Collaboration of Two AI Musicians: Challenges and Possibilities

Kit Armstrong 2024-03-28

Environment

- Fixed music scores
- Real-time interaction
- Live performance
- Acoustic communication

Objectives

- Each AI can function alone
- Together, two AIs “interpret” their respective parts in sync

Objectives

- Each AI can function alone
  - can play its part (i.e. score → audio)
  - can at the same time collaborate with a human

- Together, two AIs “interpret” their respective parts in sync

Objectives

- Each AI can function alone

  - can play its part (i.e. score → audio)

  - can at the same time collaborate with a human

- Together, two AIs “interpret” their respective parts in sync

  - each AI does not know the partner's identity

  - the overall effect should be human-like

Relevant capabilities

- Playing
- Listening

Relevant capabilities

- Playing

  - turn a score into audio in real-time,
    with flexibility in timing and dynamics
    (relatively easy for MIDI piano,
    more difficult for voice)

- Listening

Relevant capabilities

- Playing

  - turn a score into audio in real-time,
    with flexibility in timing and dynamics
    (relatively easy for MIDI piano,
    more difficult for voice)

- Listening

  - identifying input

  - score following (“alignment”)

Relevant capabilities - Playing

- For MIDI piano, we have a system that:

  - writes a score as an array of dimension 2N x 6

    each note consists of 1 note-on, 1 note-off event

    event format:
    [index, type, note, score position, time, velocity]

  - can adjust (time, velocity) of each note
    at any time, following “musical intent”

  - outputs the array in real-time
    as MIDI performance

Relevant capabilities - Listening

- Identifying input

- Score alignment

Relevant capabilities - Listening

- Identifying input

  - we want essentially an Audio to MIDI module, aka “transcription”

- Score alignment

  - from the MIDI data, we can fill in the score array

    our new “search-by-beat” protocol, used for automating data processing (for SMC submission) works reasonably well

Relevant capabilities - Listening

- Identifying input
  - we want essentially an Audio to MIDI module, aka “transcription”

Relevant capabilities - Listening

- Identifying input

  Pitch detection

Relevant capabilities - Listening

- Identifying input

  Pitch detection

  - Problems:

    noise

    echo

    single pitch

    human intonation inaccuracy

Two simultaneous parts present in audio
Pitch detection algorithm "confused"

Peaks at 200Hz, 300Hz
→ Pitch detection returns 100Hz?

MIDI generated thus will not be a good representation of intent

Relevant capabilities - Listening

- Can we circumvent this problem?
- How do human musicians do it?

Relevant capabilities - Listening

- Can we circumvent this problem?
- How do human musicians do it?

- One piece of information we have not used: the score
- I think humans do not fully transcribe; rather,

  they start from the score,
  identify salient events on both sides (score, audio),
  confirm a match upon “reasonably certain” detection

Relevant capabilities - Listening

Idea for “query-based” listening:

no separation of tasks into detection, alignment

- At every time t, calculate initial hypothesis for $P(t)$, “score position corresponding to the current time”, by extrapolating assuming constant tempo
- Salient note-on events in the vicinity of $P(t)$ are checked against the audio (“salient” = “unique”)
- Find probability that the audio contains the event
- If threshold not met, initial hypothesis holds; If threshold met, $P(t)$ is set to the score position of the matched event

Relevant capabilities - Listening

Issues

- “Probability that the audio contains the event”: absolute or time-based?
- Acoustic effects: it will always get slower

Relevant capabilities - Listening

Issues

- "Probability that the audio contains the event": absolute or time-based?
- Acoustic effects: it will always get slower

  - we do not know the attack curve

  - so when a match with score event E occurs, we can only surmise that $P(t) \geq p(E)$

  - thus $P(t)$ should be adjusted only positively

  - stability problem? "Mutual acceleration"?

Relevant capabilities - Listening

- Of course, we can do even better
- More available information: the listener's own performance
- If we can “subtract” the expected acoustic effect from the audio, we could isolate the partner's performance

Putting it together

- Two machines, each following this listening protocol
- Each perceives a series of identified couplets (time, score position)
- By linear interpolation & Kuramoto model, it generates a fitted accompaniment