AN LSTM-BASED DYNAMIC CHORD PROGRESSION GENERATION SYSTEM FOR INTERACTIVE MUSIC PERFORMANCE

Christos Garoufis, Athanasia Zlatintsi, and Petros Maragos

School of ECE, National Technical University of Athens, Greece
Robot Perception and Interaction Unit, Athena Research Center, Greece
cgaroufis@mail.ntua.gr; [nzlat, maragos]@cs.ntua.gr

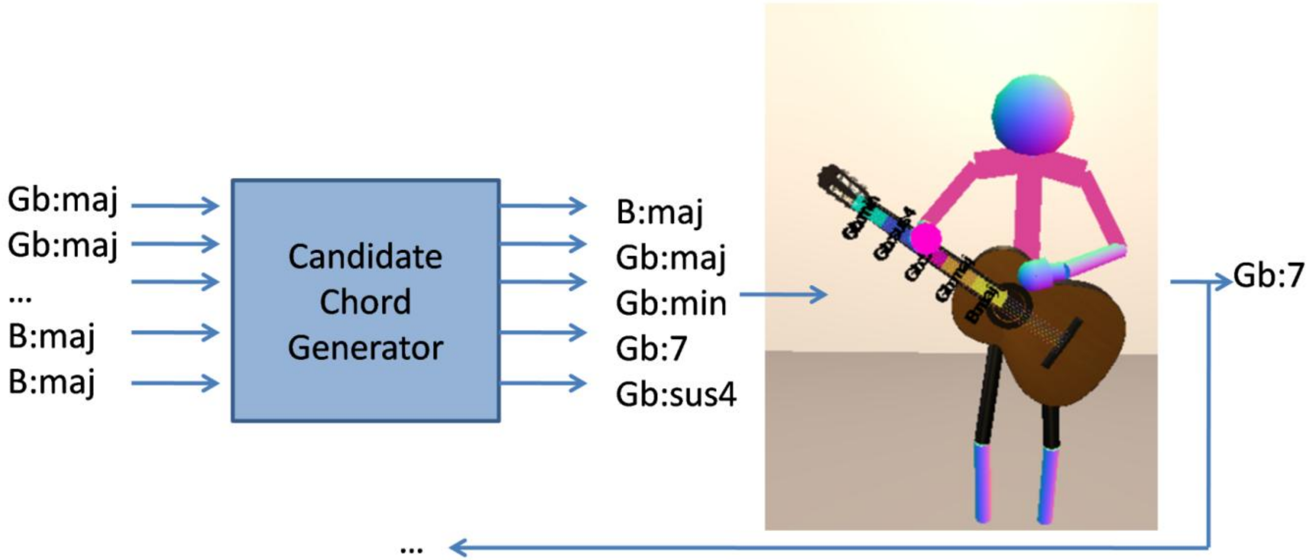
ICASSP2020 Barcelona

Overview

- Main Idea – Introduction
- System Architecture
- Methodology
- Experimental Setup
- Results & Discussion
- Conclusions & Future Work

Main Idea

- An intersection between automatic chord progression generation and interactive music performance.
- **Generative**, because…
  Candidate chords are generated automatically.
- **Interactive**, because…
  A human performer is involved.

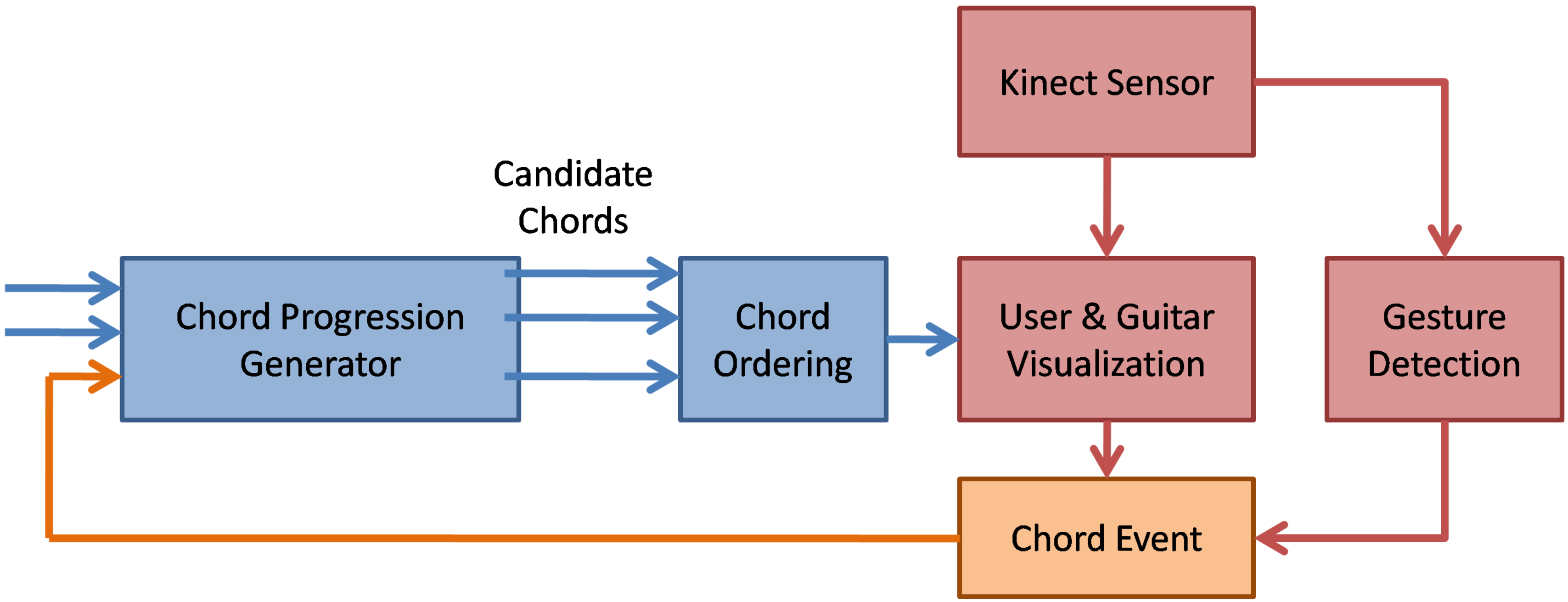


Automatic Music Generation

- **Generative music system**: A system that **algorithmically** composes music, based on some rules.
- **State of the art**: Neural network architectures that can capture **long-range temporal dependencies**, such as RNNs [1], or attention-based networks [2].
- **Interactive generative systems**: An external user can **modify** some of the music **parameters** [3].

[1] N. Boulanger-Lewandowski et al, “Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription”, ICML 2012
[2] Elliot Waite, “Generating long-term structure in songs and stories,” in Magenta Blog, 2016.
[3] C. Donahue, I. Simon, and S. Dieleman, “Piano Genie,” IUI 2019.

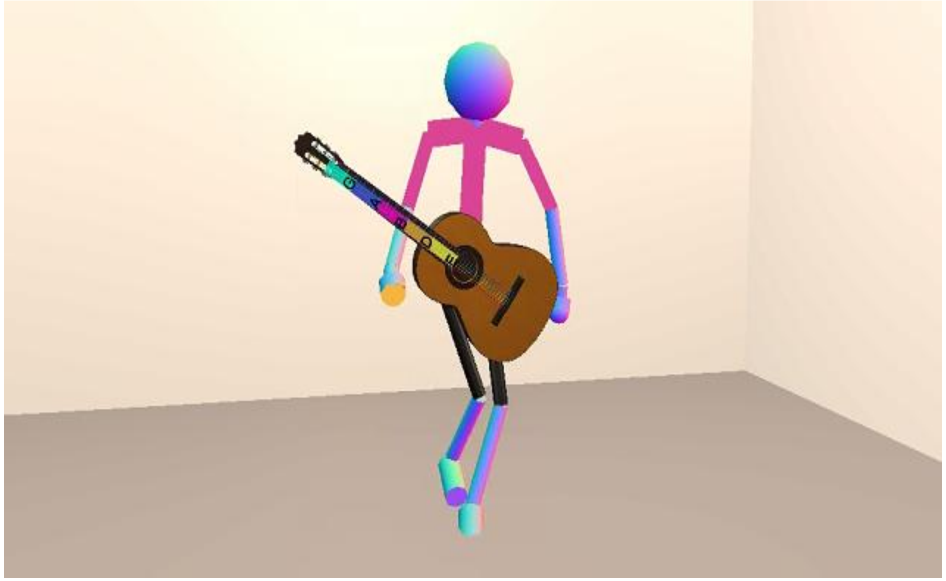
Overall system architecture



Methodology: Interaction & Visualization

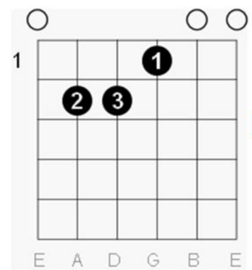
- Our interface deploys a **Kinect** sensor.
- During performance, a **skeletonized avatar** appears in the computer screen, along with a **virtual instrument**.

- **Dominant hand:** Performs **plucking gestures** to play guitar chords.
- **Subdominant hand**: Defines the played chord via its **placement** in the virtual **fretboard**.



Methodology: Data Representation and Problem Formulation

- **Data representation:** **Pianorolls**, at the time resolution the song beat dictates.



...	E	F	F#	G	G#	A	A#	B	C	C#	D	D#	E	...
...	1	0	0	0	0	0	0	1	0	0	0	0	1	...

- **Problem Formulation**: From a *NxT* sequential array of chords, **predict** the *Nx1* pianoroll that corresponds to the following chord.

- **Loss function**: **MSE** between the true and predicted pianorolls (**regression** problem formulation).

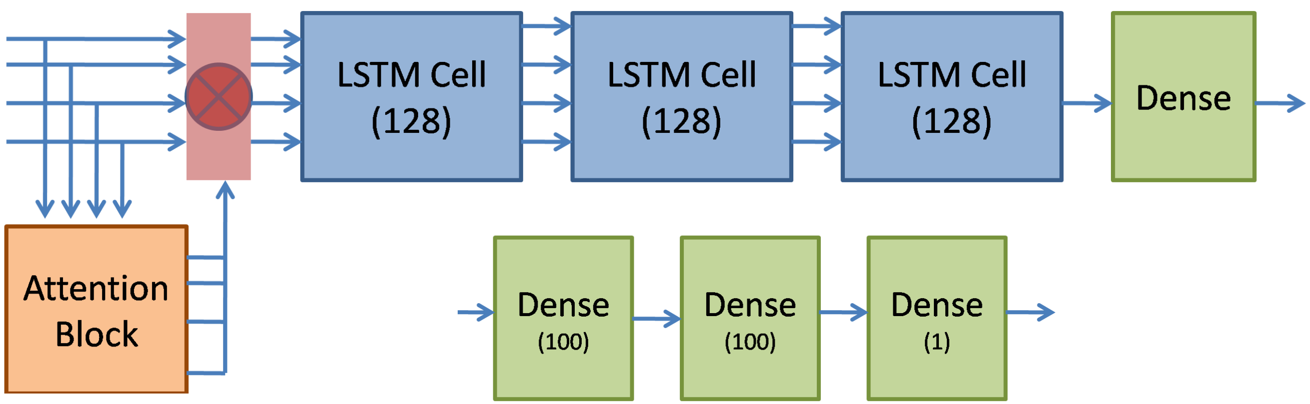
Methodology: Chord Progression Generation

- **Base architecture**: 3 **LSTM cell layers** & a fully connected output layer.

- **Proposed modifications**:

  a) **a switch detection** mechanism, predicting whether the played chord changes.

  b) a **temporal attention** layer, applied directly to the network input.



Methodology: Chord Ordering

- The network outputs a **single** pianoroll per timestep.
- Selection of a number of candidate chords (#5) based on their **Euclidean distance** to the predicted pianoroll.

- **Challenge**: How should we **position** them in the virtual fretboard?
- **Proposed solution**: training of a **genetic algorithm**, to provide a suitable chord ordering.

Experimental Setup: Data Preprocessing

- **Initial Dataset**: McGill Billboard Dataset [4]

  Data preprocessing:
- Reduction of the dataset, keeping only songs where the guitar is included in the **dominant** instruments.
- Simplification of the **chord vocabulary** (10 chord types per root chroma – total of **121** chords).
- Transformation of the chord annotations into **pianoroll** format.
- **Final Dataset Statistics**: **442** songs, **192869** chords.

[4] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis.," ISMIR 2011.

Experimental Setup: Evaluation

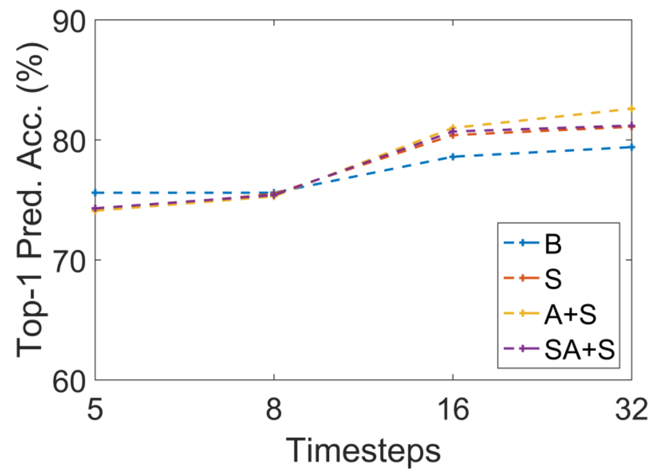
- **Objective**:
  Can we correctly **predict** the next chord in a chord sequence?
- **Training Protocol**: 20 epochs, 5-fold cross-validation.
- **Metrics**: Top-1 and top-5 prediction **accuracy** (%)

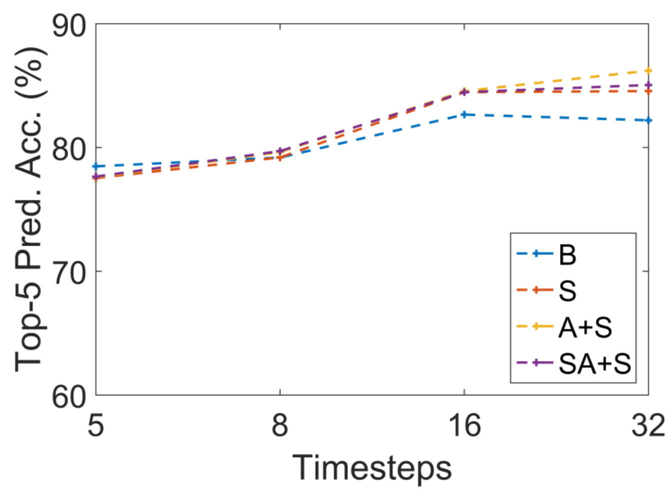
- **Subjective**:
  Are the generated chord progressions **valid** from a **musical** point of view?
- **Testing Protocol**: User evaluation tests.
- **Metrics: Coherence** and **variety** of proposed chords (5-point Likert scales)

Results & Discussion: Objective Evaluation

- For **small** sequence lengths, all architectures perform generally **equally**.

- As the input sequence length gradually **increases**, we observe an improvement due to both the **switch detection** (S) mechanism and, for **even larger** sequences, the **temporal attention (A+S).**

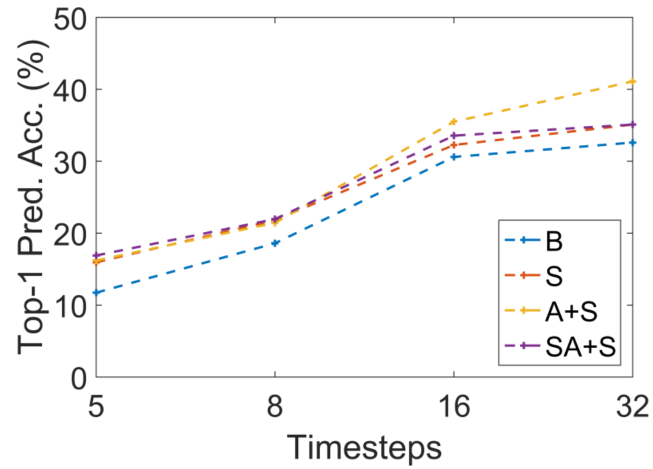


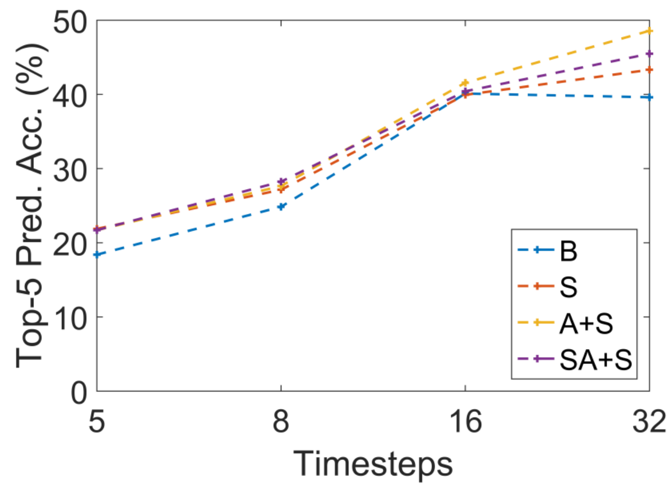


Results & Discussion: Objective Evaluation

- This improvement is more clearly **evident** considering only the cases where a **chord switch** occurs.

- Connecting the attention module to the **latent space** before the last LSTM layer (SA+S) does not perform equally well to applying directly to the **input** (A+S).





Results & Discussion: Objective Evaluation

- Inferring the **modality** (major, minor, augmented…) of a predicted chord is **easier** to inferring its **chroma**.

- Using the **attention mechanism** improves the **chroma** prediction accuracy, in contrast to the chord **modality** prediction accuracy.

Setup Used	Top-1 %			Top-5%		
	Acc.	C.Acc.	T.Acc	Acc.	C.Acc	T.Acc
B	79.40	81.26	85.00	82.20	86.28	93.81
S	81.05	82.74	85.96	84.56	**91.44**	**97.54**
A+S	**82.60**	**84.34**	**86.60**	**86.21**	91.17	97.24

Top-1 and top-5 chord prediction accuracies, regarding the chord, (Acc.) the chord chroma (C.Acc.) and the chord type (T.Acc.) for the baseline (B), switch (S) and attention+switch (A+S) architectures.

Setup Used	Top-1 %			Top-5%		
	Acc.	C.Acc.	T.Acc	Acc.	C.Acc	T.Acc
B	32.60	38.66	56.03	39.62	52.21	**80.21**
S	35.12	40.66	48.06	43.32	52.22	66.10
A+S	**41.06**	**48.67**	**51.01**	**48.58**	**58.32**	66.81

Top-1 and top-5 chord prediction accuracies, regarding the chord, (Acc.) the chord chroma (C.Acc.) and the chord type (T.Acc.) for the baseline (B), switch (S) and attention+switch (A+S) architectures, in the instances of chord change.

Results & Discussion: Subjective Evaluation

- The chord progressions generated by the **baseline** architecture were slightly more **coherent** musically than those generated by the more complex architectures.

- The **variety** of the generated chords increased significantly when the switch architecture was used, especially when temporal **attention** was also utilized.

Architecture	Mus. Coherence	Variety
B	**3.58**	1.83
S	3.33	3.08
A+S	3.08	**3.67**

Results of the subjective evaluation of our system with regards to the perceived musical coherence and variety of our system, using a 5-point Likert scale.

Conclusions

- Presentation of an interactive chord progression generation system.
- Positive results regarding the performance of our system in **chord prediction** from a given chord progression.
- Improved prediction accuracy when utilizing the **attention** module, especially in the cases a **chord switch** occurs.
- Room for improvement regarding **long-term** chord progression generation.

Future Work

- Perceptually motivated **distance metrics** for selecting candidate chords from pianorolls.
- Unification of pianoroll prediction, chord selection and chord ordering in an **end-to-end** architecture.
- Experimentation with recent breakthroughs in **natural language processing**.
- Usage of conditioning learning to condition the generated chords on a musical parameter, such as **genre**.

Thank you for your attention!
We wish everyone courage and health during the COVID19 pandemic.

For more information, demos, and current results: http://cvsp.cs.ntua.gr and http://robotics.ntua.gr