Dear Reviewers,

Thank you for your reviews. We implemented just about every single suggestion. We appreciate the thoroughness of the reviews as well.

- Re-ran the depth experiment for overfitting (confirmed)
- Renamed appropriate sections
- Implemented the majority of suggestions

We thank you for the time you have spent and we feel we have met all the review comments.

Abram Hindle and Gregory Burlet

Detailed comments are inlined with the review below:

> Dear Authors,
>
> The revised version of the paper is a significant improvement compared to the original submission. Yet, there are still some comments and suggestions by one of the reviewers, which should be taken into account prior to acceptance.
>
> Reviewer 1 (Eric Humphrey)
>
> Basic reporting
>
> Basic reporting looks great. I'd suggest the authors consider renaming the section "Transcription Evaluation" to "Experimental Method", "Experimentation", or something to that effect. Parameter sweeps and tuning are still part of finding the solution that's being evaluated, which roll up to experimentation.

- [X] Rename Transcription Eval to Experimental Method

Response: We renamed the section to Experimental Method and Evaluation.

> Experimental design
>
> Overall, the experimental design is reasonably solid. I do, however, have two smaller comments.
>
> First, regarding L396 on page 12: The text states that "recordings" are partitioned into two sets; is this split conditional on the source songs? For example, imagine two GuitarPro files A and B. A is passed through three guitar models producing A1, A2, and A3, and B is rendered to B1, B2, and B3. Can A and B both occur in the test set? In the spirit of true scientific rigor, audio rendered from the same symbolic source really shouldn't fall across partitions, and I'm curious whether or not this is considered here.

- [X] L396 P 12 Explain that the splits are per song, so we do not train on the same songs we train on and splits are across models

Response: We added more explanation in the paper about the how the preliminary splits work, Song A will be ONLY part of training or ONLY part of test.

> Second, regarding L527-528 on page 16: The authors intend to measure the effect of model depth, i.e. the number of layers, independently, but keep the width, i.e. number of nodes, of each layer constant. This isn't truly an independent assessment, because the number of parameters (and thus model complexity) is certainly increasing. To truly measure the effect of layers, the number of *parameters* should be held constant, as this would offer insight into what may be gained / lost with depth. Otherwise, one would expect that over-fitting will almost certainly happen as the number of parameters increase, especially here given the small size of the dataset and the minimal variance of the sound fonts considered.
>
> That said, I don't think this is an irrelevant experiment, but it's not testing the hypothesis set out by the authors. Rather, I'd be willing to wager that performance on the training and test sets are going in opposite directions here. Reporting performance on the training set (not given) would provide some insight into what's happening here.

- [X] L527 P 1 Clarify that the number of parameters is not held constant and that we might be obvserving overfitting.

Response: We added text to this section to discuss how we could be overfitting and how the number of parameters was not constant because we added constant size layers.

- [X] L527 P 1 Can we rerun with training set performance reported?

Response: We evaluated the training and test performance and found improving training performance versus declining test performance.

> Validity of the findings

> The majority of the findings reported by the authors are substantiated and insightful. There are a few, though, that I would suggest the authors revisit.
>
> L502-503: The authors offer that "steel samples are generally louder than the electric or nylon acoustic samples", and perhaps that's to account for a difference in performance between the reference system (Zhou) and the one proposed here. I'm curious what the rationale would be in that one system would be more or less affected by the gain of a signal? Or how could this hypothesis be tested? After all, symbolic MIDI / GuitarPro files still encode a range of velocity values, no? I'd suspect it has more to do with the timbre of the sounds than the loudness, in that the nylon and electric guitar are "closer" than the steel sound fount (damped overtone series?).

- [X] L 502-503 Mention timbre also are difference. Ask greg.

Response: Regarding the difference between steel and nylon, I think Eric is right about timbre. I believe I (Gregory) normalize the audio samples (or at least I do now) which would offset any major gain differences between soundfonts, so I think it mostly comes down to timbre, which is quite complex and consists of several perceptual attributes of the audio signal. One thing we could mention is that spectral whitening could be used in an attempt to negate the effects of timbre when testing a model trained on one string set versus another. One citation off the top of my head: "where an input signal is first spectrally flattened (â€œwhitenedâ€) in order to suppress timbral information" (Klapuri, 2006) and my source code for this here. In the paper we talked about whitening.

> L522-525: In motivating the exploration of multiple layers in the network (2-4), a parallel is drawn between deep networks and neurobiology. I would strongly advise against trying to make this link. Not only is it debatable, it's an unnecessary distraction that undermines the good work around it. It is sufficient to say "deeper models afford greater representational power and can better model complex acoustic signals" without bringing brains into the mix, and no one can take issue with the claim. Similar comments hold for lines L536-538

- [X] L522-525 536-538 Multiple layers

Response: We editted away some of the neuro-hand-waving. Thanks.

> L548-553: As a conclusion to the same section named above, the authors offer three explanations for an increase in depth leading to decreases in performance: "First, increasing the complexity of the model could have resulted in overfitting the network to the training data. Second, the issue of â€œvanishing gradientsâ€ Bengio et al. (1994) could be occurring in the network fine-tuning training procedure, whereby the training signal passed to lower layers gets lost in the depth of the network. Yet another potential cause of this result is that the pretraining procedure may have found insufficient initial edge weights for networks with increasing numbers of hidden layers."

- [X] L548-553 add pretraining found insufficient intial edge weights for

Response: We added that pretraining found insufficient intial edge weights for to the paper.

> Based on past experience, I'd bet it is exclusively due to the first reason named. All speculation would be easily resolved by including performance numbers over the training set, as well as the test set. If training accuracy increases with model complexity, then we have our answer. In a similar manner, I am suspicious that the vanishing gradient problem is to blame, or that pre-training yielded poor parameters. Again, performance on the training set would shed some light on this. Also, I'd offer that some of the narrative be adjusted to reflect that model complexity, not just depth, is being varied here.

- [X] Collect training performance.

Response: We evaluated the training and test performance and found improving training performance versus declining test performance.

> Comments for the Author
>
> Overall the article is in good shape, and I commend the authors for their diligence in continuing to improve the work. My most important feedback (regarding science and whatnot) is named above, but I've a number of much smaller notes to share. Do with them as you will.
>
> L34: It would be more slightly more convincing to find a more modern reference than (Klapuri, 2004) when referring to the state of the art being so far behind human experts, since it's almost 10 years older than the (Benetos, 2012) citation, which is used as evidence that a monophonic transcription is solved.

- [ ] Find a better reference for polyphonic transcription

Response: We have cited plenty of examples of systems in the rest of paper, the point of the Klapuri quote was that they really said it quite beautifully, even if was in the past. But we understand the concern.

> L81: It's a minor misrepresentation that Humphrey et al, 2012 & 2013 advocate the use of "deep belief networks", which are a specific kind of neural network (RBM pre-training followed by supervised fine

> tuning). Rather, the articles argue for feature learning and deep architectures generically, e.g. CNNs, DBNs, LSTMs, autoencoders, etc.

- [X] Fix reference to Humphrey

Response: We clarified that it was feature engineering with deep architectures.

> L119-127: It's somewhat unclear from this passage that the two systems presented are doing different things on different data, i.e. extensive guitar fingerings from solo guitar recordings versus guitar chord shapes over polyphonic pop/rock music.

- [X] Clarify difference between systems

Response: We added clarification to indicate the difference in difficulty of each system.

> L380-381: Why would MFCCs be a good feature for polyphonic pitch tracking?

- [X] We removed the MFCC reference, it was part of Gregory's thesis. It did not work well, but it was used in speech recognition literature.

> L485-489: Both seem like reasonable kinds of errors. Do you have any insight to the frequency or prevalence of one over the other? Personally I'd expect the duration merging kind to be more common than the thresholding issue, but that's just a hunch.

- [X] L485-489 Address octave errors.

Response: It's not exactly clear other than timbral differences between steel, electric, and nylon. Zhou does not do well on steel either. We have not modified the paper to discuss this.

> L473: Perhaps consider using the Constant-Q transform in future work, which provides a more reasonable trade-off between frequency resolution and time resolution than the DFT.

- [X] Response: Thanks for the suggestion

> L560: It's more accurate to say "faster than real-time", right? It's my understanding that the HMM decoding is non-causal, which means the full signal must be processed before an output can be given at t=0. This would be different from an "on-line" system, as in "as one plays music."

- [X] L560 clarify the system's deep learning aspect is faster than play-time. And in total it is faster than play-time, maybe not streaming low latency.

Response: We clarified in the paper what was meant, that the runtime is less than the playing time. The HMM definitely would increase the latency of the technique.

> L565: Similar to the previous comment, it's a bit of a stretch to claim that the algorithm could be achieved with a microcontroller. It's doubtful that the processing speed seen on a personal computer (with an Intel or AMD processor in the GHz) will translate well to smaller processors with less / slower RAM.

- [X] L575 -- microcontroller argument -- cite integer based bengio paper

Response: There's some interesting work in how to avoid floating point and use binary weighted networks to improve performance. We have cited it and suggested that perhaps it could end up on a microcontroller. Current cell phones have pretty impressive arm processors.

> L568: Also in the realm of tempering optimistic claims, the sentence "All that is required is a set of audio files" is a tad ironic, given how difficult it sounds to obtain data in L355-363 and

- [X] We toned down this claim.

> L590-609. Perhaps something along the lines of "When it's possible to find, curate, or synthesize data, this approach is great."

- [X] L568 We added a similar claim to the paper .

> L587: The approach used in this paper is not early stopping, but a fixed number of iterations. Early stopping requires that some measure (typically over a validation set) is computed as a function of iteration / parameters [https://en.wikipedia.org/wiki/Early_stopping].

- [X] We fixed that in the paper, we said fixed iterations.

> L590: What about sample-based synthesizers? These are at least real sound recordings, rather than algorithmically defined synthesis equations.

- [X] Clarify the dangers of sample based synths

Response: We added some arguement that sample based synths are a little dangerous due to the repetition of the signals due to the low number of samples. It might be quite easy to learn the sound font rather than the actual instrument that was recorded.

Thank you for your reviews!

We've thoroughly addressed them.