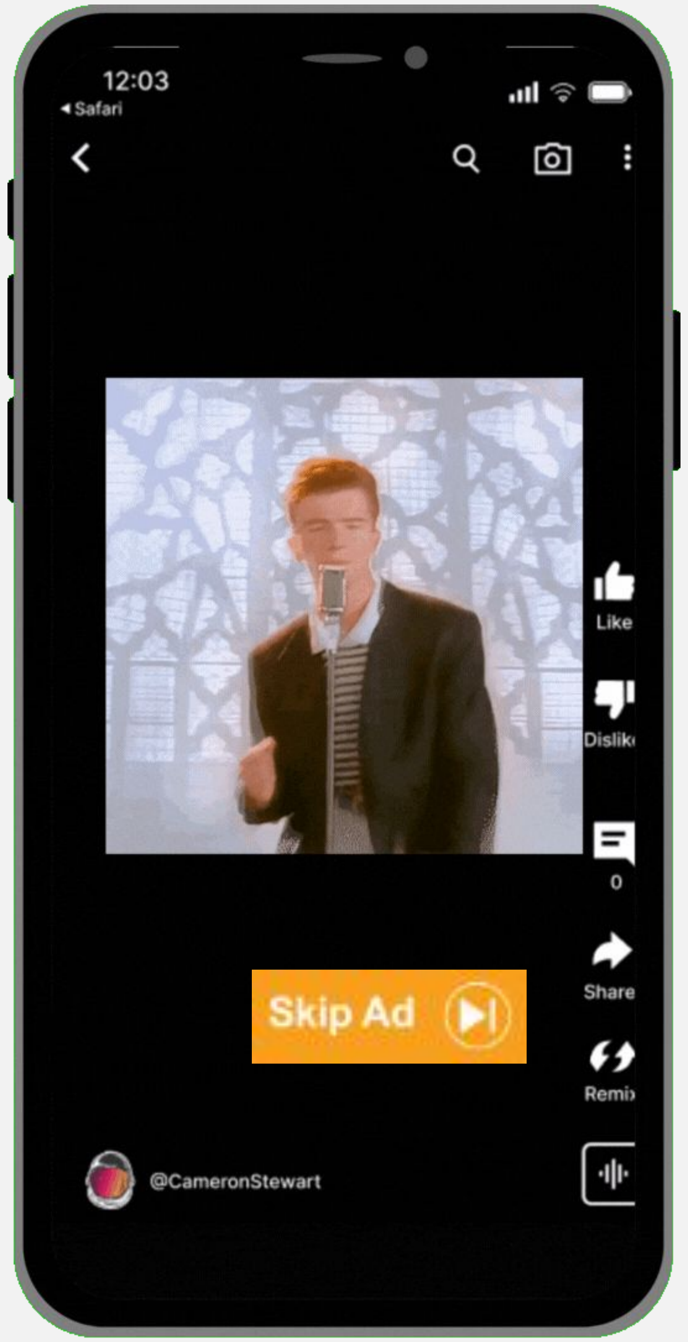
MUSIC SOFTWARE



MEDIA & MARKETING

AUTOMATED CHORUS DETECTION

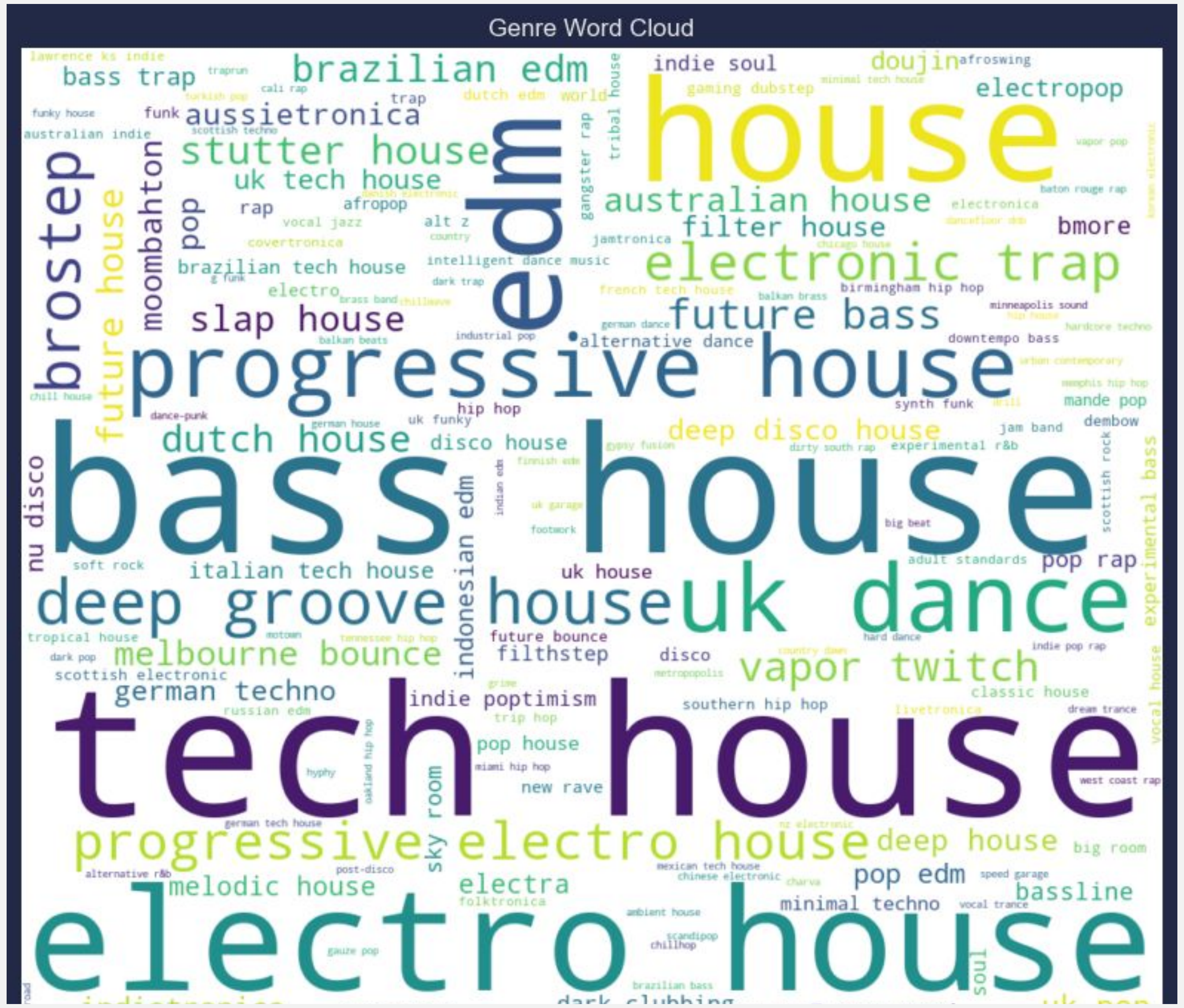
- Automatically identify the most memorable and representative parts of a song
- Automatically generate engaging content for artists on streaming platforms
- Automatically generate labeled segments (for DJing or Music Research)

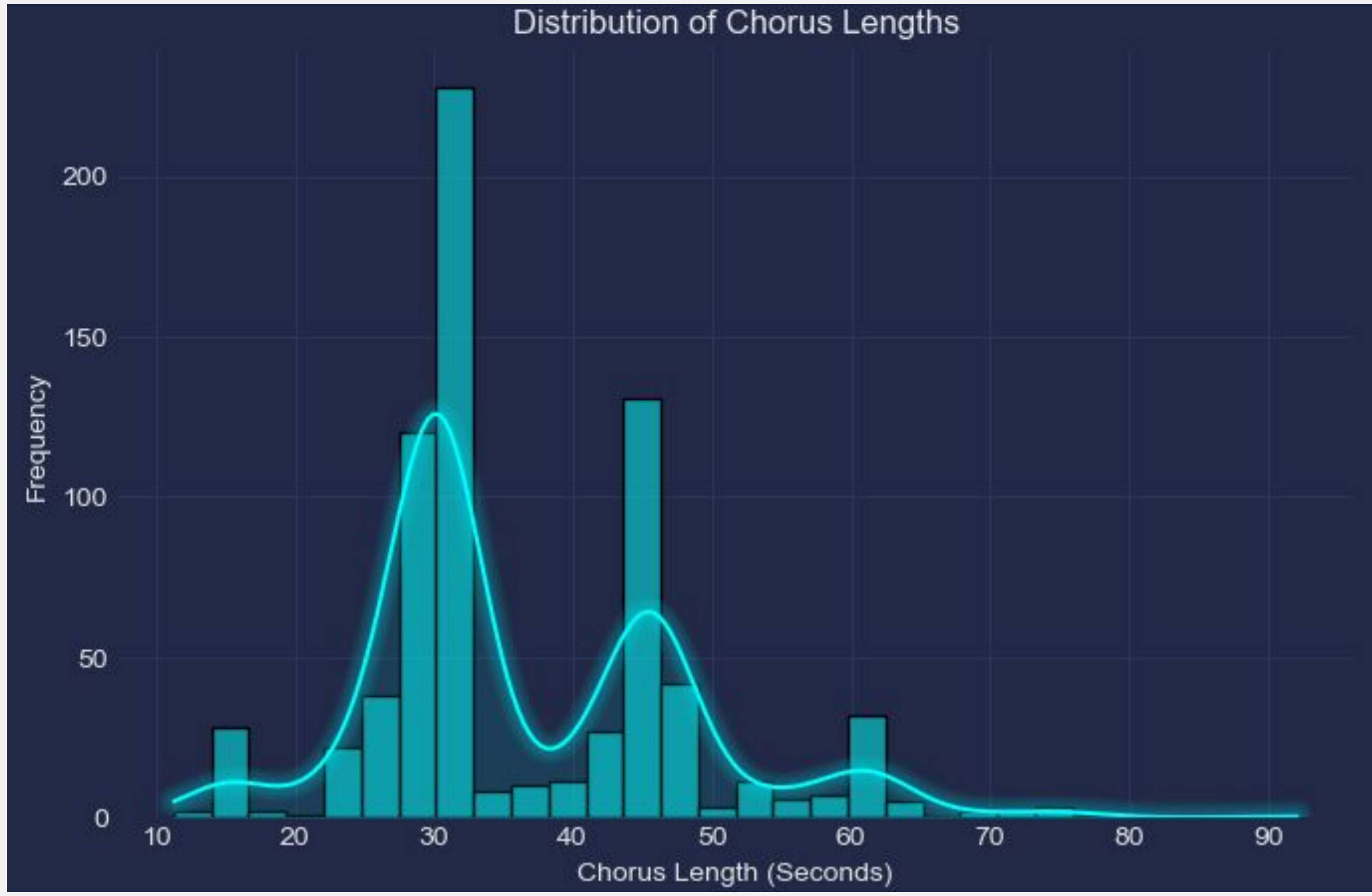


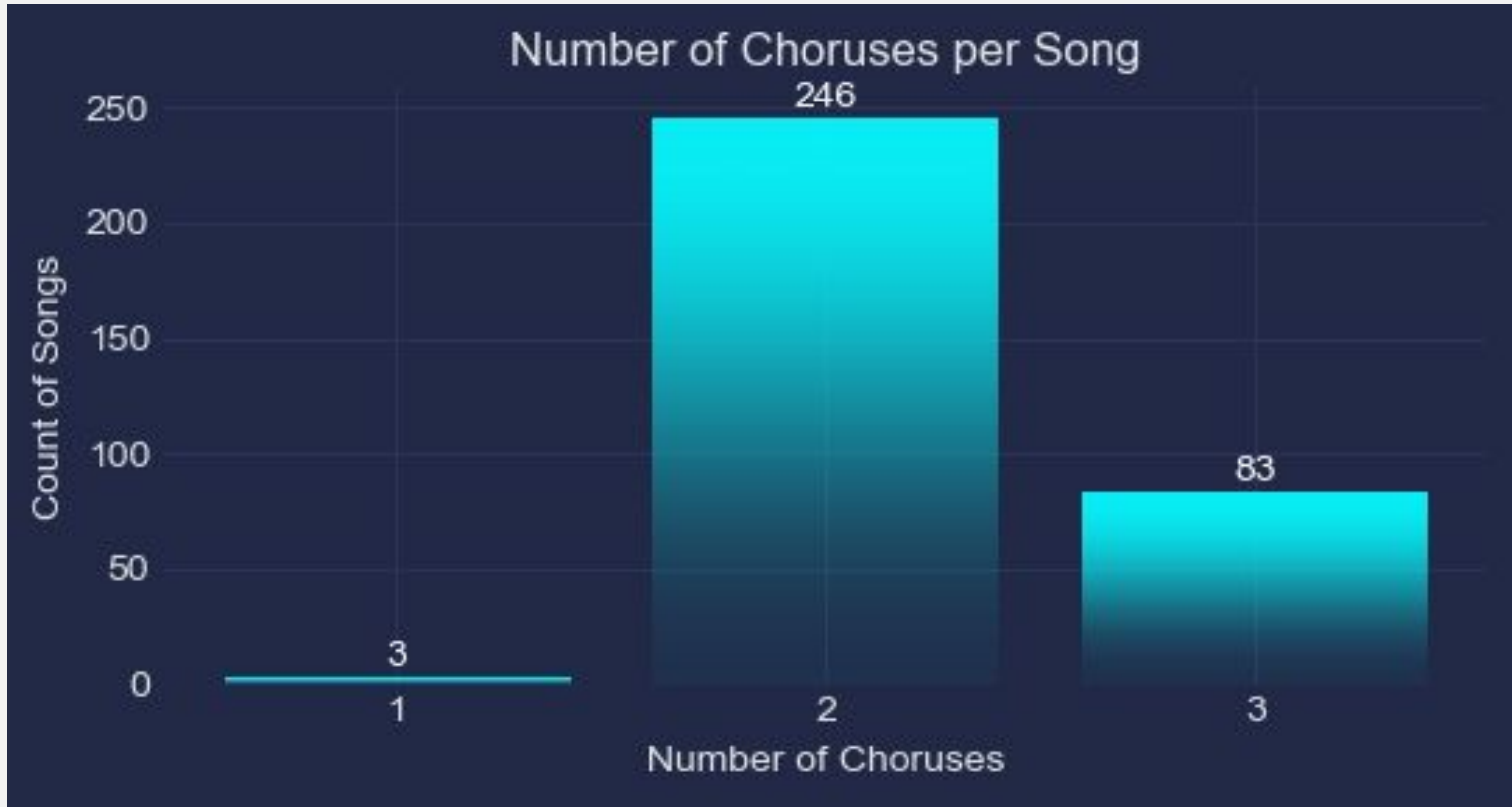
DATA COLLECTION

- **What is a chorus?**
  - A distinctive section of a song that showcases the primary or recurring theme, and significantly contributes to the song's identity and emotional impact

- Established criteria for labeling choruses (refer to [Annotation Guide])
- 332 songs from mostly electronic music genres



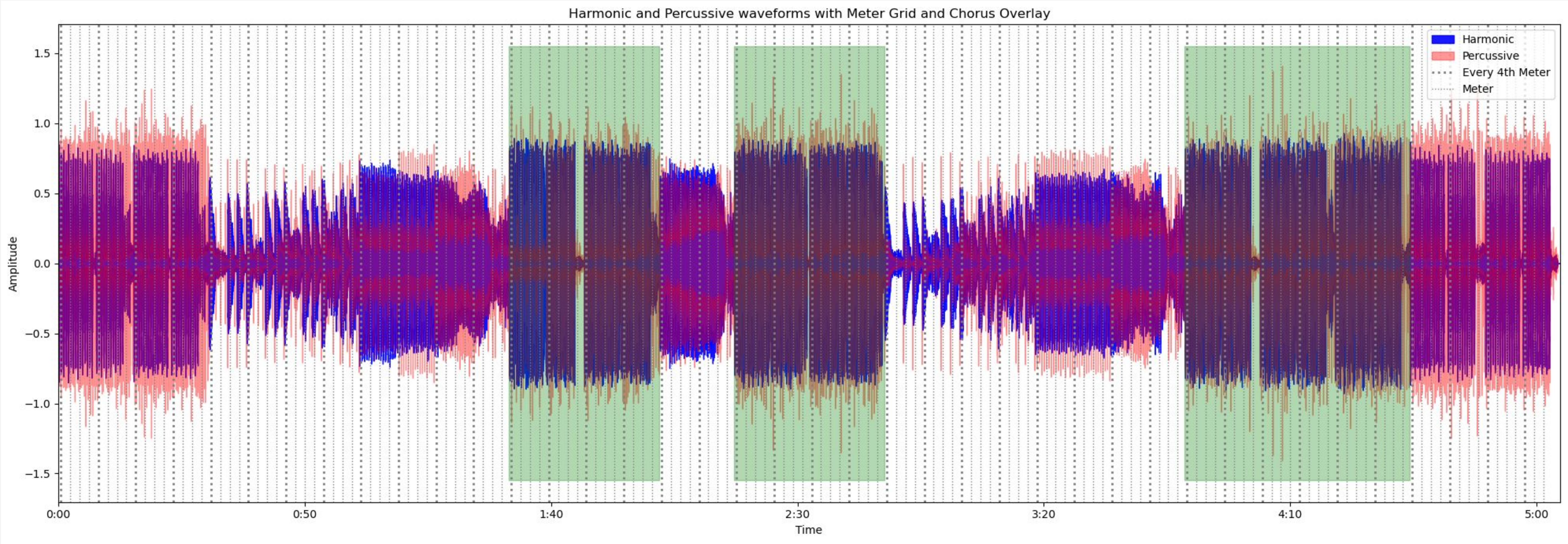




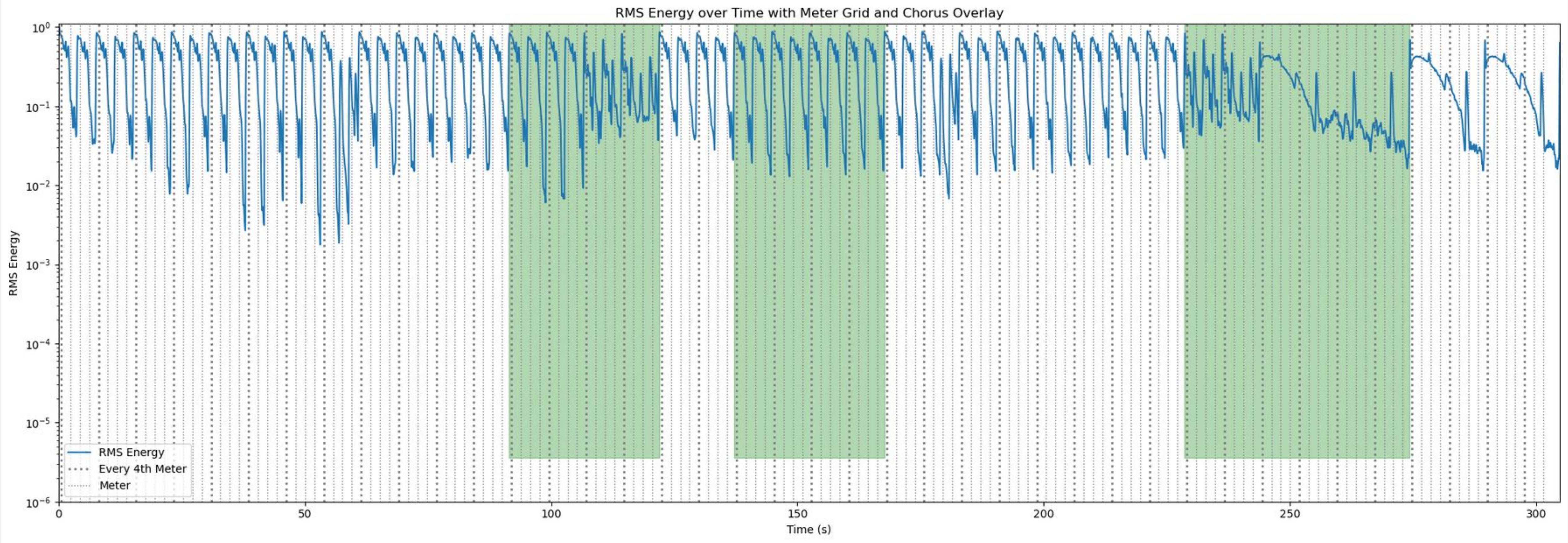
EXPLORATORY DATA ANALYSIS

- **Using visualizations to identify choruses**
  - Chorus labels overlaid in green
  - Musical meters overlaid as dotted lines. Every 4th meter emphasized

- *Which features can I visually identify the chorus with?*
- *Which features behave differently during the chorus vs. non-chorus?*
- *Does this feature align with the meter structure of the song, particularly in chorus sections?*

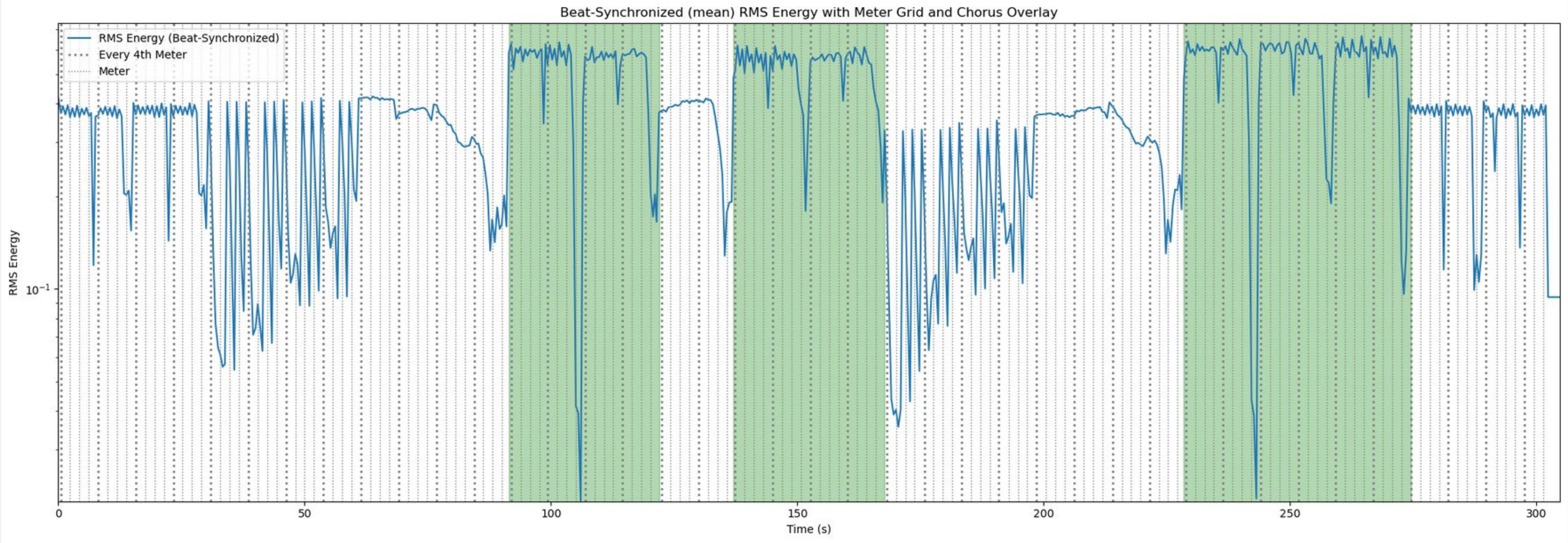


ROOT MEAN SQUARE ENERGY:



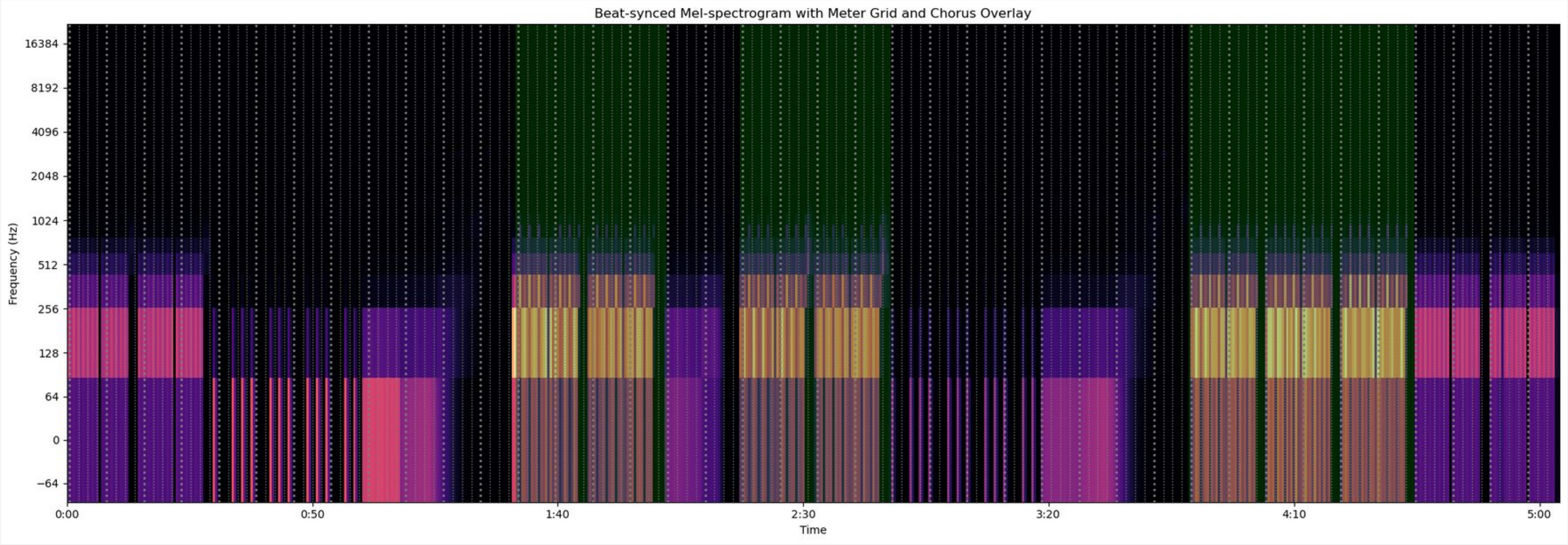
- Average magnitude (loudness) of an audio signal over a specific time window
- Low dimensional feature
  (1 x n_timesteps)
  (1 x 40000 audio frames)

BEAT-SYNCED RMS ENERGY:



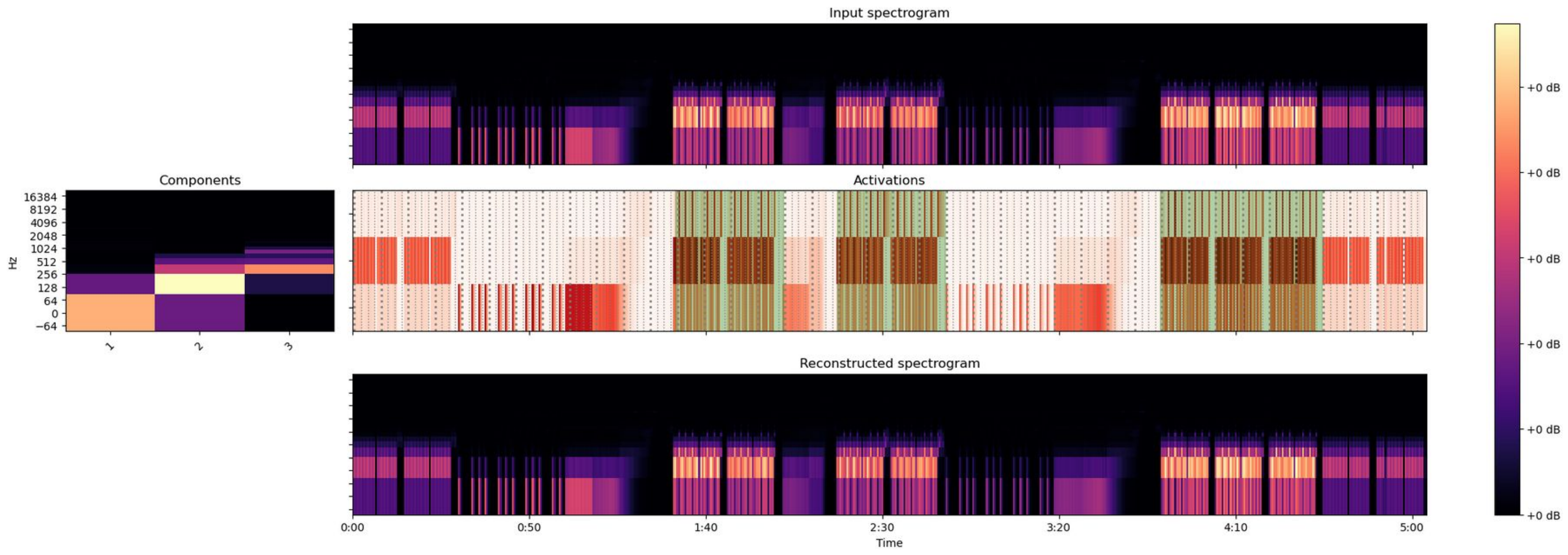
- Mean interpolation of RMS Energy between each beat
- Chorus sections become distinguishable
- Highlights the importance of rhythmic structure when analyzing audio features

MEL SPECTROGRAM:



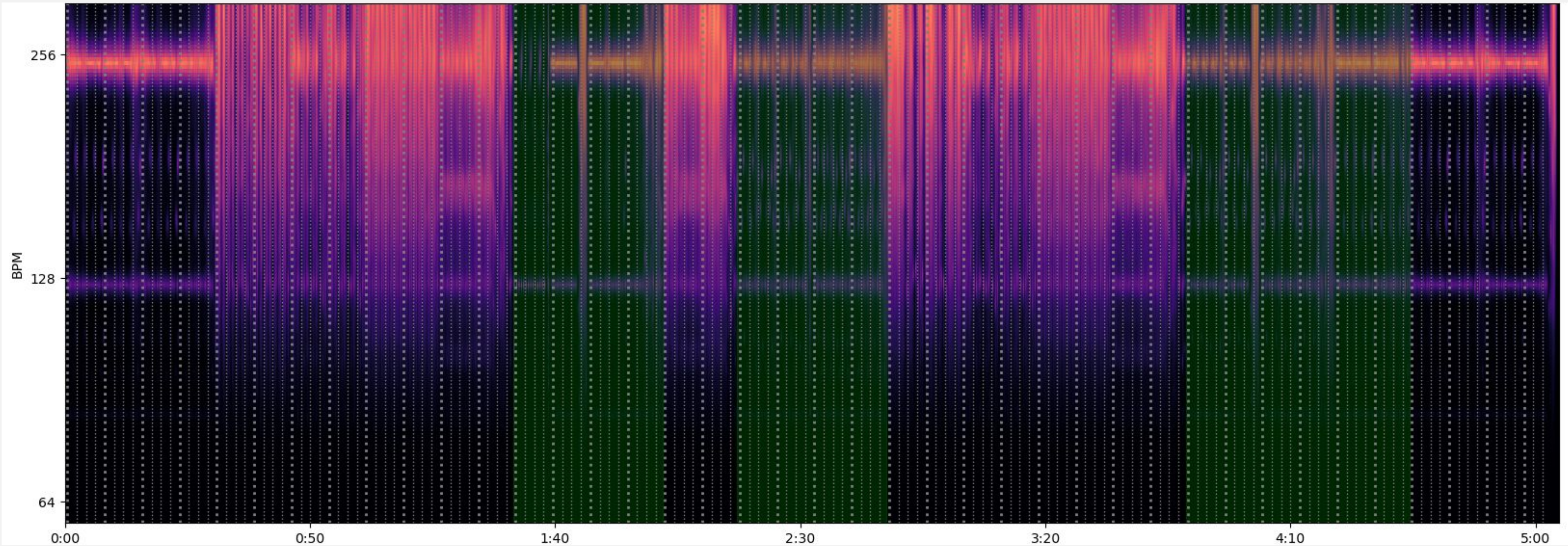
- Frequency (Hz) of an audio signal over time
- High-dimensional feature *(n_mel_bins x n_timesteps)* *(128 x 40000)*

DECOMPOSED MEL SPECTROGRAM:



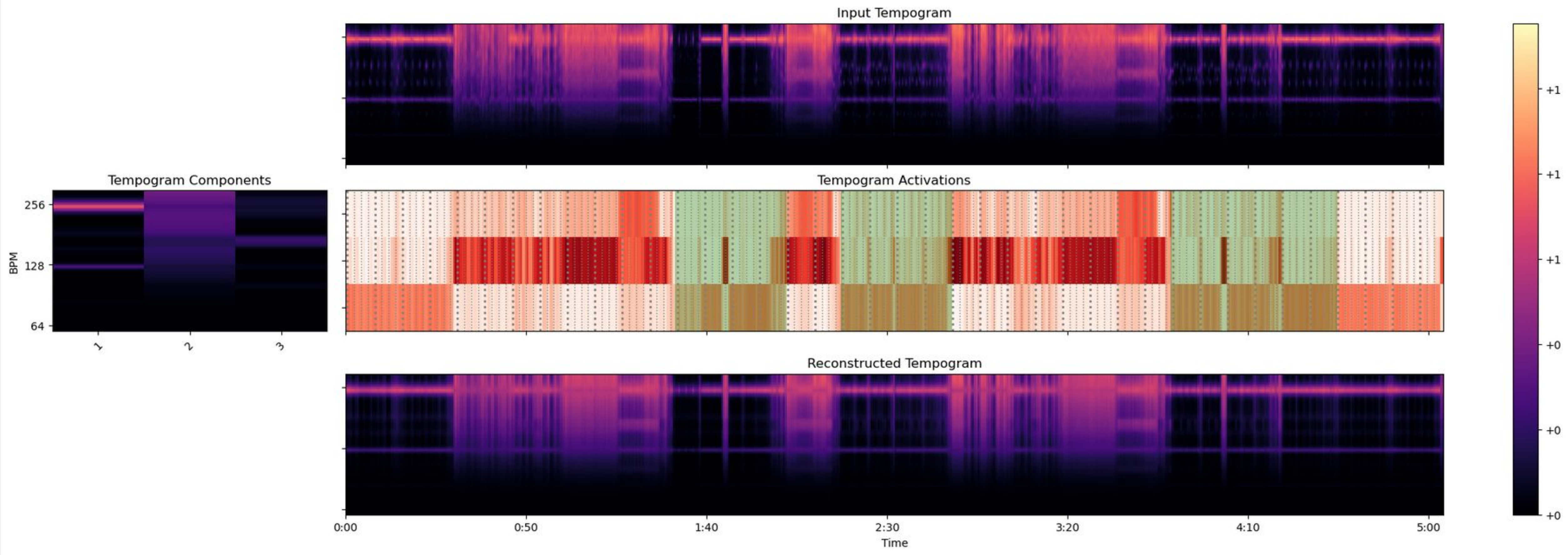
- Decompose using Non-negative Matrix Factorization (NMF) into 3 components and time-varying activations
- Reduces dimensionality, memory/compute requirements *(n_components, n_timesteps)* *(3, 40000)*

TEMPOGRAM:



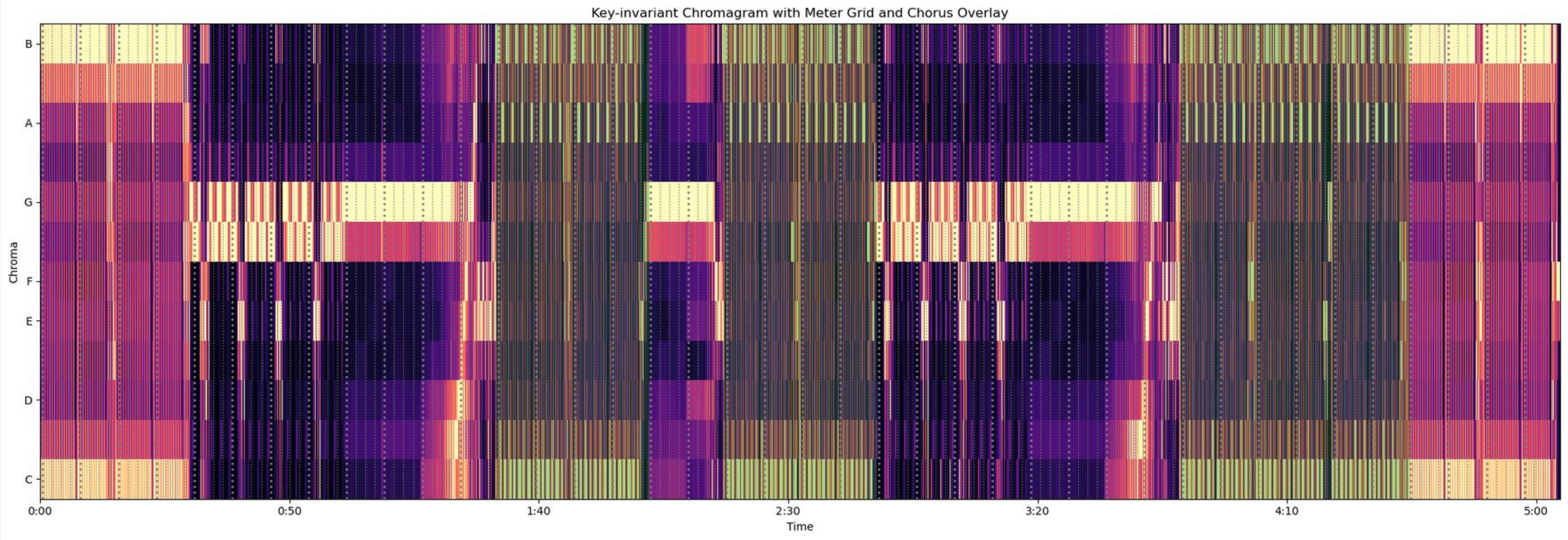
- Rhythmic/tempo stability over time
- Illuminates “conflict resolution” song structure
- High-dimensional feature *(autocorr_window_size X n_timesteps)* *(384 x 40000)*

DECOMPOSED TEMPOGRAM:



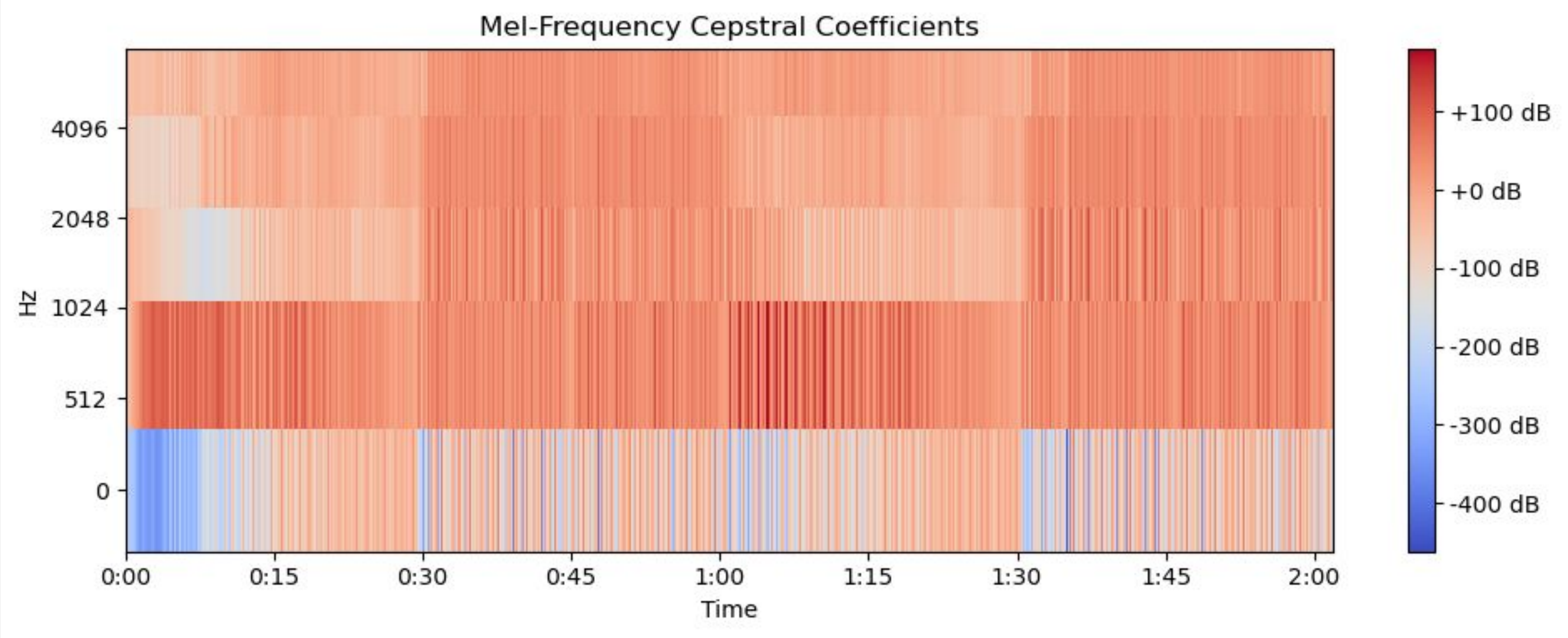
- Decomposed using NMF into 3 components and activations

CHROMAGRAM:



- Captures tonal content represented as energy across the 12 pitch classes (C, C#, D, D#, E, F, F#, G, G#, A, A#, B)

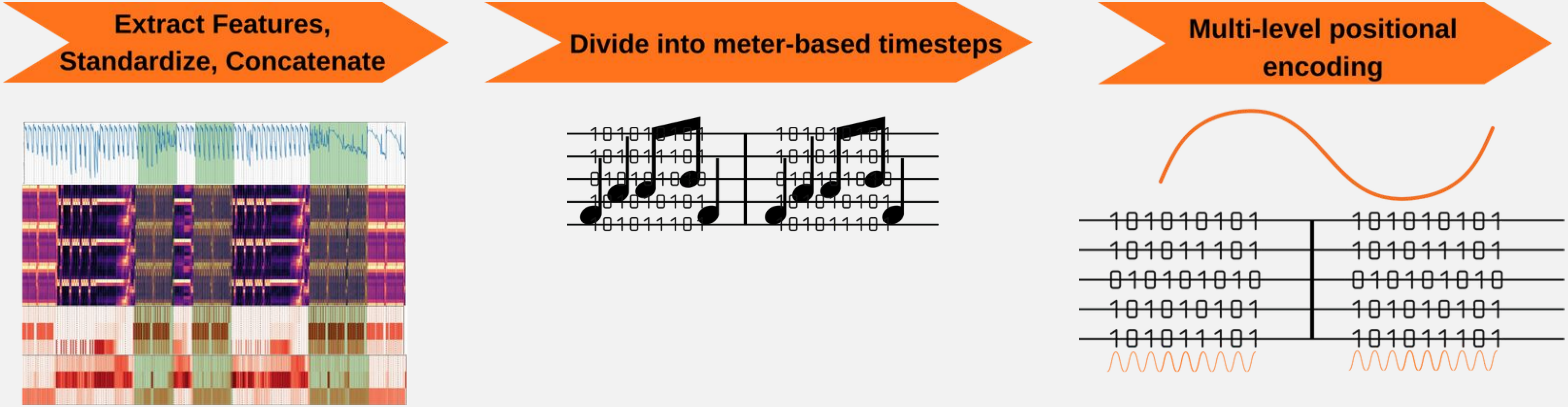
MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)



- Captures timbral characteristics of the audio signal

MODEL PREPROCESSING

- **Convolutional Recurrent Neural Network (CRNN)**
- **Features extracted:** RMS energy, chromagram activations, tempogram activations, MFCC activations, mel spectrogram activations (15 dimensions)
- **Meter-based timesteps:** Input shape = (201, 300, 15) *(meters, frames, features)*
- **Positionally encode each meter AND each frame in a meter**



MODEL ARCHITECTURE

- **CRNN Input Shape**
  - Each song contains 201 meters
  - Each meter contains 300 audio frames
  - Each audio frame consists of 15 features
  - Input shape: (201, 300, 15)

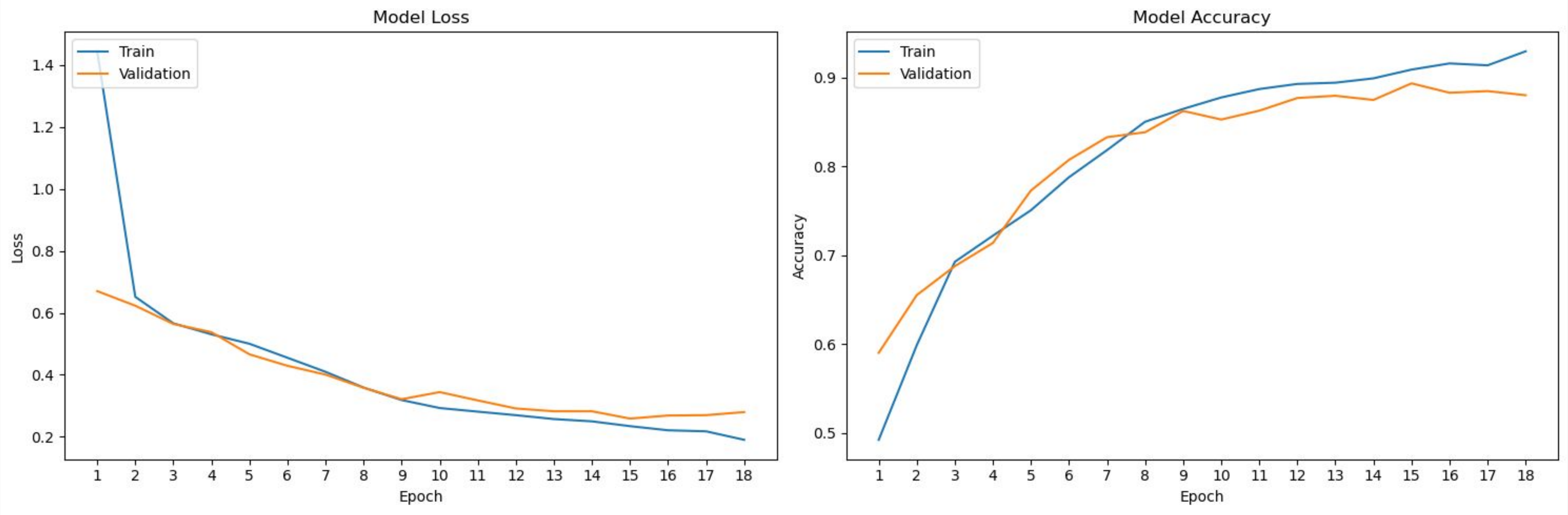
- **Three Convolutional layers to extract frame-level features from each meter**
  - Input shape: (300, 15) *(i.e. a meter)*
  - 1D convolutional layer -> ReLU activation function -> 1D max-pooling layer **(x3)**

- **One Recurrent layer for temporal summarization of the features extracted by CNN**
  - Bidirectional LSTM processes the input sequence in both forward and backward directions
  - Output shape is (201, 512) *(512 LSTM units)*

- **Dense layer applies sigmoid activation function to each time step (meter)**
  - Output shape is (201, 1) representing the probability of a chorus being present at each meter in a song

MODEL TRAINING

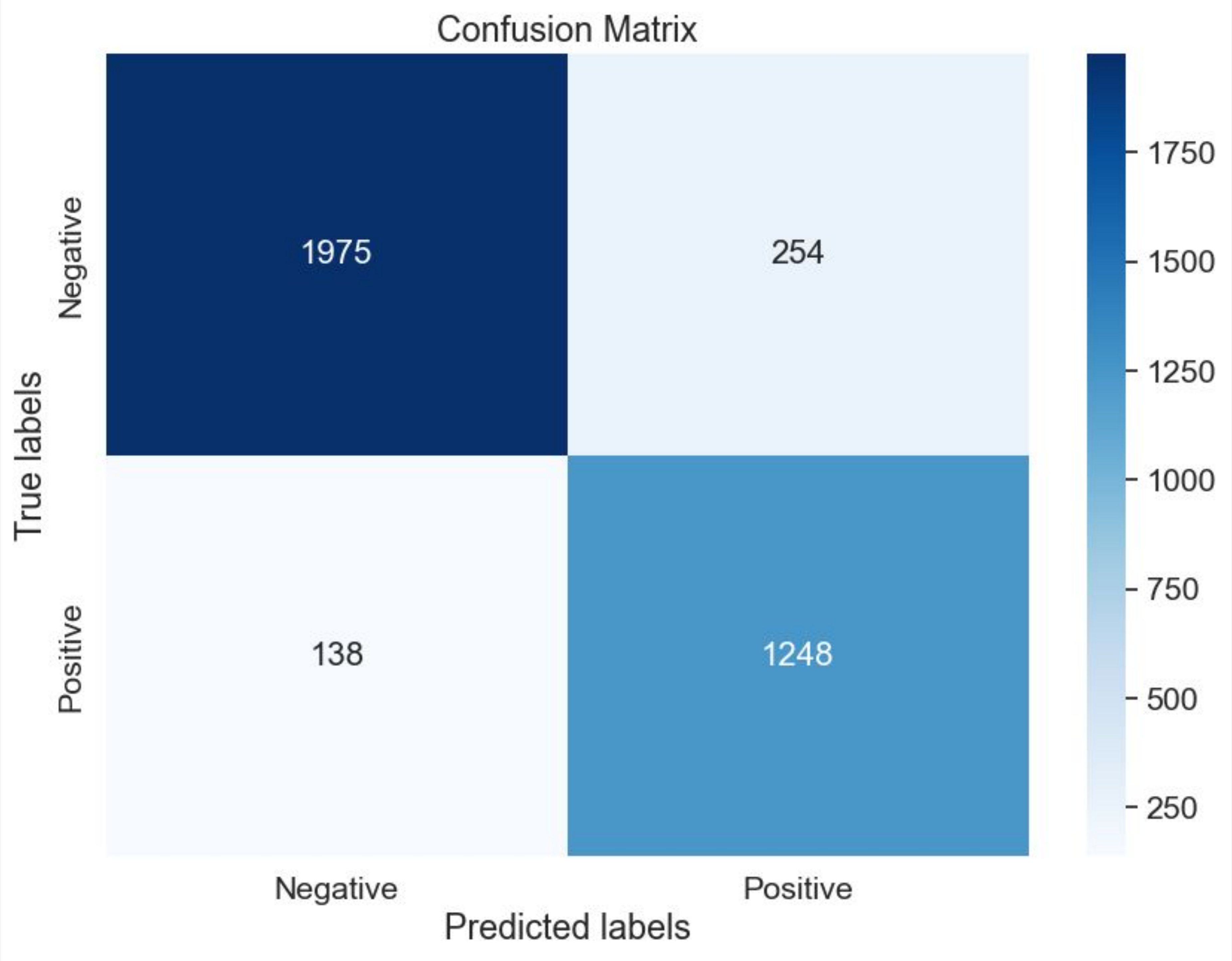
- **Trained over 50 epochs using the training and validation datasets**
- **Callbacks:**
  - **ModelCheckpoint:** Save the best model based on minimizing validation loss (binary cross entropy)
  - **EarlyStopping:** Stop training if validation loss doesn't improve for 3 epochs
  - **ReduceLROnPlateau:** Reduce learning rate if validation loss plateaus



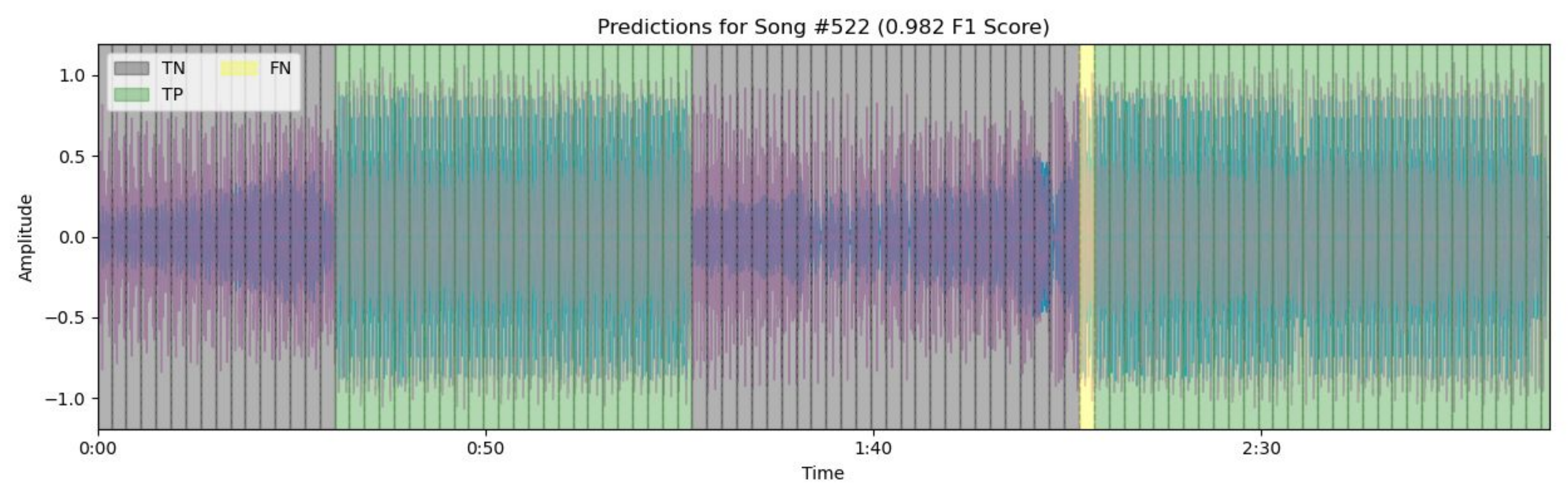
MODEL EVALUATION

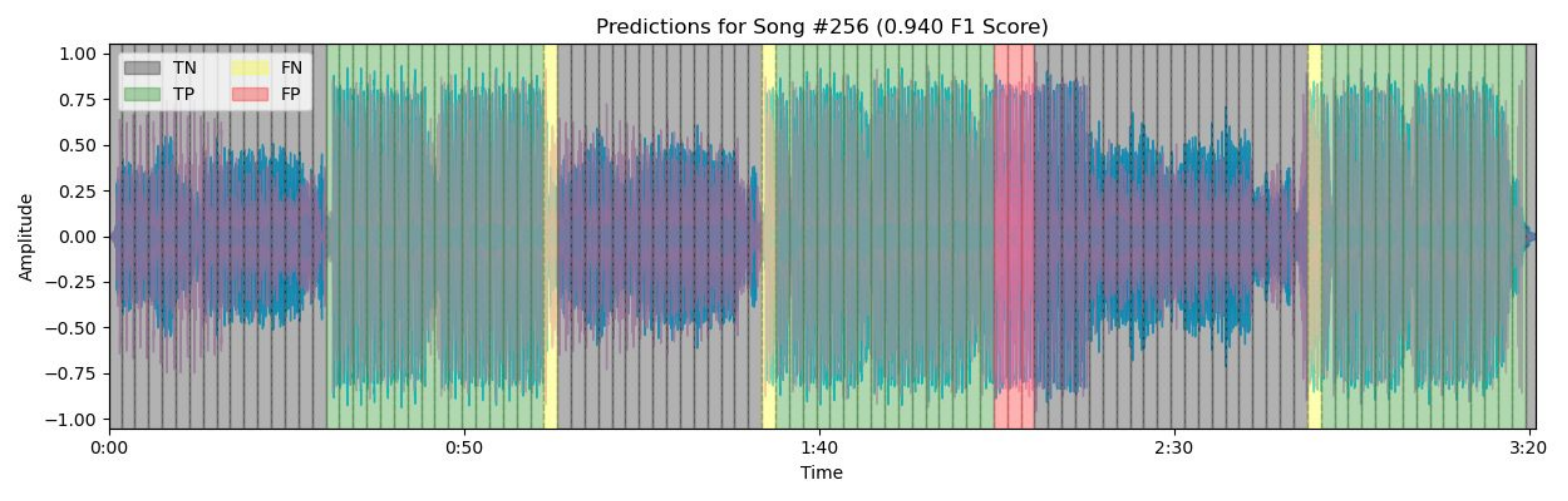
- **Evaluated model on unseen test set of 50 songs**

- **Results:**
  - **Accuracy:** 0.891
  - **F1 Score:** 0.864
    - **Precision:** 0.831 (17% false positive rate)
    - **Recall:** 0.900 (10% false negative rate)

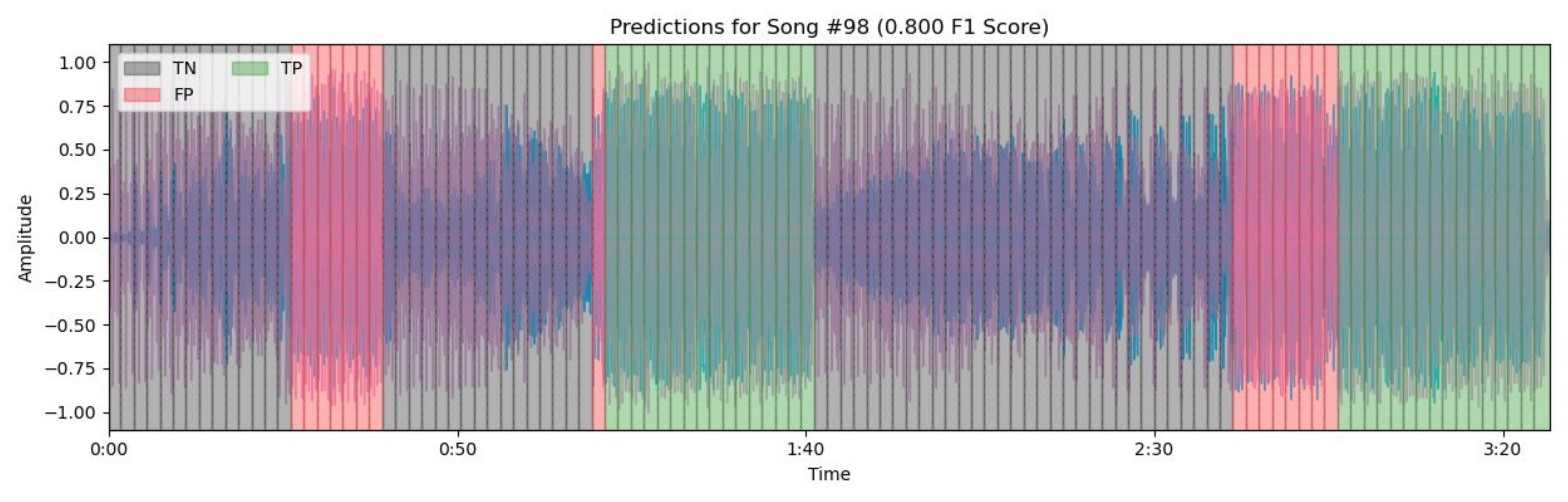


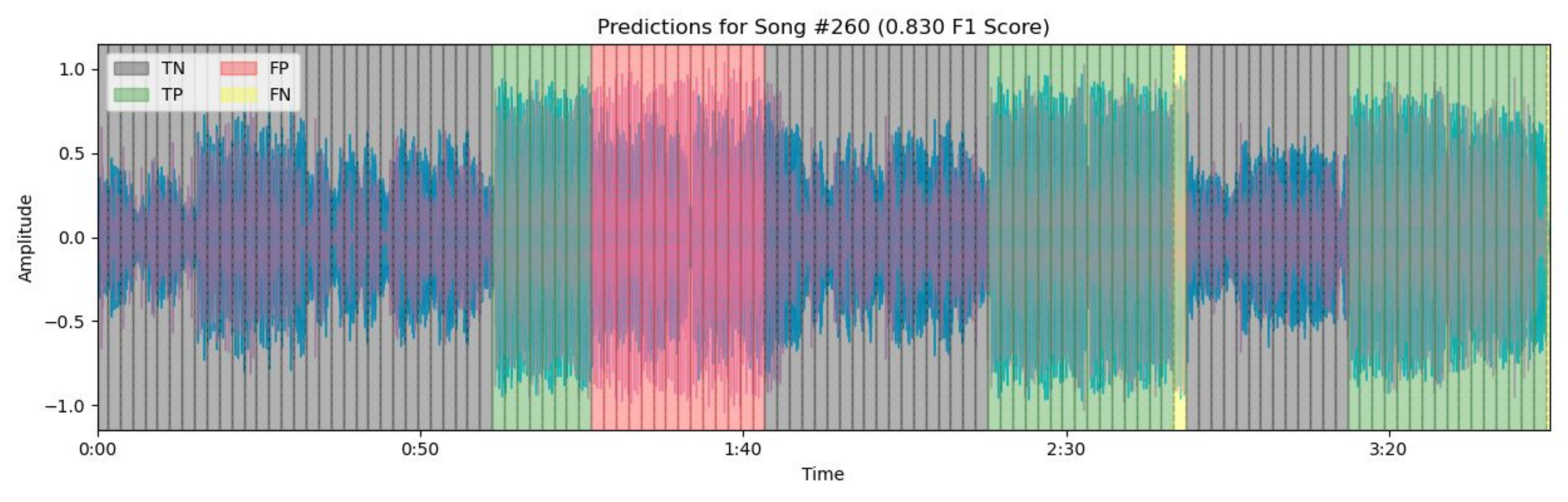
TEST PREDICTIONS



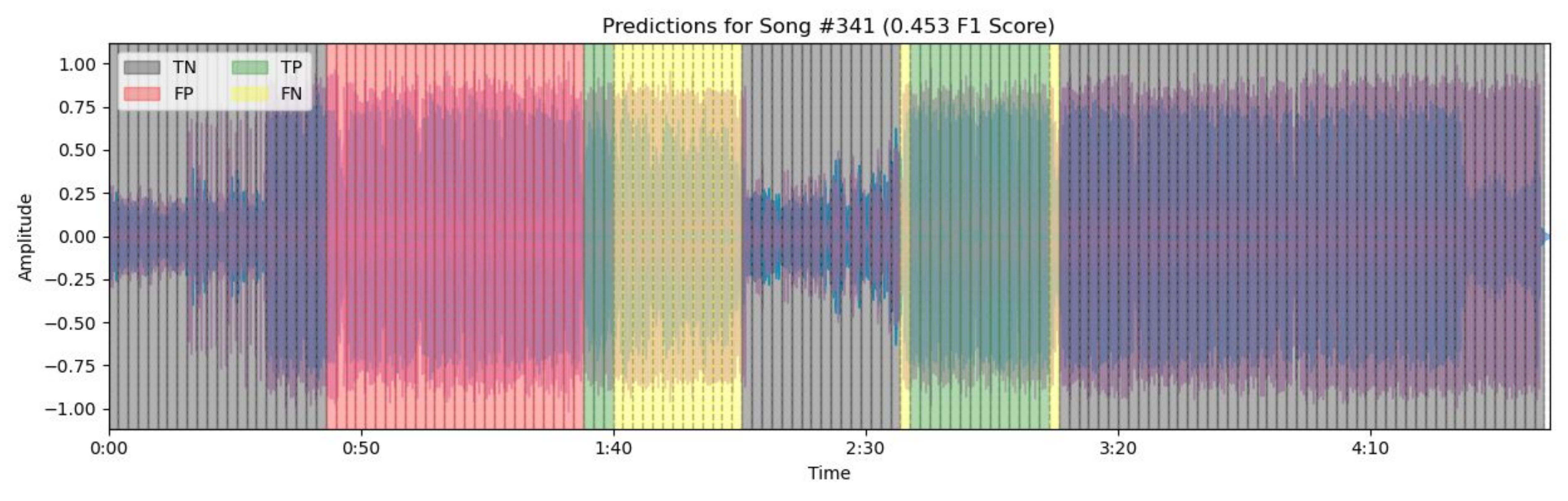


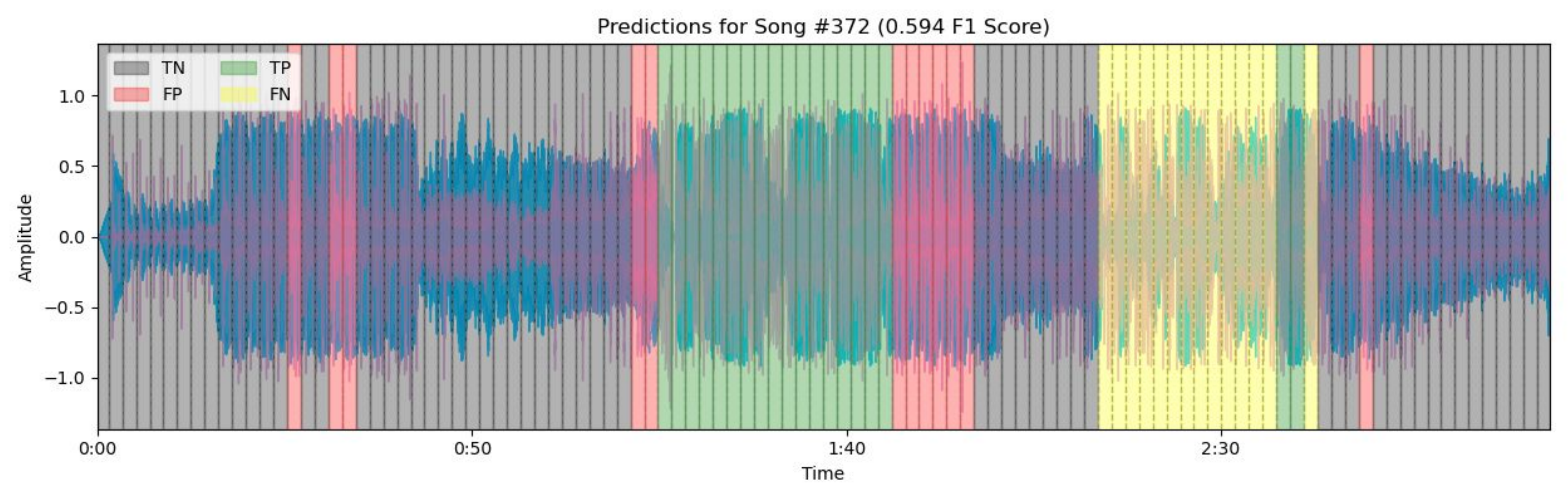
TEST PREDICTIONS





TEST PREDICTIONS





FUTURE CONSIDERATIONS

- **Prediction post-processing**
- **Experiment on wider variety of genres**
- **Experiment using multiple time-resolutions (e.g. frames, beats, meters)**
- **Tuning feature weights and hyperparameters (e.g. filter size, batch size, LSTM units)**