RNN-BASED MUSIC LANGUAGE MODELS FOR IMPROVING AUTOMATIC MUSIC TRANSCRIPTION

Siddharth Sigtia*, Emmanouil Benetos†, Srikanth Cherla†, Arter d'Avila Garcez†, Tillman Weyde† and Simon Dixon*

* EECS, Queen Mary University of London

†Computer Science Department, City University London

Introduction

- We investigate the problem of incorporating **symbolic** priors into automatic music transcription (AMT) systems.
- An accurate model of higher-level symbolic music can potentially help improve transcription by providing a measure of expectations of predicted notes.
- Training accurate statistical models for predicting musical score is a harder problem than language modeling for speech.
- It is not immediately obvious how to combine the two sources of information together into a transcription system.
- We investigate one possible architecture using a recurrent neural network (RNN) **music language model** (MLM) and a spectrogram factorisatoin based acoustic model.
- The proposed architecture results in a 3% improvement in F-measure on the Bach-10 dataset.

Language Modeling

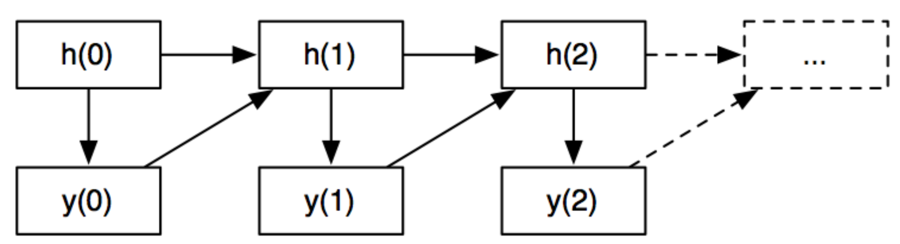

Figure 1: Generative RNN architecture

- RNNs are powerful models for polyphonic music prediction systems [1].
- One limitation is that the outputs of an RNN define unimodal distributions over output variables.
- This assumption of independence is violated by polyphonic music.

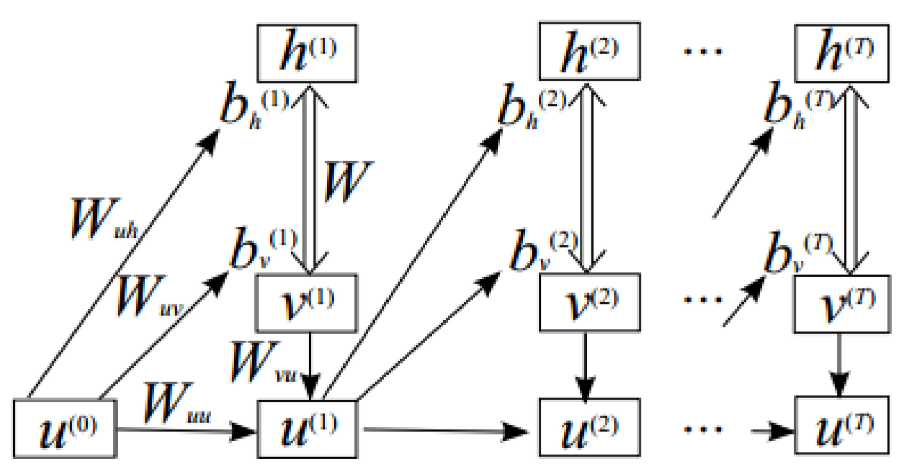

Figure 2: RNN-NADE architecture

- Multimodal conditional distributions can be modeled by allowing the RNN to predict parameters of a high-dimensional density estimator like the RBM or NADE.
- We use the NADE because calculating probabilities is tractable and it can be trained with Hessian Free (HF) optimisation.

Acoustic Modeling

- We utilise the multiple-instrument transcription system based on probabilistic latent component analysis (PLCA).
- The input spectrogram $V_{\omega,t}$ is approximated as $P(\omega,t)$ (ω: frequency, t: time).

PLCA Model

$$P(\omega,t) = P(t)\sum_{p,f,s} P(\omega|s,p,f)P_t(f|p)P_t(s|p)P_t(p)$$

$P(t)$: Signal energy (known quantity).
$P(\omega|s,p,f)$: template for instrument s, pitch p, log-frequency shifting f.
$P_t(f|p)$: Time dependent pitch shifting (semitone range). $P_t(s|p)$: Time dependent source contribution per pitch. $P_t(p)$: Pitch activation probability of p at t.

- The unknown model parameters are estimated using the Expectation-Maximisation (EM) algorithm.
- Using fixed sound templates $P(\omega|s,p,f)$, 20-30 iterations of the EM algorithm are sufficient for convergence.

Proposed Transcription System

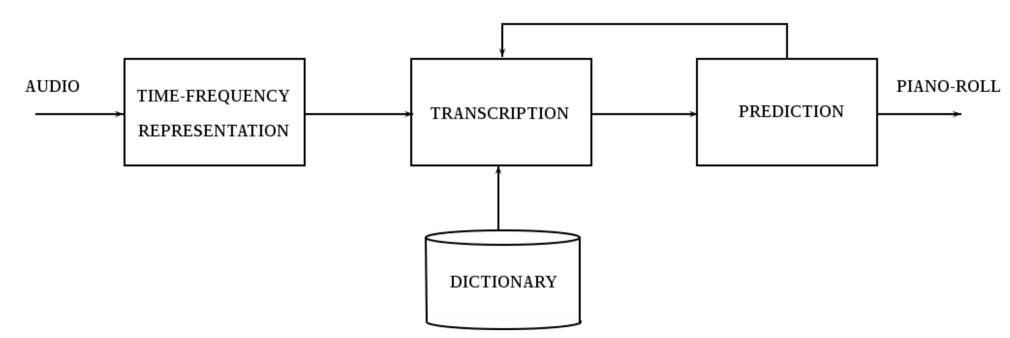

Figure 3: Proposed Transcription Architecture

- The outputs of the PLCA acoustic model form a multinomial distribution.
- The Dirichlet distribution is conjugate to the multinomial distribution and a Dirichlet prior can be used to combine the two sources of information.
- Dirichlet prior for pitch activation: $\alpha_t(p) = P_t(p)P_{MLM}(p,t)$
- The recording is re-transcribed using the following equation.

$$P_t(p) \propto \sum_{\omega,f,s} P_t(p,f,s|\omega)V_{\omega,t} + \kappa\alpha_t(p) \qquad (1)$$

- κ is a parameter that controls the degree of influence of the prior. κ is decressed from 1 to 0 over subsequent iterations.
- Therefore, the transcription yields a symbolic prediction, which improves the subsequent re-transcription of the input.

Validation

Model	*Pre*
RNN (SGD)	67.89%
RNN (HF)	69.61%
RNN-NADE (SGD)	68.89%
RNN-NADE (HF)	**70.61**%

Table 1: Validation results for MLMs

- The language models are trained on the Nottingham dataset, a collection of 1200 folk melodies.
- We evaluate the performance of the RNN and the RNN-NADE models on a music prediction task for validation.
- Both models are trained in 2 different ways; Stochastic Gradient Descent (SGD) and Hessian Free (HF) Optimisation.
- Table 1 enumerates the expected precision for a music prediction task.

Results

Configuration	*F*	*Pre*	*Rec*
Configuration 1	62.02%	58.51%	66.12%
Configuration 2 - NADE	62.62%	59.70%	65.92%
Configuration 3 - NADE	64.08%	61.96%	66.44%
Configuration 2 - RNN	62.29%	59.08%	65.98%
Configuration 3 - RNN	63.85%	61.14%	66.90%
Configuration 2 - NADE-HF	62.20%	59.14%	65.68%
Configuration 3 - NADE-HF	**65.16**%	**62.80**%	**67.78**%
Configuration 2 - RNN-HF	62.44%	59.28%	66.07%
Configuration 3 - RNN-HF	62.87%	60.03%	66.11%

Table 2: Transcription results using various system configurations.

- Transcription experiments are performed on the Bach-10 dataset, a multi-track collection of multiple-instrument polyphonic music.
- We evaluate three different configurations for transcription experiments

Configurations

- Configuration 1: PLCA Acoustic model only.
- Configuration 2: Predictions from acoustic model as inputs to MLM.
- Configuration 3: MLM used to re-transcribe recordings according to Equation 1.

Discussion

- From table 2, we observe the RNN-NADE MLM performs best when used in configuration 3.
- When using the MLMs to provide priors for re-transcription, the F-measure improves by 3% over an acoustic only transcription configuration.
- Training the MLMs with HF appears to help improve transcription results.

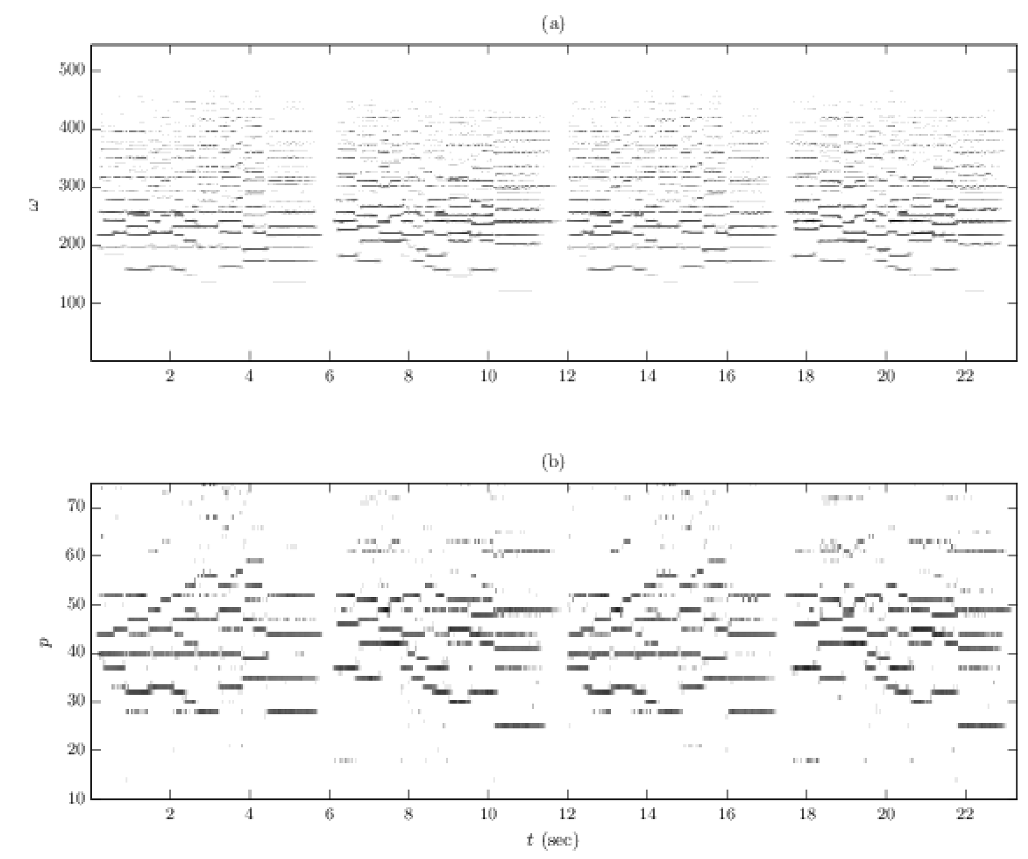

Figure 4: (a) The spectrogram $V_{\omega,t}$ for a recording. (b) The pitch activation $P(p,t)$ using the transcription-prediction system with the NADE-HF.

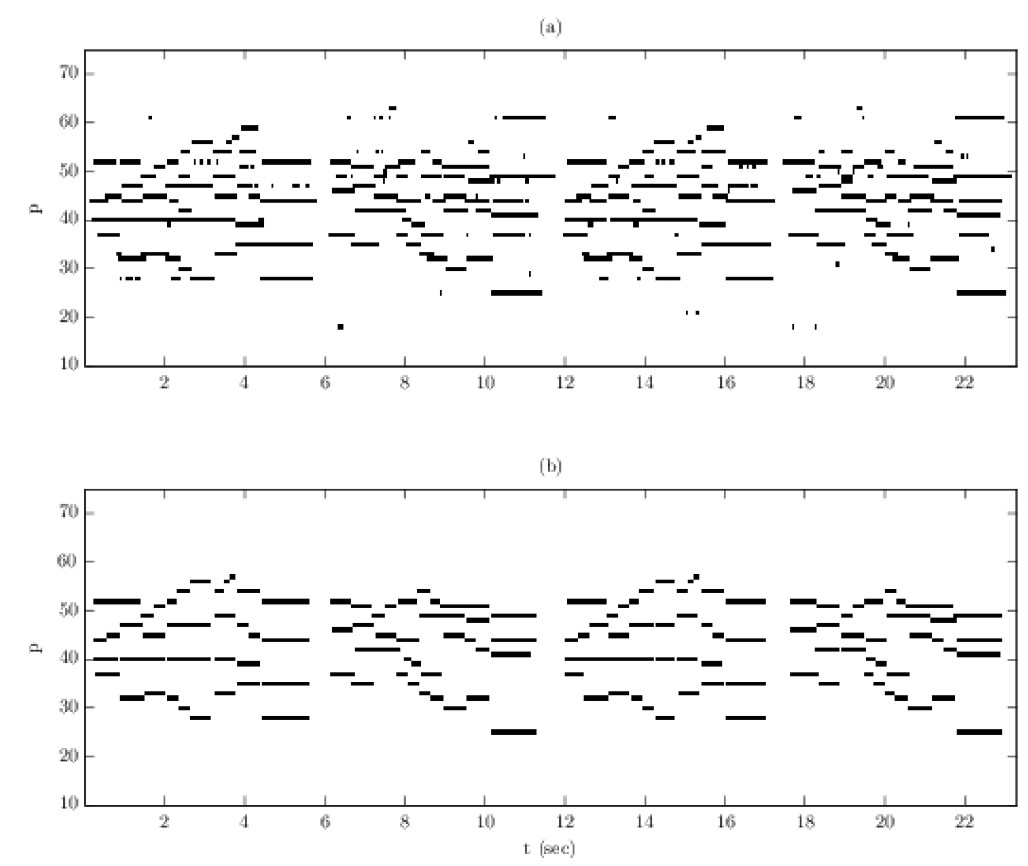

Figure 5: (a) Post-processed output of the transcription-prediction system with the NADE-HF. (b) The pitch ground truth of the recording.

References

[1] Nicolas Boulanger-lewandowski, Yoshua Bengio, and Pascal Vincent.
Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription.
In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1159–1166, 2012.