Audio to transform (translation, rotation, scale) of an Actors Bone

Md.Zeeshan Data Scientist, ML & DL developer from ramstein.ml

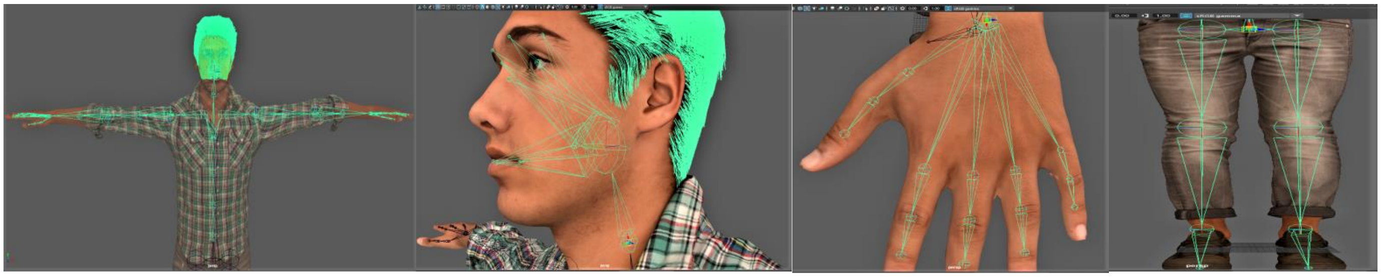


Fig.1 Given input audio and a reference mocap data, we predicts the transforms of all bones

Predicted audio from **dt_tts** model, we predict the transforms or floats of actor bones specifically of face for deriving the animation. Trained on many hours of mocap data, a convolutional neural network learns the mapping from raw audio features to bone transforms. We synthesize 3D bone transforms and animate it with proper textured 3D actors pose matching to change what he appears to be saying in a target animation to match the input audio track. Our approach produces three-dimensional realistic results.

CCS Concepts: *Computing methodologies: Bone transform based rendering, Animation, Rigging, Shape modeling,

Additional keywords and phrases: Transforms, Maya, Unreal Engine, Tesla P100 GPU, mocap,

ACM Reference format:
Md.Zeeshan Data Scientist, ML & DL developer
Audio to transform (translation, rotation, scale) of actors bone
DOI:

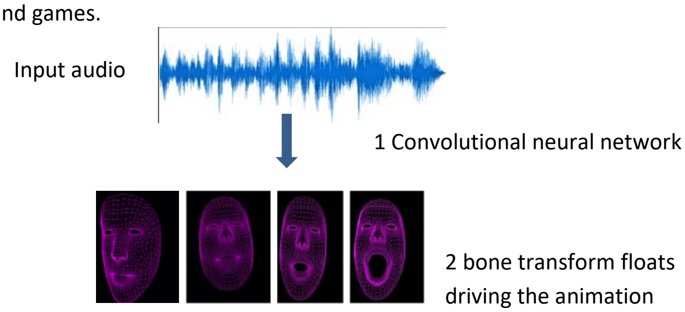
Abstract

In this futuristic era, everyone has heard of the faceAR, DeepFakes, FaceApp and many more. Why not, we are using these things to increase our understanding of any unknown context. Why not, we are using it in our education system. For solving the big hurdle behind the project eduFUTURE.ai i.e. mapping the mocap (motion-captured) data of a professor delivering the lecture in a class to a 3D modeled, rigged animated meshed character, we have created a convolutional neural network model that learns the mapping from mocap data to actors bone transforms. From experiments, it has been seen that CNN outperforms the RNN even at the sequence to sequence task.

Video lectures are present in abundance amount but the mocap data of those video lectures is 10 times ahead in the form of precise data. High quality and a large amount of data are one of the requirements of best argmax predicting ML model, so we have used here the mocap data.

Despite the availability of such promising data, the problem of generating bone transforms from audio is extremely difficult, due in part to the technical challenge of mapping from a 1D signal to a 3D transform (translation, rotation, scale) float values, but also due to the fact that humans are extremely attuned to subtle details in expressing emotions; many previous attempts at simulating talking body have produced results that look uncanny. In addition to generating realistic results, this paper represents the first attempt to solve the audio speech to character bone transform prediction problem by analyzing a large corpus of mocap data of a single person. As such, it opens to the door to modeling other public figures, or any 3D character (through analyzing mocap data).

Audio to bone transform, aside from being interesting purely from a scientific standpoint, has a range of important practical applications. The ability to generate high-quality textured 3D character from audio could significantly reduce the amount of bandwidth needed in video coding/transmission (which makes up a large percentage of current internet bandwidth). For hearing-impaired people, video synthesis from bone transform could enable lip-reading from over-the-phone audio. And digital humans are central to entertainment applications like movies special effects and games.



Input audio

1 Convolutional neural network

2 bone transform floats driving the animation

Our approach is based on synthesizing animation of the whole body from audio. For that, we have trained our model on raw audio data and captured motion data as the label. After an exhaustive amount of training on Tesla P100, we are getting here astonishing results on inferencing the model. Predicted bone transforms is being directly imported into a UE4 character through the MAYA_Live_LInk 2018 plugin or for directly into the Unreal Engine, we can use the block Transform(modify) Bone in the blueprint editor for transforming the x, y, z coordinates of the bone and that'll finally drive the animation.

2 Related Works:

2.1: Mode-Adaptive Neural Networks for Quadruped Motion Control by HE ZHANG† , University of Edinburgh SEBASTIAN STARKE†

Motion control using neural networks has recently demonstrated success in producing high-quality animation for biped locomotion with clear cycles [Holden et al. 2017]. However, merely applying the same framework to quadrupeds fails because defining single phase for all four legs is not possible in the transition between gaits with very distinct footfall patterns. This also makes manual phase labeling of unstructured quadruped motion data with complex gait transitions impractical.

In this paper, we propose a novel network architecture called Mode-Adaptive Neural Networks (MANN) that can learn a locomotion controller from a large amount of unstructured quadruped motion capture data. The system is composed of the motion prediction network and the gating network. At each frame, the motion prediction network computes the character state in the current frame given the state in the previous frame and the user-provided control signals. The gating network dynamically updates the weights of the motion prediction network by selecting and blending what we call the expert weights, each of which specializes in a particular movement. This architecture provides flexibility such that the system can learn consistent features across a wide range of non-periodic actions and periodic unlabeled gait types. This framework can release the developers from the tedious and difficult process of phase labeling, where the unstructured quadruped motion capture data of different gait types must be aligned along the timeline. In particular, our model does not require individual labels for different gaits which are often difficult to distinguish even for humans, and thus avoids gait mislabeling during data preprocessing.

The contributions of the paper can be summarized as follows:

- The first systematic approach for constructing data-driven quadruped character controllers that can synthesize animations in production-quality with a wide variety of locomotion modes and transitions between them.
- A novel end-to-end neural network architecture that can learn from unstructured quadruped motion capture data without providing labels of the phase and locomotion gaits.
- A comprehensive evaluation of the proposed architecture through comparison with existing approaches.

2.2: Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion by TERO KARRAS, NVIDIA and with low latency. Our deep neural network learns a mapping from input waveforms to the 3D vertex coordinates of a face model, and simultaneously discovers a compact, latent code that disambiguates the variations in facial expression that cannot be explained by the audio alone. During inference, the latent code can be used as an intuitive control for the emotional state of the face puppet. We train our network with 3–5 minutes of high-quality animation data obtained using traditional, vision-based performance capture methods. Even though our primary goal is to model the speaking style of a single actor, our model yields reasonable results even when driven with audio from other speakers with different gender, accent, or language, as we demonstrate with a user study. The results are applicable to in-game dialogue, low-cost localization, virtual reality avatars, and telepresence.

2.3: A Deep Learning Approach for Generalized Speech Animation SARAH TAYLOR, University of East Anglia TAEHWAN KIM, YISONG YUE, California Institute of Technology MOSHE MAHLER, JAMES KRAHE, ANASTASIO GARCIA RODRIGUEZ, Disney Research JESSICA HODGINS, Carnegie Mellon University IAIN MATTHEWS, Disney Research

We introduce a simple and effective deep learning approach to automatically generate natural looking speech animation that synchronizes to input speech. Our approach uses a sliding window predictor that learns arbitrary nonlinear mappings from phoneme label input sequences to mouth movements in a way that accurately captures natural motion and visual coarticulation effects. Our deep learning approach enjoys several attractive properties: it runs in real-time, requires minimal parameter tuning, generalizes well to novel input speech sequences, is easily edited to create stylized and emotional speech, and is compatible with existing animation retargeting approaches. One important focus of our work is to develop an effective approach for speech animation that can be easily integrated into existing production pipelines. We provide a detailed description of our end-to-end approach, including machine learning design decisions. Generalized speech animation results are demonstrated over a wide range of animation clips on a variety of characters and voices, including singing and foreign language input. Our approach can also generate on-demand speech animation in real-time from user speech input.

2.4: Synthesizing Obama: Learning Lip Sync from Audio SUPASORN SUWAJANAKORN, STEVEN M. SEITZ, and IRA KEMELMACHER-SHLIZERMAN, University of Washington

Given audio of President Barack Obama, we synthesize a high quality video of him speaking with accurate lip sync, composited into a target video clip. Trained on many hours of his weekly address footage, a recurrent neural network learns the mapping from raw audio features to mouth shapes. Given the mouth shape at each time instant, we synthesize high quality mouth texture, and composite it with proper 3D pose matching to change what he appears to be saying in a target video to match the input audio track. Our approach produces photorealistic results.

3 AUDIO TO VIDEO

Given a source audio track of a professor delivering lecture, we seek to synthesize a corresponding bone transforms. To achieve this capability, we propose to train on many hours of motion captured data of professors to learn how to map audio input to bone transforms output. ⍰

This problem may be thought of as learning a sequence to sequence mapping, from audio to bone transforms, that is tailored for one specific individual. This problem is challenging both due both to the fact that mapping goes from a lower dimensional (audio) to a higher 3-dimensional (transforms) float values, but also the need to avoid the uncanny valley, as humans are highly attuned to lip motion.

To make the problem easier, we focus on synthesizing the parts of the face that are most correlated to speech. At least for delivering the lectures, we have found that the content of any lecture correlates most strongly to the region around the mouth (lips, cheeks, and chin), and also aspects of head motion – professors head stops moving when they pauses their speech.

The overall pipeline works as follows (Fig. 2): Given an audio from dt_tts model, we first extract audio features to use as input to a convolutional neural network that outputs bone transform for every frame mapped to 3D character(section 3.1).

3.1: Creating mocap data synced with audio: