Deep Auditory Hallucinations: Multi-Modal Music Generation from Actions

Si Chen
Georgia Institute of Technology
sichen@gatech.edu

Abstract

Music education has proven academic and social benefits, but is costly and time intensive. This paper aims to hallucinate an instrument based on the marimba from the user's actions. To this end, we present a novel dataset, current progress in generating pitched notes that temporally correspond to the action, velocity, and location of a strike, and future direction for the generation of music from video.

1. Introduction

Multi-modal learning is an integral human ability; transferring learned concepts from one modality to another with an independent abstract representation gives us great flexibility in applying learned concepts to new tasks. For example, while most do not know how to play the piano, it is easy to mimic pressing keys and visually comprehend the continuous progression of pitch on a keyboard. The ability to generate across modalities enhances human creativity by enabling non-experts to engage with the arts at a higher level and gain the benefits that come with creative fields.

In United States high schools, participating in music ensembles is a privilege segregated by race, socioeconomic, and immigration status [1]. The learning curve for musical instruments is steep and the price of involvement can be extremely prohibitive; as a result, the students who are able to reap the academic and social benefits of a music education are disproportionately represented in the highest socioeconomic classes of the American high school population [3].

In order to address this disparity, we look to make music education more accessible. This paper looks to simplify the skills necessary to learn instruments into a gesture-based action recognition and music generation system so that non-musicians can focus on the creativity behind musicianship rather than the economic hurdles that are a significant aspect of secondary music education.

2. Deep Action-Based Instruments

While recent work [4] shows promise in moving from the visual to the auditory domain, it focuses on noise-like audio synthesis from objects being struck. Our work focuses on generating sounds in a similar, but distinct, context: synthesizing pitched audio from actions. For non-musicians who are familiar with the general motions of instruments but have no expertise, we believe that miming the actions of playing an instrument and hearing the representative sounds is an incredibly compelling creative experience.

Music education in Venezuela's world-renowned *El Sistema* program begins with children learning the basics of body expression and rhythm before they ever touch an instrument [6]. With a vision-based instrument that understands actions, spatial relevance, and color's correlation with instrument materials, not only would adult non-musicians be able to interact with music using various household objects as pieces of a novel instrument, but children would be able to engage with the physicality of music education even when traditional instruments are not available. By distilling the complexities of music education into visual attributes, we hope to democratize the ways in which a music education can be pursued. This project takes the first initiative towards this by correlating pitch (audio vectors) with actions (RGB video) across viewpoints.

2.1. Dataset Collection

A novel dataset was created comprising of multi-view videos and audio of a marimba being struck. Unlike other instruments, the musical information produced by a marimba is easily encoded visually. Unlike a violin, where a millimeter difference on the fingerboard can produce different pitches and multiple fingers occlude one another, a marimba has discrete pitches and only involves the use of a mallet (or several mallets) that do not occlude the majority of the view. In a marimba, pitch is 1) spatially relative, with lower to higher pitches transitioning smoothly across the instrument and 2) geometrically relative, with larger bars producing lower tones than smaller bars. A note's 1) decay is inferred by the length of time the mallet stays on the bar, 2) amplitude is linked with velocity of the mallet, and 3) tim-

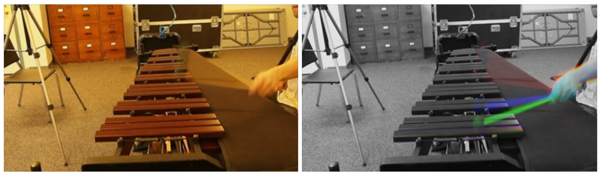

Figure 1. RGB frame (left) and space-time computation (right).

ber is correlated with the instrument color (rich, sonorous – brown wooden marimba; glistening, gossamer – grey metal vibraphone).

The resulting dataset consists of videos from six distinct viewpoints of the marimba being hit at random velocities, locations, and decay lengths. In our initial experiments, these six angles are synced to the same audio recording and the dataset consists of approximately 250,000 frames.

2.2. Network Architecture

The initial architecture explored expands upon [4], using AlexNet pre-trained on Imagenet, but fine-tuned on the marimba dataset, to predict audio features from RGB frames from the test set and space-time images. Space-time images, computed by concatenating the previous, current, and subsequent frame as greyscale images into a single RGB image, is a simple method of incorporating temporal information into our network that is more efficient than approaches such as optical flow. The audio feature predicted is a vector of size 18, with each index representing the percentage that each note contributes to the overall produced sound. From this audio feature, we can both use nearest neighbor search to find the nearest note in our database of sounds as well as synthesize the note using 1) frequency modulation synthesis or 2) LSTM-based temporal generative methods for instruments with more complex sounds. While methods such as WaveNet [5] have gained notoriety, these simpler methods of audio generation are solid baselines for comparison, especially for the marimba, whose waveform is simpler than that of other instruments.

2.3. Results & Future Work

From this architecture, 2 shows the ability of our network to predict the audio feature vector regression accurately across ten seconds of test frames. While these results are encouraging, we hope to incorporate temporal information with a wider field of view across frames [7] on this dataset before moving to more complex musical instruments, such as the violin and guitar. Additionally, we hope to analyze videos of professional musicians and correlate their fine-grained actions with emotive qualities of their music, incorporating more fine-grained action recognition into a broader music generation methodology that can recurrently create melodies dependent on the physical expression of emotions. The ultimate goal of this work is to create simplified musical representations of these instruments that can be performed by non-musicians.

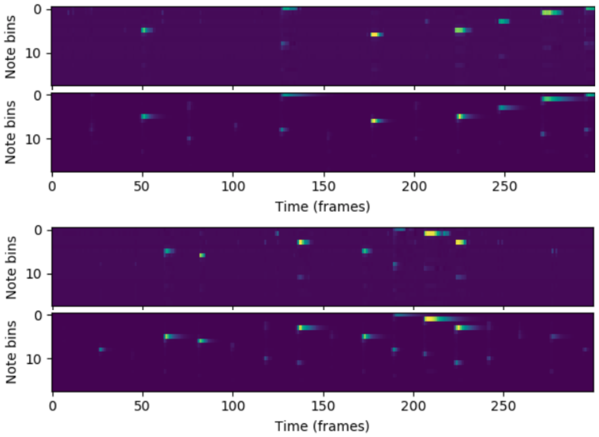

Figure 2. Two preliminary audio feature regressions: the top graph of each pair is the regression, the bottom is the ground truth.

3. Conclusion

While Americans are avid consumers of the arts, only 12% of adults practice an instrument and many adults feel hindered by their lack of formal education in this field [2]. Given the importance of music education in adolescent development, the contribution of this generative multi-modal learning method across the audio-visual domains is twofold: 1) it creates latent representations of complex musical ideas (e.g. timbre) and correlates them with actions 2) it has important sociological implications in providing access to music education to non-musicians.

References

[1] K. Elpus and C. R. Abril. High school music ensemble students in the united states a demographic profile. *Journal of Research in Music Education*, 59(2):128–145, 2011.

[2] N. E. for the Arts. How a nation engages with art. 2013.

[3] S. Hallam. The power of music: Its impact on the intellectual, social and personal development of children and young people. *International Journal of Music Education*, 28(3):269–289, 2010.

[4] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016.

[5] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang. Fast wavenet generation algorithm. *arXiv preprint arXiv:1611.09482*, 2016.

[6] E. Sistema. El sistema in venezuela. 2014.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.