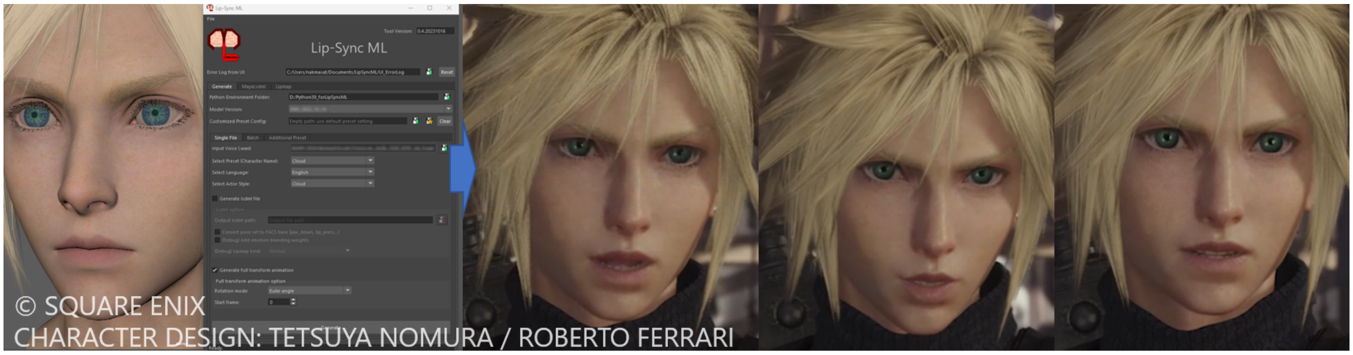
Lip-Sync ML: Machine Learning-based Framework to Generate Lip-sync Animations in FINAL FANTASY VII REBIRTH

Masato Nakada
SQUARE ENIX CO., LTD.
Tokyo, Japan
nakmasat@squre-enix.com

Leandro Graciá Gil
SQUARE ENIX CO., LTD.
Tokyo, Japan
graclean@squre-enix.com

Akira Iwasawa
SQUARE ENIX CO., LTD.
Tokyo, Japan
aiwasawa@squre-enix.com

Ryo Hara
SQUARE ENIX CO., LTD.
Tokyo, Japan
hararyo@squre-enix.com



Tool Version: 0.4.20231016

Error Log from UI: C:/Users/nakmasat/Documents/LipSyncML/UI_ErrorLog

Python Environment Folder: D:/Python39_forLipSyncML

Model Version:

Customized Preset Config: Empty path: use default preset setting.

Input Voice (.wav):

Select Preset (Character Name): Cloud

Select Language: English

Select Actor Style: Cloud

□ Generate lsdml file

Output lsdml path: Output file path

□ Convert pose set to FACS base (jaw_down, lip_press...)

□ (Debug) Add emotion blending weights

(Debug) Lipmap kind: Normal

☑ Generate full transform animation

Full transform animation option

Rotation mode: Euler angle

Start frame: 0

© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI

Figure 1: Lip-sync animation generated by Lip-Sync ML. From the voice audio of a character, the tool (left) can generate the animation, and it can be played in a game engine (right.)

ACM Reference Format:
Masato Nakada, Leandro Graciá Gil, Akira Iwasawa, and Ryo Hara. 2024. Lip-Sync ML: Machine Learning-based Framework to Generate Lip-sync Animations in FINAL FANTASY VII REBIRTH. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH Talks '24), July 27 - August 01, 2024.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3641233.3664723

1 INTRODUCTION

In recent years, many AAA video game titles have been released, bringing new characters with them. These characters utter memorable and engaging lines, and lip-sync animations are crucial to make these utterances believable. In FINAL FANTASY VII REBIRTH (hereinafter called *the current title*), creating high-quality lip-sync animations were essential to pursue the realism of the characters. Meanwhile, a large number of character voices was created, so we had to find efficient ways to create lip-sync animations for them.

We present a machine learning-based framework called Lip-Sync ML to generate lip-sync animations. This framework can generate lip-sync animations from only audio input using machine learning. We used lip-sync animations from cutscenes in our previous game title: FINAL FANTASY VII REMAKE (hereinafter called *the previous title*) as training data.

2 PREVIOUS WORKS

Phoneme-based methods such as JALI[Edwards et al. 2016] were proposed in the past. They require audio and dialogue text as input to obtain time-series data of phonemes to match the audio. According to the timing of the phoneme, they express a lip-sync animation by playing poses that correspond to phonemes.

In *the previous title*, we used a phoneme-based method to create lip-sync animations for simple event scenes. We used phoneme alignment by Sppas[Bigi 2015], and mouth poses are blended according to the phonemes to play a lip-sync animation. We treat the collection of mouth poses used for this blending as an asset called *Lipmap*. Fig. 2 shows some poses stored in a Lipmap. It stores bone transformations for each pose.

However, it became problematic when there was an improvised voice or a breathing sound that was not in the dialogue text. In this case, the phonemes would not be placed correctly, resulting in wrong poses which did not match the voice. Manual adjustments were required to avoid this problem.

In recent years, machine learning-based methods such as NVIDIA's Omniverse Audio2Face[Karras et al. 2017] have also been proposed. Our method belongs to these. One of the fundamental differences

SIGGRAPH Talks '24, July 27 - August 01, 2024, Denver, CO, USA
The definitive Version of Record was published in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH Talks '24), July 27 - August 01, 2024*, https://doi.org/10.1145/3641233.3664723.

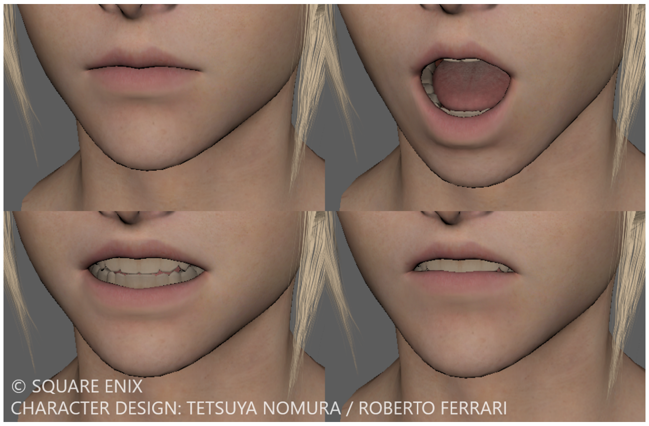


Figure 2: Some poses stored in a Lipmap. It has 8 poses including the ones shown above.

between Audio2Face and our system is the format of the output animations. Audio2Face generates animations as vertex displacements, which is generally not supported in game engines.

3 MACHINE LEARNING

We used synchronized audio and lip-sync animation data from the cutscenes of *the previous title* to train a machine learning model that transforms input audio into bone 3D transforms and pose weights.

Our model first transforms audio into log mel spectrograms, and processes them as images by applying custom residual convolutional networks in a way inspired by how both absolute and relative pitch work. We iteratively reduce the frequency dimension of the data by using strides during convolutions, while also applying dilations at different scales. The resulting features are then processed using a Transformer model with relative attention[Huang et al. 2018] to produce voice features.

These voice features are then grouped and averaged to match the desired frames per second of the animation, and a learnable style system is applied to them. This style system allows us to control different aspects of the generated animation, such as target character or language, in order to get more refined results.

From this, our machine learning model generates two kinds of animation outputs: high quality bone-based animation, and light and versatile pose weights animation to blend poses in a Lipmap. Generated pose weights are saved as an asset called *Lsdml* (Lip-Sync Data from Machine Learning.) To generate high quality bone-based animation, we select a target facial skeleton and add the generated results to the character's resting pose for every bone and frame. To train pose weights, we blend their Lipmap poses with the generated weights and compare the result with the training animation data. This is possible because pose blending is a differentiable operation.

4 WORKFLOW

To make this machine learning model easier to use, we developed a tool for Autodesk Maya®[Autodesk, INC. 2024]. This tool can generate lip-sync animations from input audio using the machine learning model, edit the generated assets, and export them within Maya. We also developed a plugin called *LsdmlPlayer* that allows us to play lip-sync animations from Lsdml and Lipmap assets in a game engine. Fig. 3 shows the workflow of using the tool and the plugin. We would like to present them in more detail and show how they are used in practice during our talk.

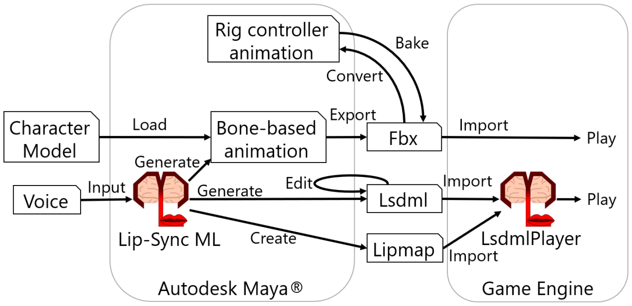


Figure 3: Workflow with Lip-Sync ML

5 USE CASE

In *the current title*, we developed another tool to convert generated bone-based animation into animation for the controllers of our character rig. This made possible for animators to use our generated animations as the starting point for creating lip-sync animations for cutscenes.

Pose weight animations are used for the lip-sync animation of simple event scenes, battle actions, and field actions. In *the previous title*, which featured approximately 96,000 audio clips in simple event scenes and action sequences, we had to manually correct the generated lip-sync animations of over 100 audio clips. In contrast, with the implementation of Lip-Sync ML in *the current title*, the generated animations of only 10 to 20 audio clips required manual fixes despite having about 1.4 times more audio clips in similar scenes and action sequences compared to *the previous title*.

6 CONCLUSION

The machine learning-based framework Lip-Sync ML can generate lip-sync animations automatically from only audio input. It can match mouth poses to improvised voices and breathing sounds. By using Lip-Sync ML in addition to the work done by animators, we were able to create believable lip-sync animations for characters in FINAL FANTASY VII REBIRTH efficiently.

REFERENCES

Autodesk, INC. 2024. *Maya.* https://www.autodesk.com/products/maya/overview

Brigitte Bigi. 2015. SPPAS - MULTI-LINGUAL APPROACHES TO THE AUTOMATIC ANNOTATION OF SPEECH. *The Phonetician. Journal of the International Society of Phonetic Sciences* 111-112, ISSN:0741-6164 (2015), 54–69. https://hal.science/hal-01417876

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization. *ACM Trans. Graph.* 35, 4, Article 127 (jul 2016), 11 pages. https://doi.org/10.1145/2897824.2925984

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. arXiv:1809.04281 [cs.LG]

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.* 36, 4, Article 94 (jul 2017), 12 pages. https://doi.org/10.1145/3072959.3073658