Towards a Generalized Performance Tutor

Joe Barrow
jdb7hw@virgnia.edu

Harang Ju
hj4hz@virginia.edu

May 1, 2016

Abstract

1 Introduction

Human instructors are expensive, and have limited time to spend with each student. Classrooms scale poorly; increasing the number of students decreases the teacherś availability for personal instruction. The goal of computer-assisted instruction is to use technology to improve both quality and availability of instruction. In this paper, we outline and implement a system that provides automatic feedback in order to help students improve performance-based activities. Such a system would allow students to practice and receive feedback on performances irrespective of instructor availability or classroom size.

Performance tutoring provides a constrained version of computer-assisted instruction. We define a performance as an activity conducted through time according to a script, or score. Further, the solution outlined in this paper requires that the performance has discrete, separable states as outputs. In this paper, we explore two types of performances: musical performance and pronunciation. In a musical performance, notes act as the discrete, separable outputs. In speech, the discrete outputs are obvious, as sounds dont consistently map to letters in a script and speech tends to run together. Thus, we instead treat phonemes, the atomic sounds that comprise a language, as the discrete outputs of speech.

The overall goal of a performance tutor is to provide feedback based on a performance and a script. The script provides the expected outputs, and a tutor uses it to find areas in which the performer deviated from the script. The minimum requirements for such a system would be to provide precise feedback on exactly where a student made mistakes. An ideal performance tutoring system goes beyond this in two ways: first, it would use information about individual mistakes to find specific areas of focus that would help the student improve, and second, its bank of scripts is easily extensible. One such example of the first system is the 1990 Piano Tutor Project from Dannenberg et al at Carnegie Mellon University. The Piano Tutor Project was built as a holistic tutoring framework, with a lesson bank and an overall awareness of a students progress.

In this paper, we focus more on the generalization of a performance tutoring system. Primarily, we focus on building an architecture that could be extended to almost any type of performance. As a result, however, we investigate how a novel script can be used to judge performances against. The goal is to provide a generalized architecture for feedback generation that can be applied across performance types. To accomplish this, we discuss the specifics of building a piano tutor, and then discuss the necessary adaptations to extend the software, first, to a pronunciation tutor, and second, to much different performance types.

For brevity, musical notes are referred to by their pitch and octave number, e.g. A4 refers to the pitch A in the 4th octave of an 88-key piano. Accidentals are denoted by a # for a sharp, and b for a flat. Similarly, symbols from the International Phonetic Alphabet (IPA) will be used to denote individual phonemes.

2 Overview

3 Segmentation

The first step in the error detection process is to find a maximum likelihood segmentation of the audio data. In an effort to keep segmentation as homogenous as possible across domains, we chose as our recurrent neural networks (RNNs) as our classifier. First, we felt that RNNs would provide smoothness and resiliency against noisy data, as a prediction depends both on the current frame of data and the previous state of the network. For example, the presence of background noise could force a simple frame-wise, or

non-recurrent, predictor to misclassify a frame in the middle of a sustained note, despite the acoustic evidence that the note is being sustained. Second, RNNs have achieved state-of-the-art results on several segmentation tasks, including phoneme segmentation. This allows us to keep a relatively stable architecture across performance domains.

For the piano tutoring system, we trained our classifier on the monophonic subset of the MAPS dataset. The MAPS dataset is a collection of piano notes and transcriptions, and features roughly an hour of high quality monophonic music, including chromatic scales, notes played with different dynamics, repeats, and trills.

In the end, the data was only minimally preprocessed. The audio was split up into overlapping 100ms frames with a hop size of 12.5ms, and the Short-Time Fourier Transform (STFT) was taken of each frame. The length of the frames was necessitated by frequency resolution requirements; at the very bottom of a piano, only about 1.6Hz separates an A#0 from an A0. We attempted several different combinations of feature extraction techniques including: a semitone filterbank, a Constant-Quality Transform (CQT) with a quality-factor set to a single semitone, the log energy of a frame, the RMS energy of a frame, and the first order derivative of the filterbank, CQT, and energies. We achieved the highest accuracy on our validation set when using the raw audio data, so our final design forewent the additional preprocessing steps.

Using the raw audio data as features, we trained a recurrent neural network with a single hidden layer. The hidden layer consisted of 128 Long-Short Term Memory (LSTM) cells, and the output of the network was a probability distribution over 89 classes, one for each note on the piano and one for silence. After training, the audio was then classified by selecting the maximum likelihood class at every timestep. Over a testing set of the MAPS database, representing a randomly sampled 10% of the audio, our RNN achieved an average of 82.85% accuracy on the testing set.

4 Error Finding

4.1 Hidden Markov Model

4.2 Alternative Models

4.3 Providing Feedback

5 Generating Examples

5.1 From Experts

5.2 From Scores

6 Discussion

6.1 Extending to Language

6.2 Extending to Other Performance Types

7 Related Work

Performance instruction is not a new field of research. Both music and pronunciation tutoring tools have been well-researched. In the early 1990s, researchers from Carnegie Mellon University built and tested a holistic piano instruction system called the Piano Tutor Project. However, the research done by the Piano Tutor team was focused mostly on lesson selection and curriculum analysis, which is reflected in the published works. Computer-Assisted Pronunciation Training (CAPT) is a well-researched field, both in its effectiveness and its implementation.

Similarly, there are several commercial systems for both music and pronunciation tutoring. In music, there is the recently released Yousician, which provides lessons for piano, guitar, bass, and ukelele. Yousician performs computations in real-time, but its errors are binary; Yousician only cares if you did or did not play the correct note when they were expecting it. For pronunciation training, English Computerized Learning provides accent reduction software as well as hours of lessons and other content.

However, we believe that the system outlined in this paper has a number of advantages over the related works. Our system theoretically generalizes to almost any performance-based activity, and allows for scripts to be quickly and accurately generated either by experts or algorithms.

8 Conclusions and Future Work

8.1 Increasing the Dataset

We believe that the classifier can achieve a much higher accuracy if given enough monophonic data as input. MAPS currently provides roughly an hour of monophonic music data, but the data is clean and lacks actual songs. More, and more realistic, data could improve the usefulness of a piano tutor in real-world environments. Additionally, more data could allow us to use a bidirectional RNN, which could be instrumental in solving the errors brought about by the sound a piano key makes when being struck.

8.2 Examples From Scores

In the future, we want to be able to construct an HMM directly from a printed score of music. This greatly increases the potential of the system, as thousands of HMMs could be constructed quickly added from scanned music.

8.3 Extensions to Video

For this to be a truly generalized performance tutor, it must be able to handle all different classes of performances. The architecture outlined in this paper can be generalized to many different auditory or written performances, but there the classifier must be changed if it is to work for video performances. Video performances could include any type of action with discrete states, including martial arts routines or yoga. We propose that a recurrent convolutional neural network, if provided with sufficient training data, could be substituted out for the current fully-connected RNN, and thus require minimal changes to the overall architecture. Expert scores could then be generated as a list of moves and their durations, which is similar to the data contained in a MIDI file.

9 Availability

The source code, as well as the history of the project, is being hosted on GitHub, and can be found at: http://github.com/jbarrow/Capstone/

10 Acknowledgements

We would like to thank Professor Kevin Sullivan from the University of Virginia for the unyielding support and guidance he has provided throughout the course of the project.

11 Works Cited