Audio: Generation & Extraction

Charu Jaiswal

Music Composition – which approach?

- Feed forward NN can’t store information about past (or keep track of position in song)
- RNN as a single step predictor struggle with composition, too
    - Vanishing gradients means error flow vanishes or grows exponentially
    - Network can’t deal with long-term dependencies

- But music is all about long-term dependencies!

Music

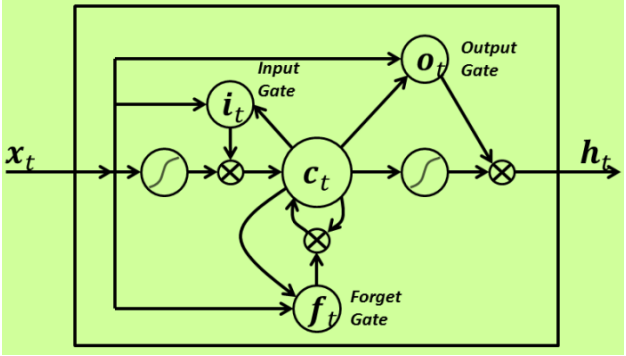
- Long-term dependencies define style:
    - Spanning bars and notes contribute to metrical and phrasal structure

- How do we introduce structure at multiple levels?
    - Eck and Schmidhuber →LSTM

Why LSTM ?

- Designed to obtain constant error flow through time
- Protect error from perturbations
    - Uses linear units to overcome decay problems with RNN

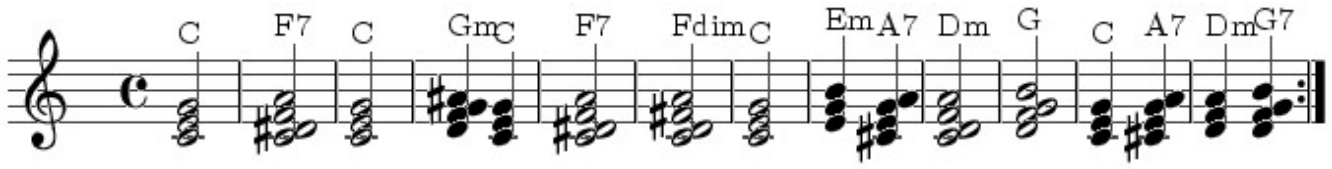
    - Input gate: protects flow from perturbation by irrelevant inputs
    - Output gate: protects other units from perturbation from irrelevant memory
    - Forget gate: reset memory cell when content is obsolete



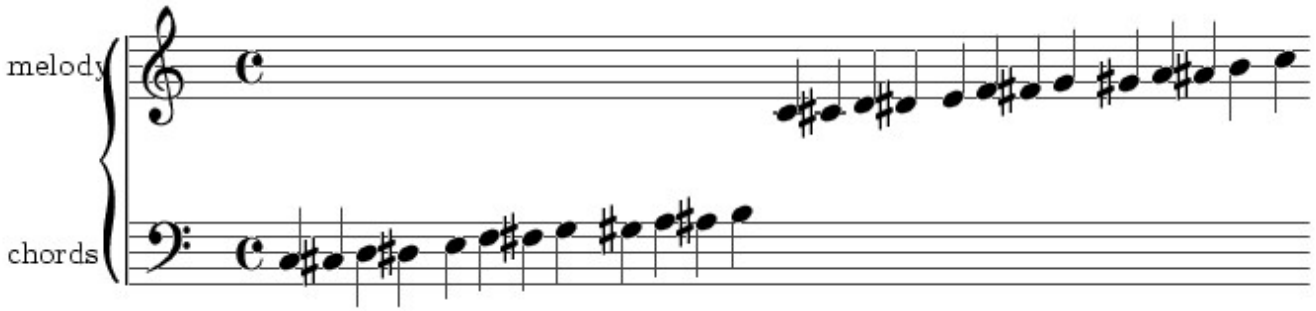
Hochreiter & Schmidhuber, 1997

Data Representation

Chords :



Notes:



Only quarter notes

No rests

Training melodies written by Eck

Dataset of 4096 segments

Experiment 1- Learning Chords

- Objective: show that LSTM can learn/represent chord structure in the absence of melody
- Network:
  - 4 cell blocks w/ 2 cells each are fully connected to each other + input
  - Output layer is fully connected to all cells and to input layer
- Training & testing: predict probability of a note being on or off
  - Use network predictions for ensuing time steps with decision threshold
  - CAVEAT: treat outputs as statistically independent. This is untrue! (Issue #1)
- Result: generated chord sequences

Experiment 2 – Learning Melody and Chords

- Can LSTM learn chord & melody structure, and use these structures for composition?
- Network:
  - Difference for ex1. : chord cell blocks have recurrent connections to themselves + melody; melody cell blocks are only recurrently connected to melody
- Training: predict probability for a note to be on or off

Sample composition

- Training set: http://people.idsia.ch/~juergen/blues/train.32.mp3

- Chord + melody sample: http://people.idsia.ch/~juergen/blues/lstm_0224_1510.32.mp3

Issues

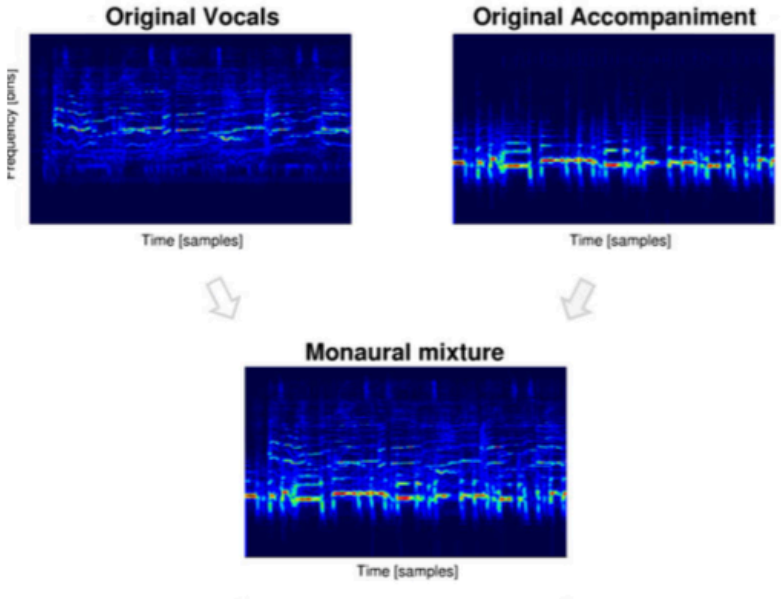
- No objective way to judge quality of compositions
- Repetition and similarity to training set
- Considered notes to be independent
- Limited to quarter notes + no rests
- Uses symbolic representations (modified sheet notation) → how could it handle real—time performance music (MIDI or audio)
  - Would allow interaction (live improvisation)

Audio Extraction (source separation)

- How do we separate sources?
- Engineering approach: decompose mixed audio signal into spectrogram, assign time-frequency element to source
  - Ideal binary mask: each element is attributed to source with largest magnitude in the source spectrogram
  - This is then used to est. reference separation

DNN Approach

- Dataset: 63 pop songs (50 for training)
  - binary mask computed: determined by comparing magnitudes of vocal/non-vocal spectrograms and assigning mask a ‘1’ when vocal had greater mag

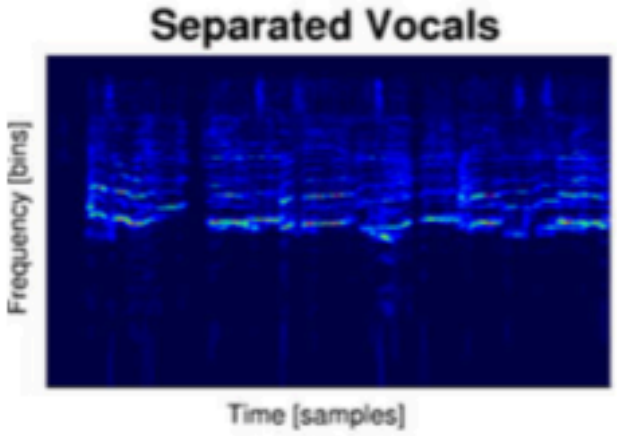


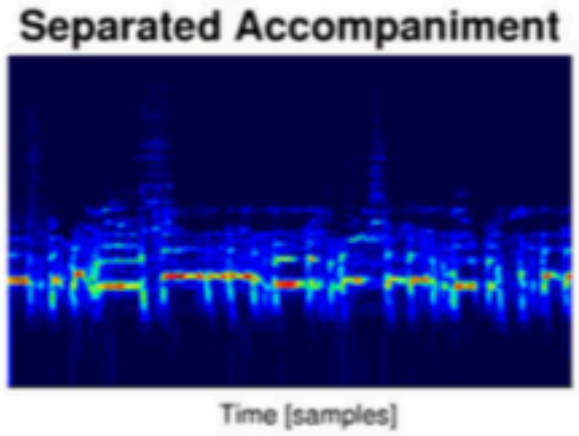
DNN

- Trained a feed-forward DNN to predict binary masks for separating vocal and non-vocal signals for a song
  - Spectrogram window was unpacked into a vector
  - Probabilistic binary mask: testing used sliding window, and output of model described predictions of binary mask in sliding window format
  - Confidence threshhold (alpha): Mv binary mask

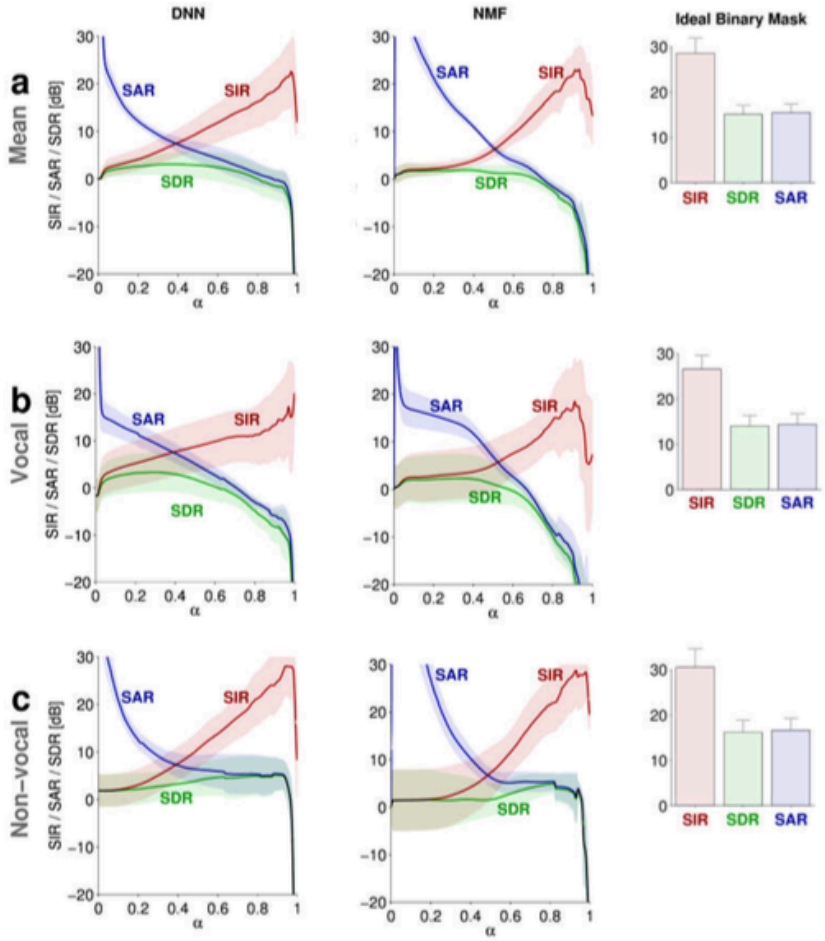
$$M^{V}{}_{t,f} = \begin{cases} 1 & for \quad \frac{1}{T}\sum_{i=0}^{T} S_{t+i,f} > \alpha \\ 0 & for \quad \frac{1}{T}\sum_{i=0}^{T} S_{t+i,f} \leq \alpha \end{cases} \qquad (1)$$

Separation of sources using DNN





Separation quality as a function of alpha



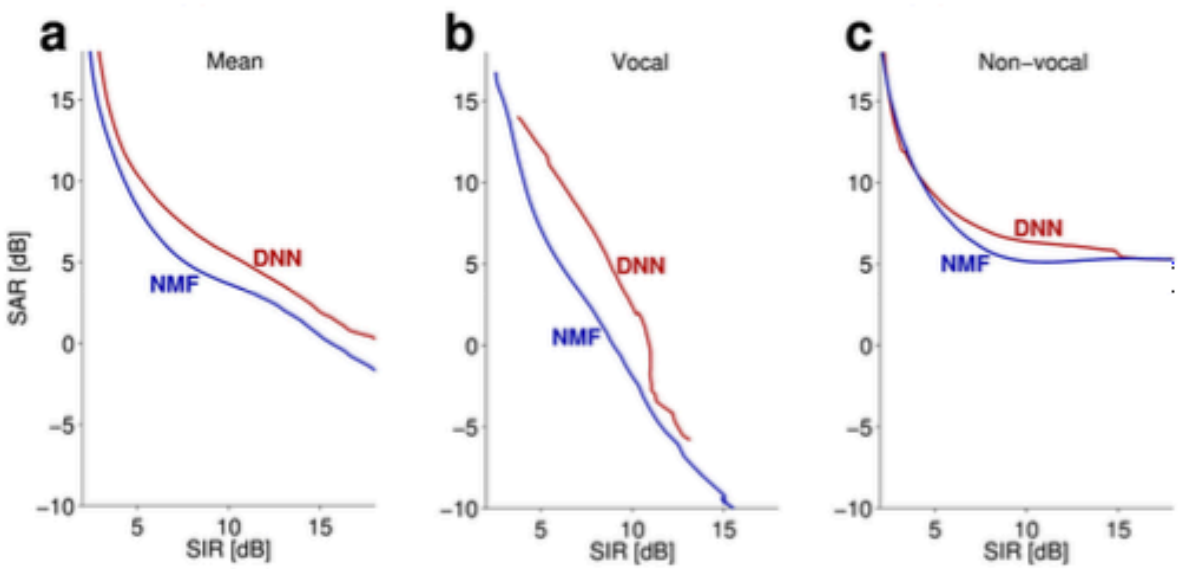
SIR (red) = signal-to-interference ratio

SDR(green) = signal-to-distortion

SAR(blue) = signal-to-artefact

SAR and SIR can be interpreted as energetic equivalents of positive hit rate (SIR) and false positive rate (SAR)

Like-to-like Comparison



Plots mean SAR as a function of mean SIR for both models

DNN provides ~3dB better SAR performance for a given SIR index mean, ~5dB for vocal and and only a small advantage for non-vocal signals

DNN seems to have biased its learnings toward making good predictions via correct positive identification of vocal sounds

Critique of Paper + Next Steps

- DNN seems to have biased its learnings toward making good predictions via correct positive identification of vocal sounds
- Only a small advantage to using DNN vs. traditional approach
- Expand data set