Audio To Body Dynamics

Eli Shlizerman, Lucio Dery, Hayden Schoen, Ira Kemelmacher-Shlizerman

E-mail: shlizee@fb.com, derylucio@fb.com, haydenschoen@fb.com kemelmi@fb.com

Abstract. We present a method that gets as input an audio of violin or piano playing, and outputs a video of skeleton predictions which are further used to animate an avatar. The key idea is to create an animation of an avatar that moves their hands similarly to how a pianist or violinist would do, just from audio. Aiming for a fully detailed correct arms and fingers motion is a goal, however, its not clear if body movement can be predicted from music at all. In this paper, we present the first result that shows that natural body dynamics can be predicted at all. We built an LSTM network that is trained on violin and piano recital videos uploaded to the Internet. The predicted points are applied onto a rigged avatar to create the animation.

1. Introduction

Traditionally, researchers have tackled the problem of lip syncing from speech [1] since the correlation between spoken audio and lip movements is stronger. In this work, we seek to explore whether body movement can be predicted computationally from a music signal. To achieve this, we gather training data from a given video of a person playing a piano or violin by featurizing the audio over a per-frame time horizon and extracting body landmarks from frames using a pose-estimation model. We then train a LSTM model that learns a transformation from audio features to body landmarks. We train separate networks for violin and piano.

2. Method

2.1. Data set

We construct a data set of videos downloaded from the Internet. Selected videos range from 3 min to 50min. We found that using recitals or single person shows are optimal sets. Total of 3.6 hours of violin recitals was collected and 4.4 hours of piano recitals.

2.2. Audio Preprocessing

Mel Frequency Cepstral Coefficents (MFCC) features have been shown to be useful a wide range of audio tasks including classifying and identifying different musical instruments [2]. For this problem, we compute the features on stereo 44.1Khz sample rate audio, perform RMS normalization to 0 db using FFMPEG, and choose the window length as 1000/videofps with fps= 24, i.e., 41.66ms.

2.3. Keypoint Estimation

We generate keypoints by running OpenPose which provides face, body, and hands keypoints [3] and MaskRCNN [4], on selected frames. We use the DeepFace [5] face recognition algorithm

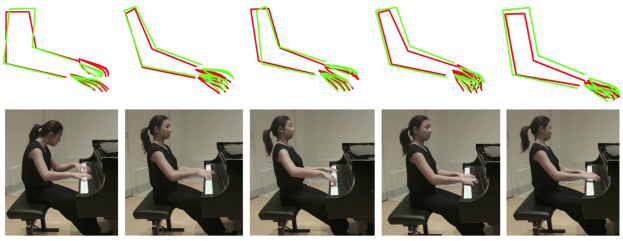

Figure 1: Piano test results. In row 1 we show predicted points from audio (in green) overlaid on top of ground truth points (red). Row 2 shows the corresponding frame for context. Note that we dont expect for them to fit exactly but just aiming for similar hands and fingers configurations

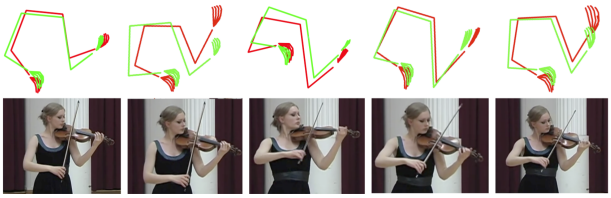

Figure 2: Violin test results. We encourage the reader to watch the supplementary videos (with audio on) too. In row 1 we show predicted points from audio (in green) overlaid on top of ground truth points (red). Row 2 shows the corresponding frame for context.

to select frames that contain the person of interest. We account for full body translation, or rotation across frames by transforming to a cordinate set centered on a reference set of keypoints. We perform PCA on the selected keypoints.

2.4. Audio Features to Keypoints

We chose to use a unidirectional single layer LSTM with time delay (inputs lag behind outputs). We also add a fully connected layer at the output which we found to increase performance. The parameters that we used are hidden state of 200, trained with truncated back propagation with time steps of 400, time delay of 5 timesteps, dropout of 0.4, learning rate of $5e^{-3}$. The number of PCA components we typically use is 10. We ran the training for 300 epochs. The network is implemented in Caffe2, and uses ADAM optimizer

3. Results

Figures 2 and 1 show the results of training our LSTM. More video examples can be found here

References

[1] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[2] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of mfcc. *Journal of Computer science and Technology*, 16(6):582–589, 2001.

[3] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. computer vision and pattern recognition (cvpr). In *2016 IEEE Conference on*, volume 2, 2016.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[5] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.