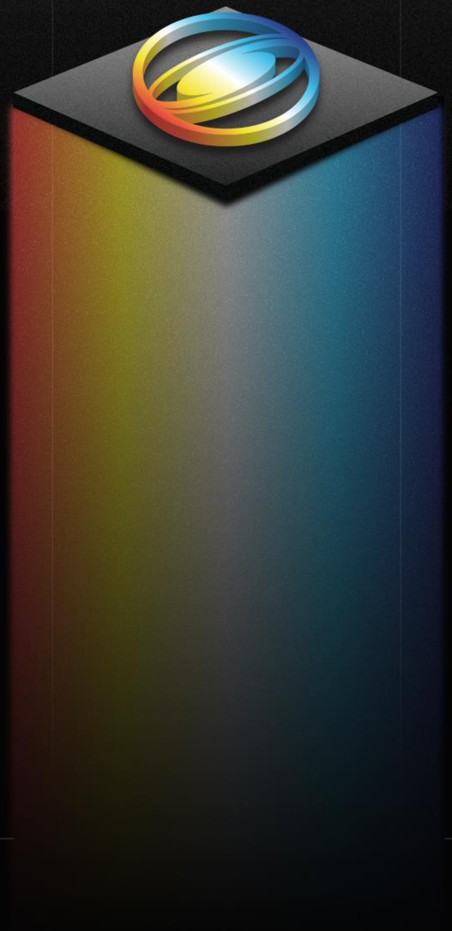
SIGGRAPH 2024

DENVER+ 28 JUL — 1 AUG

THE PREMIER CONFERENCE
& EXHIBITION ON
COMPUTER GRAPHICS &
INTERACTIVE TECHNIQUES

LIP-SYNC ML:

MACHINE LEARNING-BASED FRAMEWORK TO GENERATE LIP-SYNC ANIMATIONS IN FINAL FANTASY VII REBIRTH



AGENDA

- Introduction
- Workflow
- Results
- Machine learning details
- Conclusion

MASATO NAKADA

SQUARE ENIX CO., LTD.

INNOVATION TECHNOLOGY
DEVELOPMENT DIVISION

R&D ENGINEER

LEANDRO GRACIÁ GIL

SQUARE ENIX CO., LTD.

INNOVATION TECHNOLOGY
DEVELOPMENT DIVISION

SENIOR MACHINE LEARNING EXPERT

FINAL FANTASY VII REBIRTH (2024)

© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI

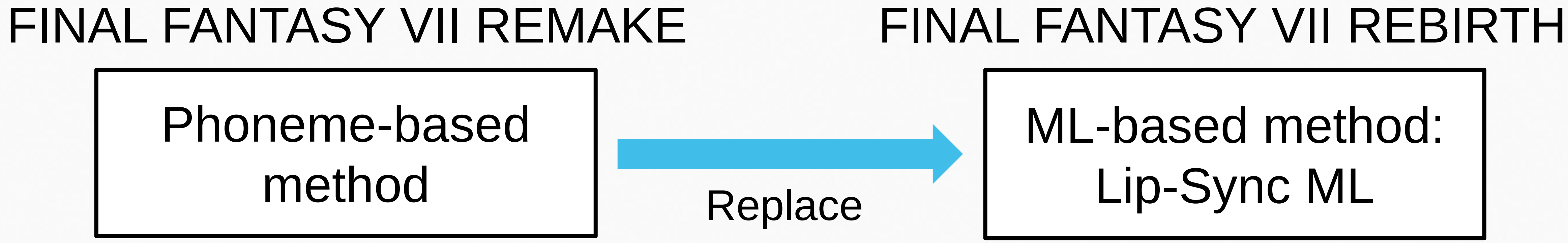
The second game of FINAL FANTASY VII remake trilogy

Previous title: FINAL FANTASY VII REMAKE (2020)

MISSION ON LIP-SYNC ANIMATION

- **Quality**
  - Express attractive characters and their dialogues.
  - Provide a game experience
    with the quality of our past CGI movie.
- **Efficiency**
  - Create a large amount of lip-sync animations
    only from audio.

OUR APPROACH IN FINAL FANTASY VII REBIRTH



FINAL FANTASY VII REMAKE

Phoneme-based method

Replace

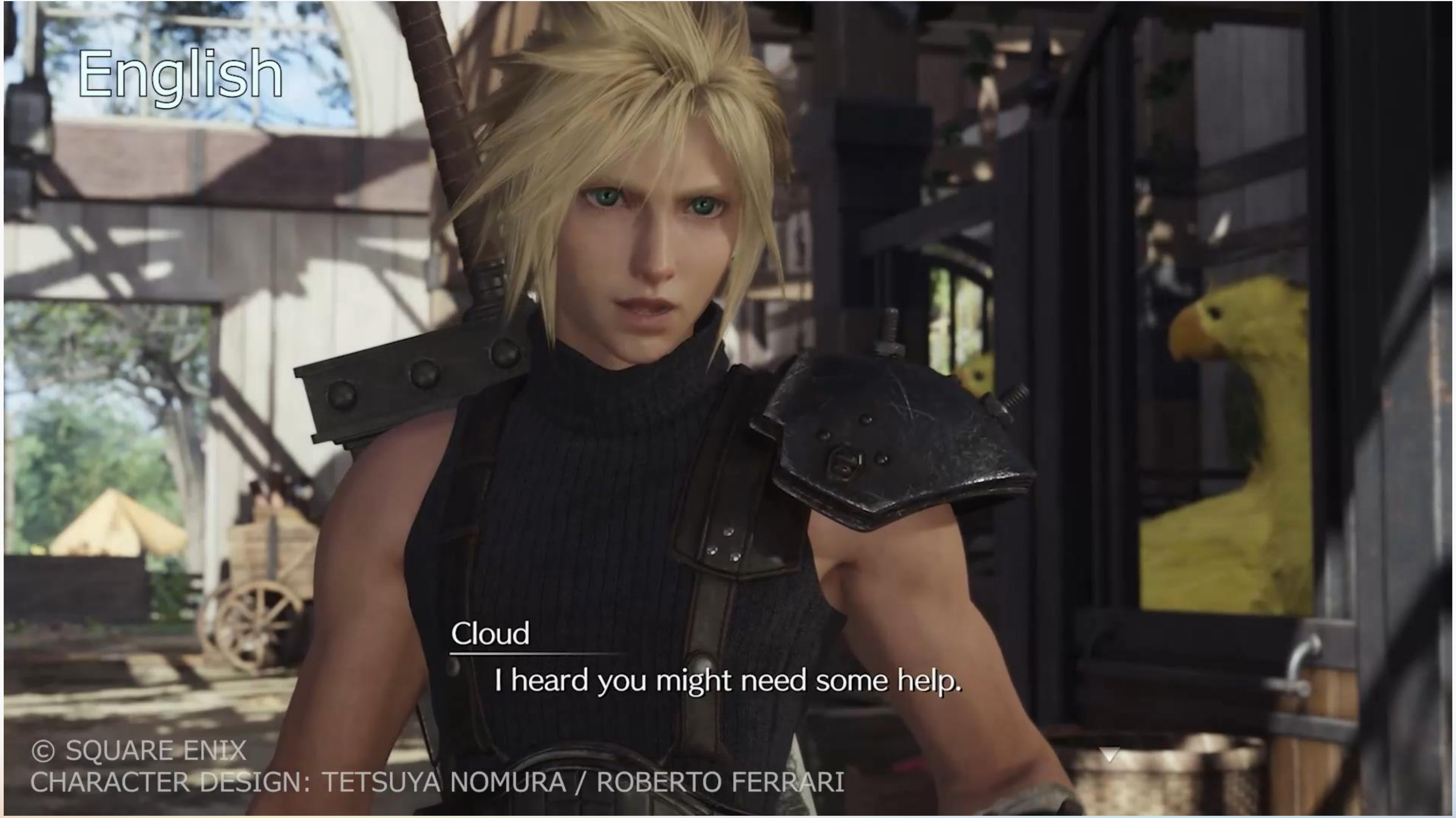
FINAL FANTASY VII REBIRTH

ML-based method: Lip-Sync ML

Use machine learning approach.

- Input: Audio
- Output: Lip-sync animations
- Training data: Cutscenes in FINAL FANTASY VII REMAKE

RESULTS IN THE GAME (ENGLISH)



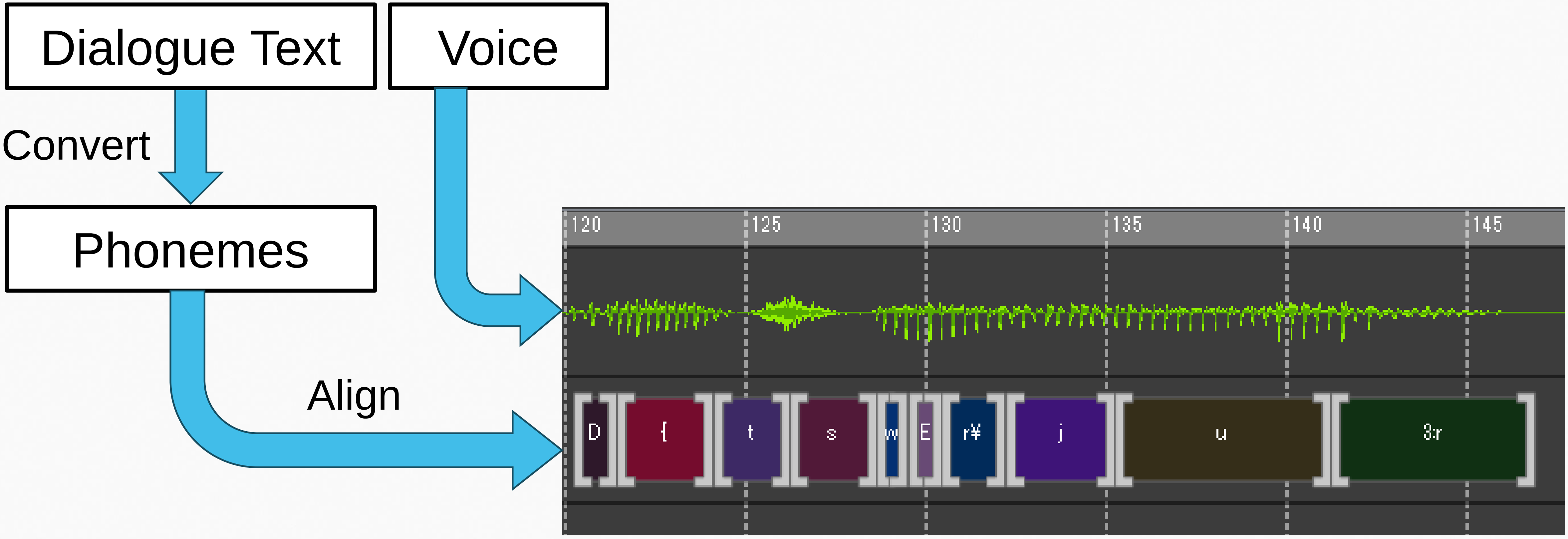
English

Cloud

I heard you might need some help.

© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI

PREVIOUS PHONEME-BASED METHOD



Dialogue Text

Voice

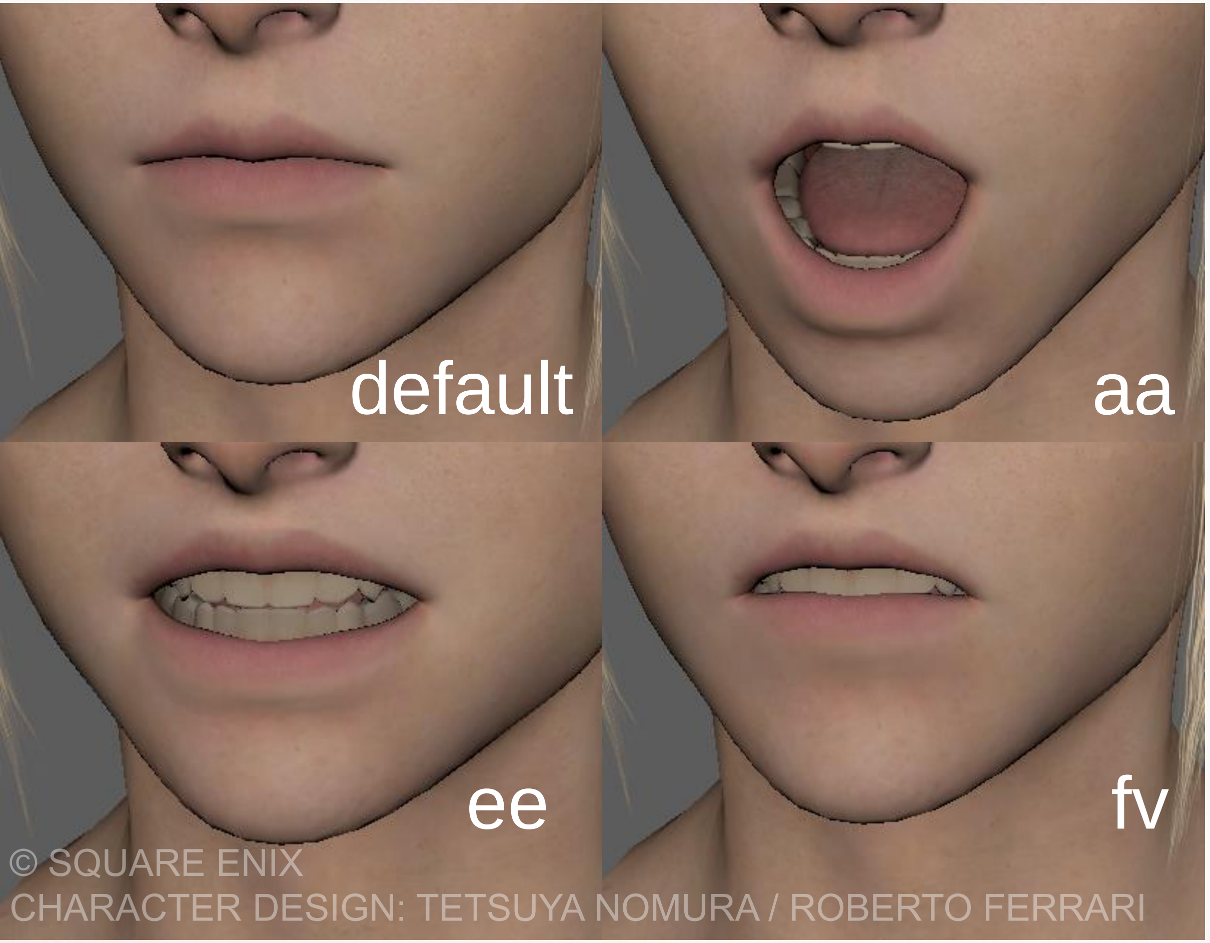
Convert

Phonemes

Align

Blend poses in the Lipmap corresponding to each phoneme.

LIPMAP



Lipmap is an asset
storing multiple mouth poses
for certain phonemes.

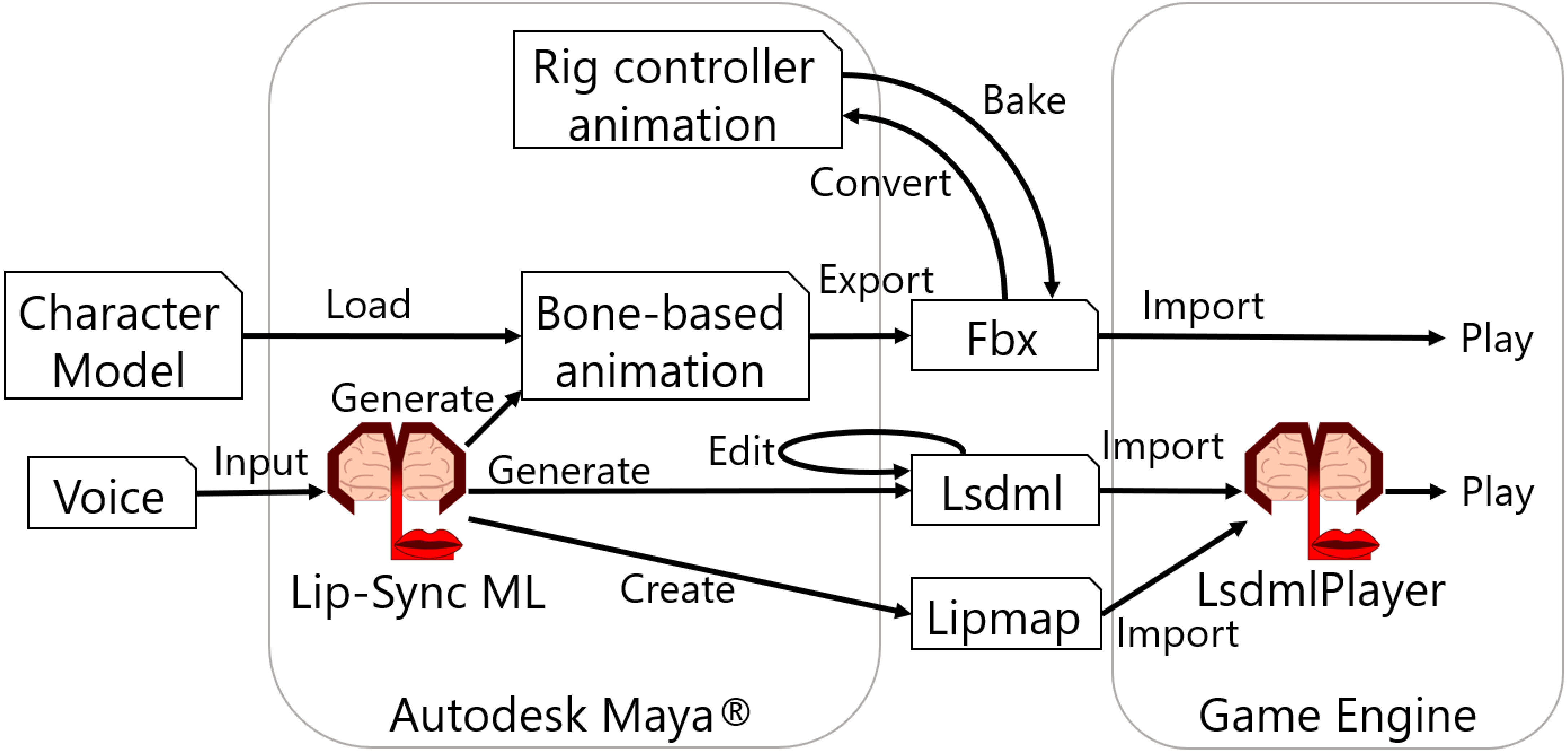
We choose 8 poses (incl. default)
for FINAL FANTASY VII REMAKE
and FINAL FANTASY VII REBIRTH.

LIP-SYNC ML

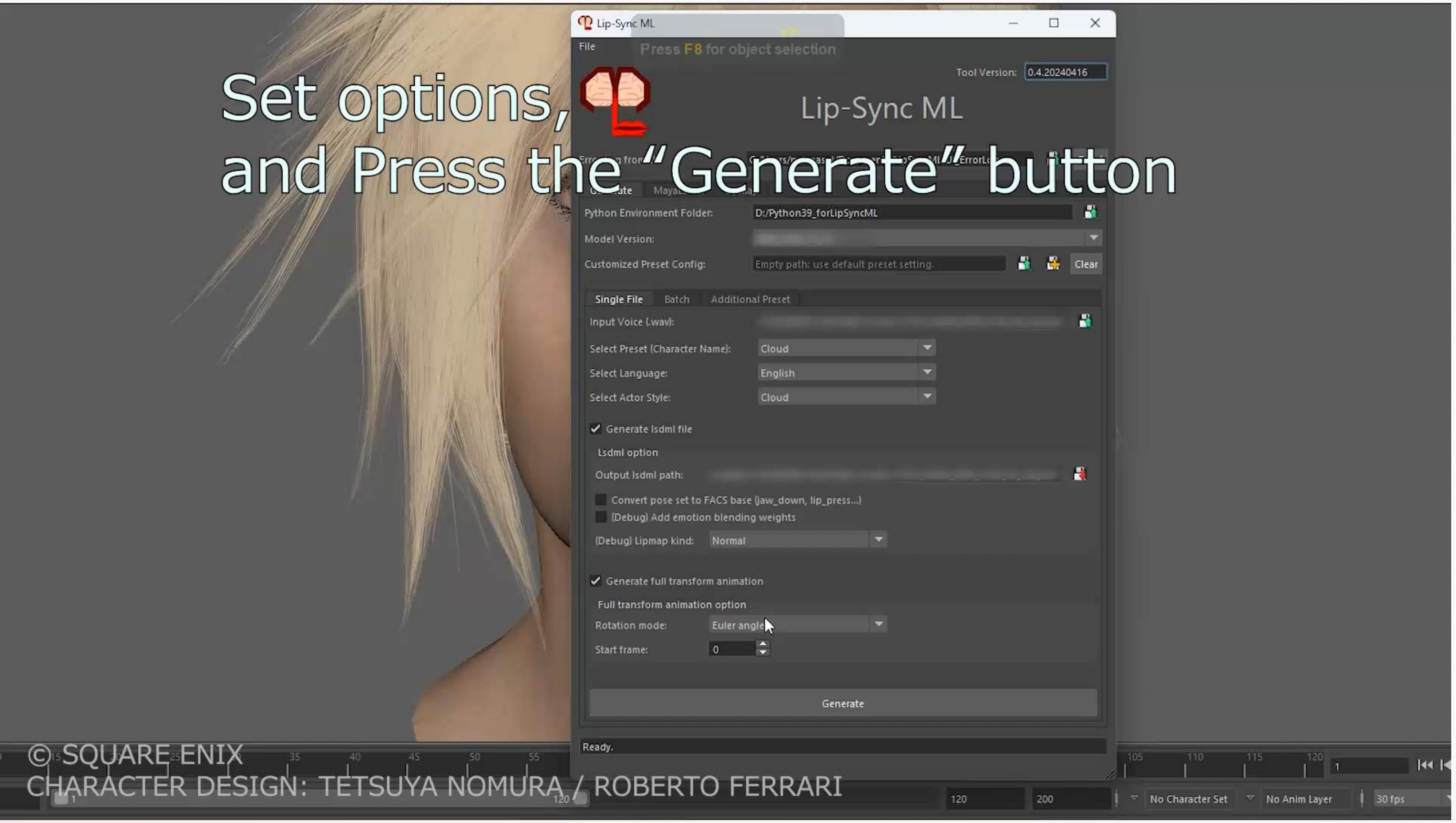
- Input: Audio
- Output: Lip-sync animations

	Bone-based animation	**Pose weight animation (Lsdml)**
Feature	Better quality	Easy to synthesized with animations of other facial behaviors
Usage	Base keyframes to create lip-sync animations for cutscenes	Used in simple event scenes, combat and field actions

WORKFLOW


Rig controller animation
Bake
Convert
Character Model
Load
Bone-based animation
Export
Fbx
Import
Play
Generate
Voice
Input
Edit
Generate
Lsdml
Import
Play
Lip-Sync ML
Create
Lipmap
Import
LsdmlPlayer
Autodesk Maya®
Game Engine

GENERATE ANIMATION



Set options,
and Press the "Generate" button

Lip-Sync ML

Tool Version: 0.4.20240416

Lip-Sync ML

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

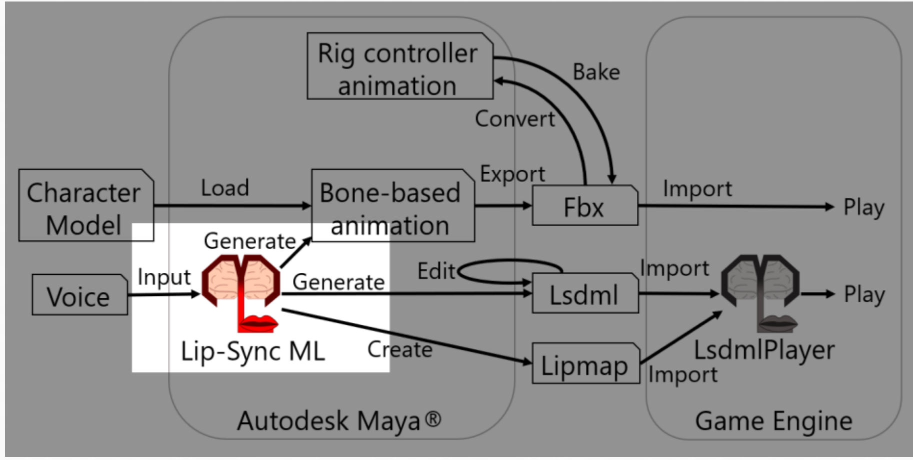
lsdml option
Output lsdml path:
□ Convert pose set to FACS base (jaw_down, lip_press...)
□ (Debug) Add emotion blending weights
(Debug) Lipmap kind: Normal

☑ Generate full transform animation

Full transform animation option
Rotation mode: Euler angle
Start frame: 0

Ready.

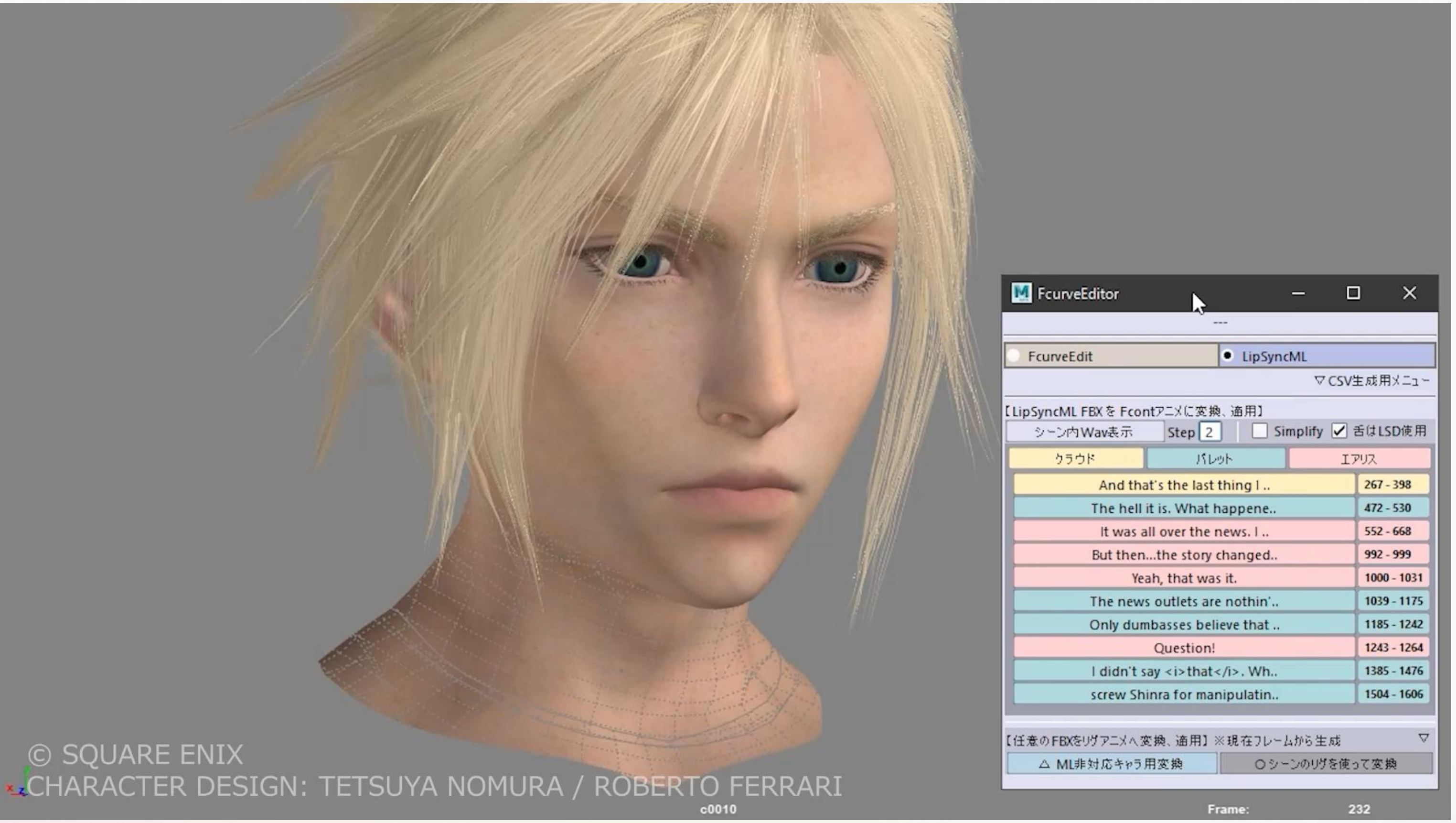
© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI



Perform machine learning generation process.

Support batch processing to deal with multiple voices.

CONVERT BONE ANIMATION INTO RIG CONTROLS



FcurveEditor

○ FcurveEdit ● LipSyncML

▽CSV生成用メニュー

【LipSyncML FBXを Fcontアニメに変換、適用】

シーン内Wav表示 Step 2 □ Simplify ☑ 舌はLSD使用

クラウド バレット エアリス

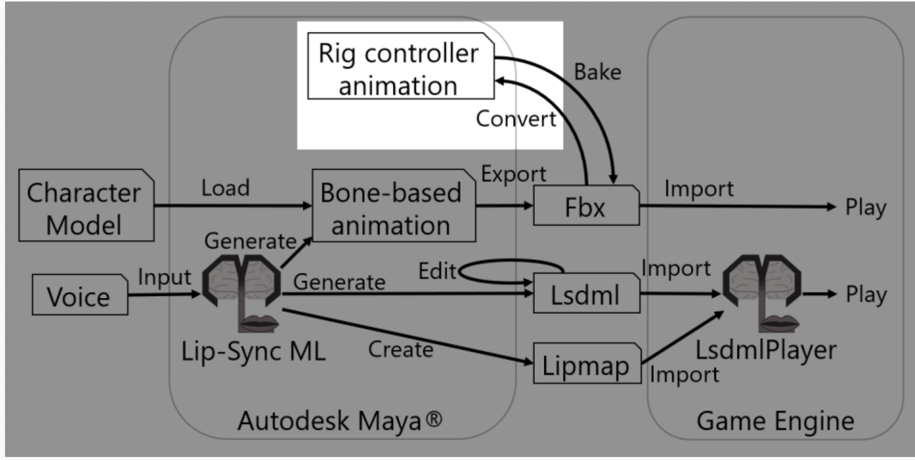
And that's the last thing I ..	267 - 398
The hell it is. What happene..	472 - 530
It was all over the news. I ..	552 - 668
But then...the story changed..	992 - 999
Yeah, that was it.	1000 - 1031
The news outlets are nothin'..	1039 - 1175
Only dumbasses believe that ..	1185 - 1242
Question!	1243 - 1264
I didn't say <i>that</i>. Wh..	1385 - 1476
screw Shinra for manipulatin..	1504 - 1606

【任意のFBXをリグアニメへ変換、適用】※現在フレームから生成 ▽

△ ML非対応キャラ用変換 ○シーンのリグを使って変換

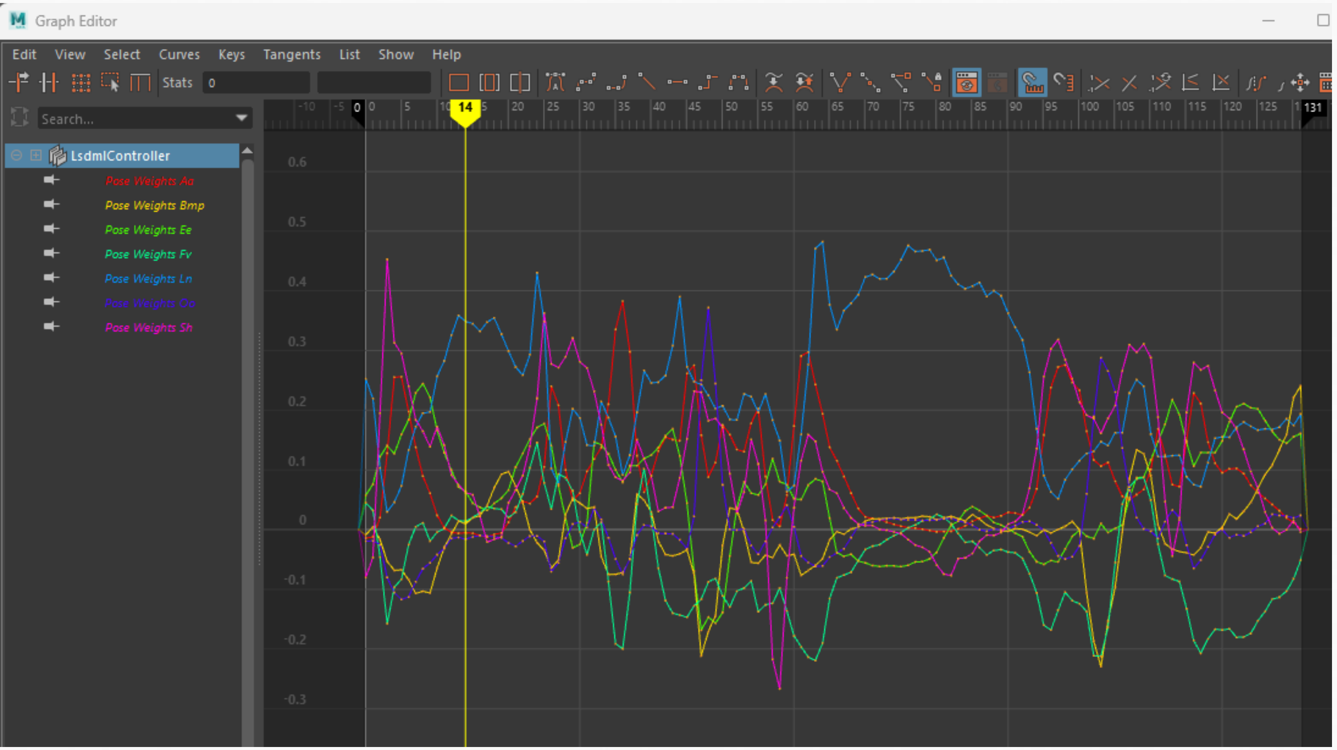
© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI

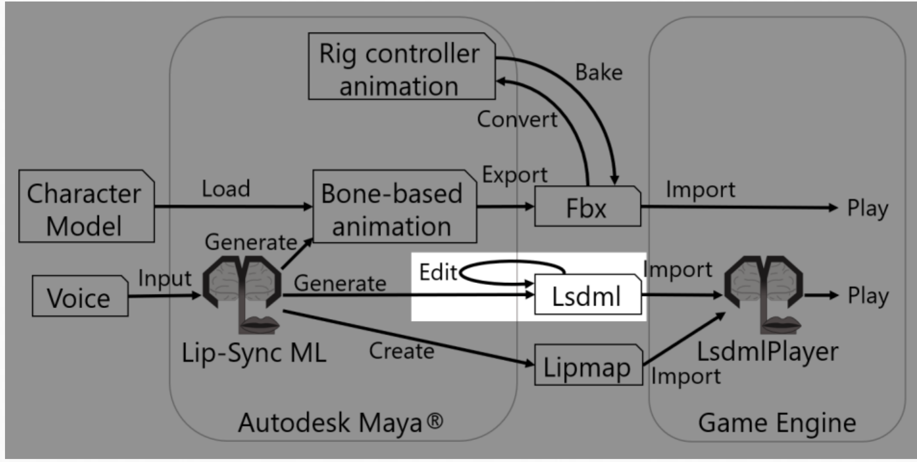
c0010 Frame: 232



Determine rig parameters in each animation frame based on affected bones.

EDIT LSDML





Edit Lsdml as animation curves in Maya if necessary.

In game engine,
LsdmlPlayer plays the animation.

COMPARISON WITH THE PREVIOUS METHOD

Lip-Sync ML
(Lsdml)
Previous
phoneme-based method
© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI

COMPARISON WITH THE PREVIOUS METHOD

This table is about animations with voices

for simple event scenes and action sequences.

	FINAL FANTASY VII REBIRTH	**FINAL FANTASY VII REMAKE**
Method	Lip-Sync ML (Lsdml)	Phoneme-based method
The number of audio clips	About 136,000	About 96,000
The number of clips corrected manually	**10 ~ 20**	**More than 100**

FUTURE WORK ON THE TOOL

Make setting up the tool easier.

- Users have to install necessary environment for machine learning currently.

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

LIP-SYNC ML: MACHINE LEARNING-BASED FRAMEWORK TO GENERATE LIP-SYNC ANIMATIONS IN FINAL FANTASY VII REBIRTH

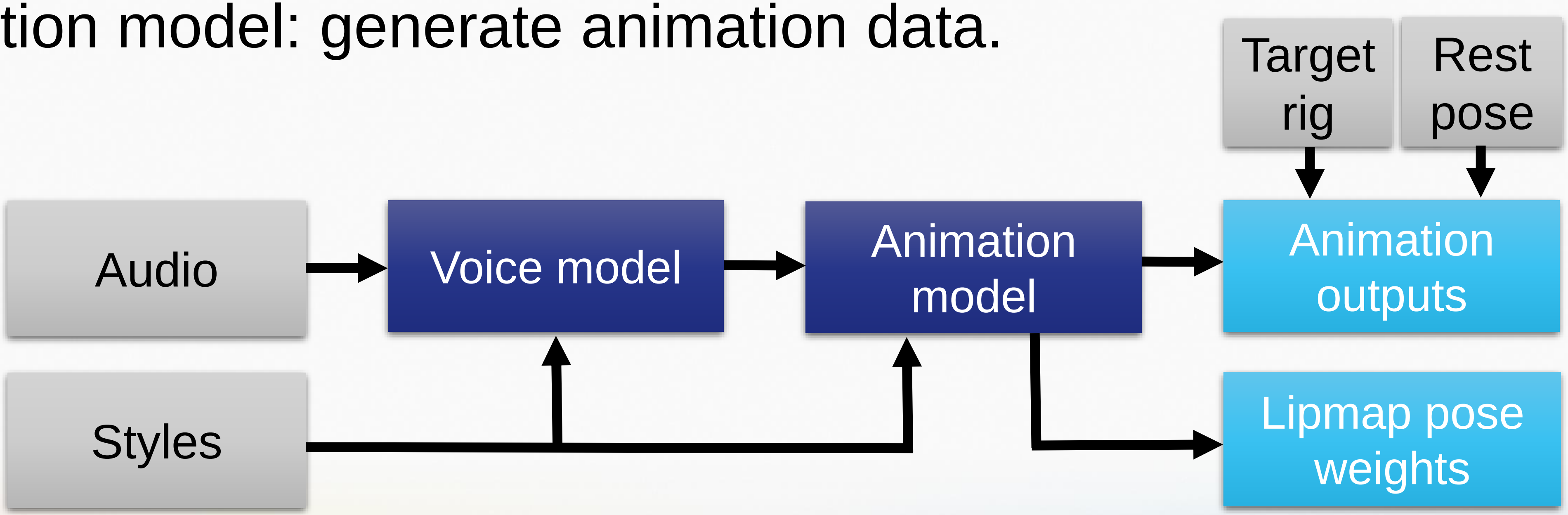
MACHINE LEARNING DETAILS

TRAINING DATA

- About 3.5 hours of cutscene data from FINAL FANTASY VII REMAKE.
  - 53 different characters.
  - 3 facial skeletons: main characters, mob characters, Red XIII.
  - 2 languages: Japanese and English.
- Bone transform animation data (was easier to collect).
- Data augmentation: random speed and pitch changes.

MODEL ARCHITECTURE

- Designed as 2 sub-models, trained end-to-end.
- Voice model: transform audio into voice features.
- Animation model: generate animation data.



LEARNED STYLES

- Multiple independent styles: language, actor, rig.
- Each style has a set of exclusive values (Actor: Cloud, Aerith, Tifa...).
- Styles can also be left empty.

Language	Japanese	English	
Actor	Cloud	Aerith	Tifa
Rig	Main Chara	Mob Chara	RedXIII

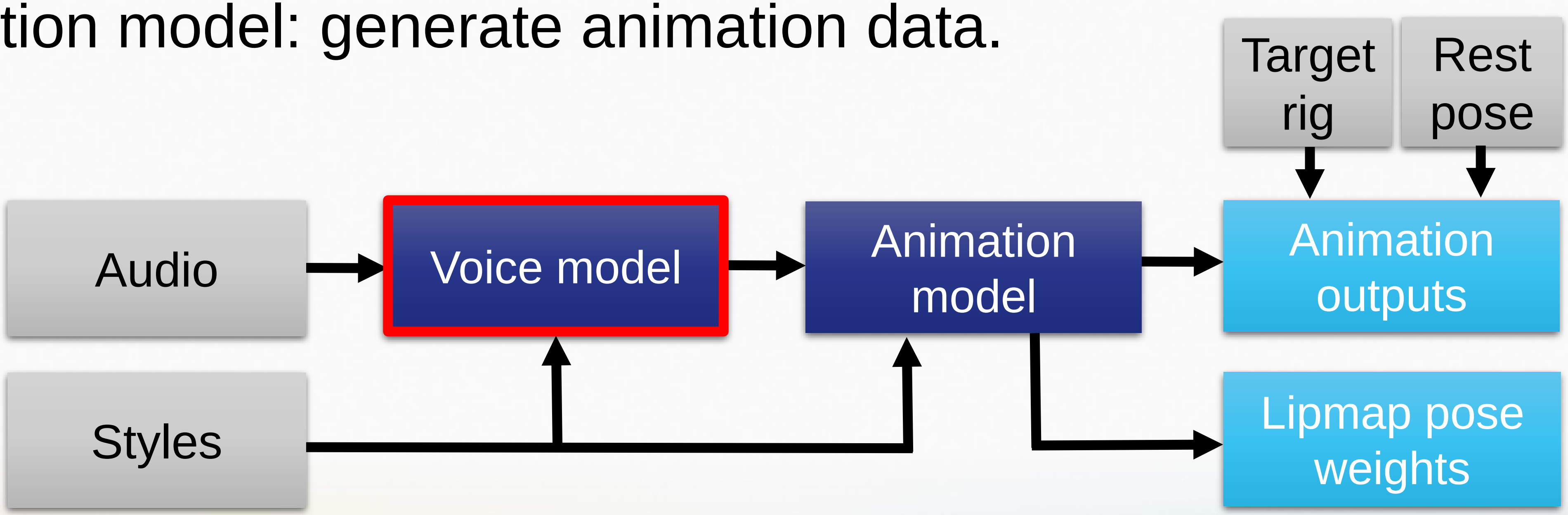
Language=None

Actor=Cloud

Rig=MainChara

MODEL ARCHITECTURE

- Designed as 2 sub-models, trained end-to-end.
- Voice model: transform audio into voice features.
- Animation model: generate animation data.

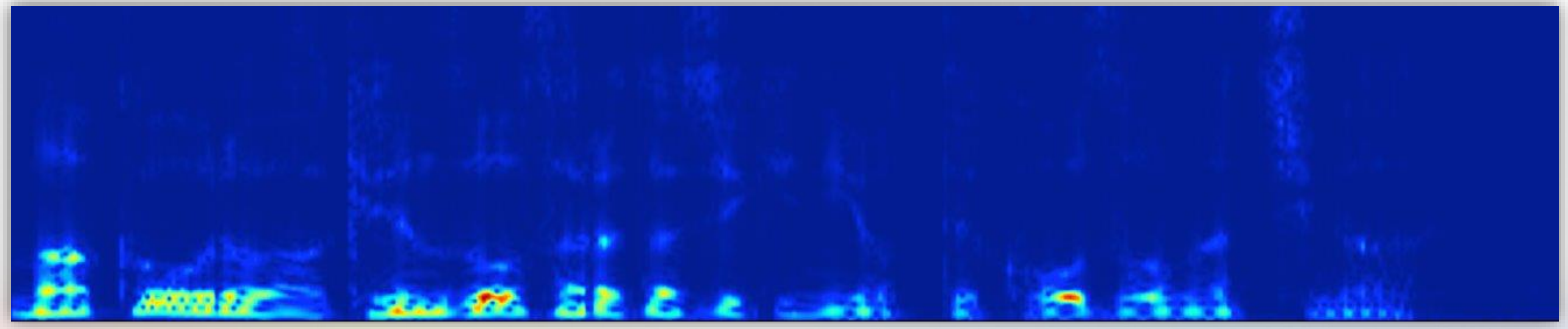


AUDIO PROCESSING

- Convert audio to mono.
- Resample to 19200 Hz (makes converting to 30 and 60 fps easier).

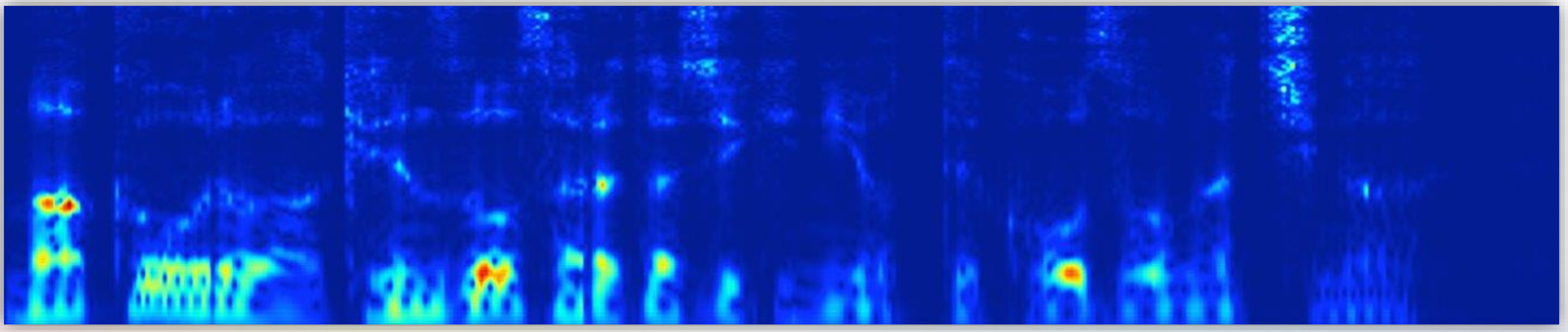
SPECTROGRAM GENERATION

- Sample rate: 19200 Hz, window size: 200, stride: 160.
- Produces 120 spectrogram frames per second.
- Frequency range: 80 to 8000 Hz.



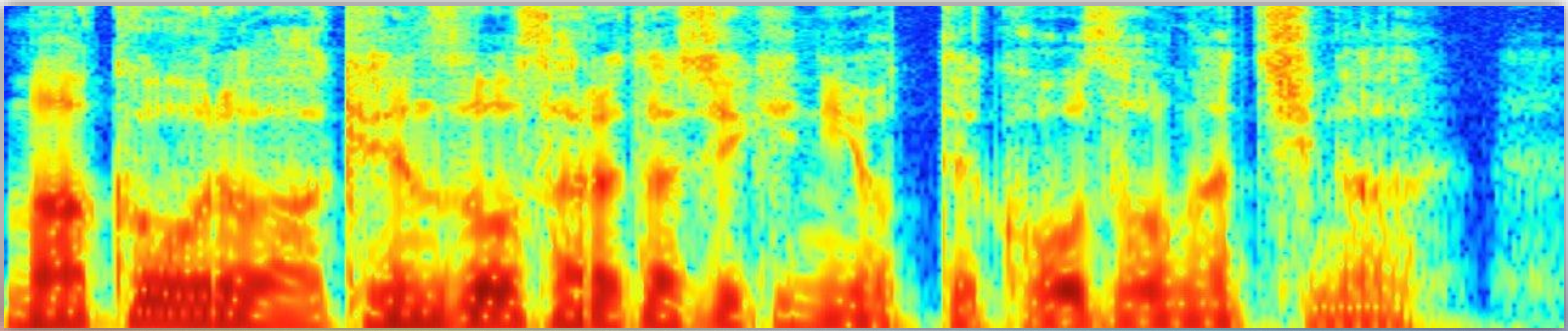
MEL SCALE CONVERSION

- Convert vertical axis to mel scale (logarithmic) using 256 bins.
- Closer to human perception of pitch.
- Changes in pitch become closer to vertical translations.



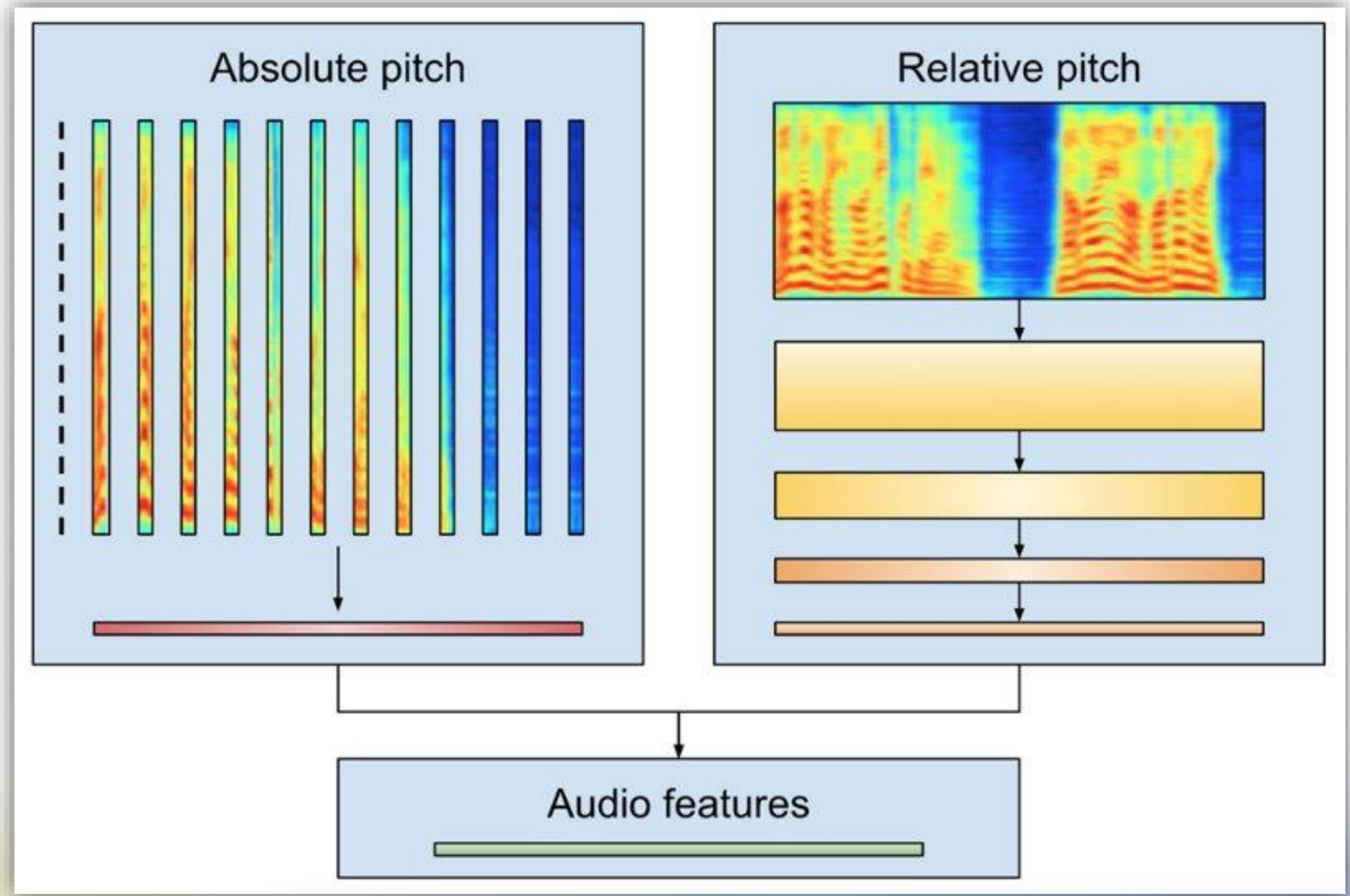
LOG MEL SPECTROGRAM

- Apply logarithm to the values (amplitude).
- Closer to human perception of volume.
- Can be seen as an image of 120 frames per second (time) x 256 mel bins (pitch), with a single value per pixel (volume).



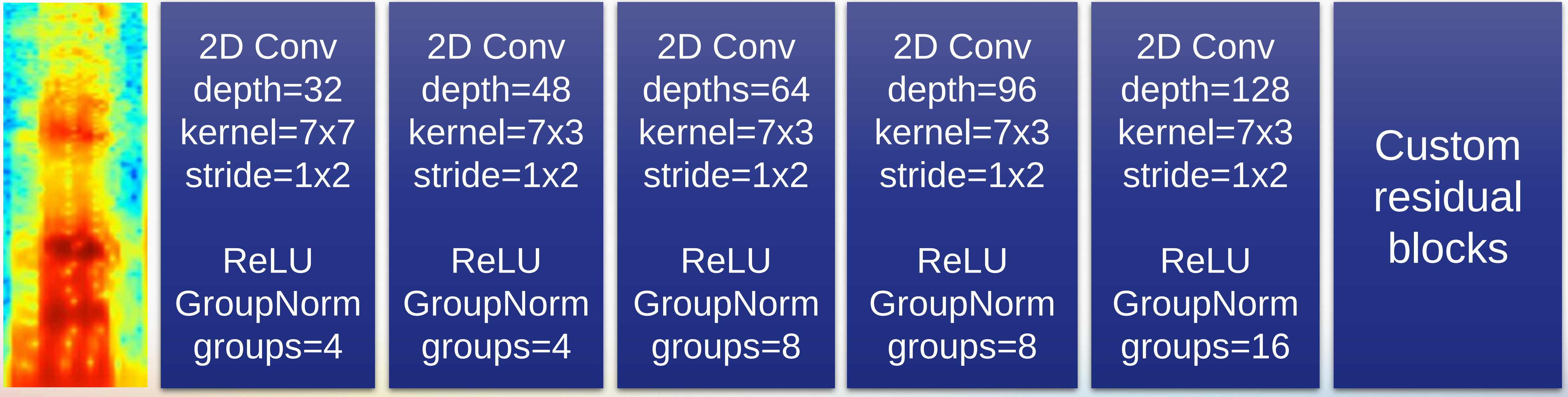
AUDIO FEATURES

- **Absolute pitch**: use each column directly as a feature vector (bag of features).
- **Relative pitch**: process as a 2D image (pitch invariance).



RELATIVE PITCH PROCESSING

- Transfer information from height into feature depth.
- Progressively apply 2D convolutional networks with a stride of 2 in the vertical axis to halve height, while increasing feature depth.



2D Conv
depth=32
kernel=7x7
stride=1x2

ReLU
GroupNorm
groups=4

2D Conv
depth=48
kernel=7x3
stride=1x2

ReLU
GroupNorm
groups=4

2D Conv
depths=64
kernel=7x3
stride=1x2

ReLU
GroupNorm
groups=8

2D Conv
depth=96
kernel=7x3
stride=1x2

ReLU
GroupNorm
groups=8

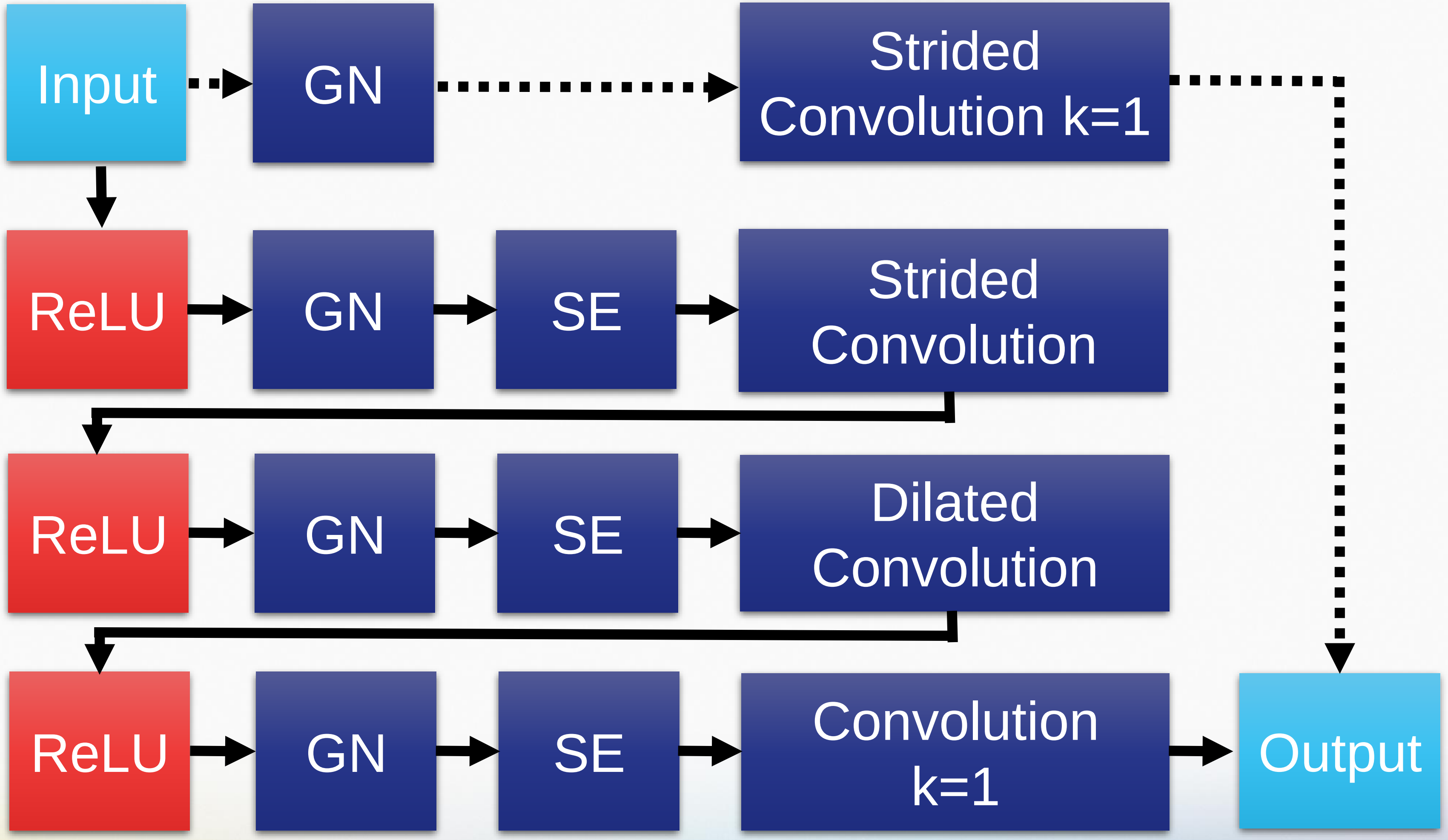
2D Conv
depth=128
kernel=7x3
stride=1x2

ReLU
GroupNorm
groups=16

Custom residual blocks

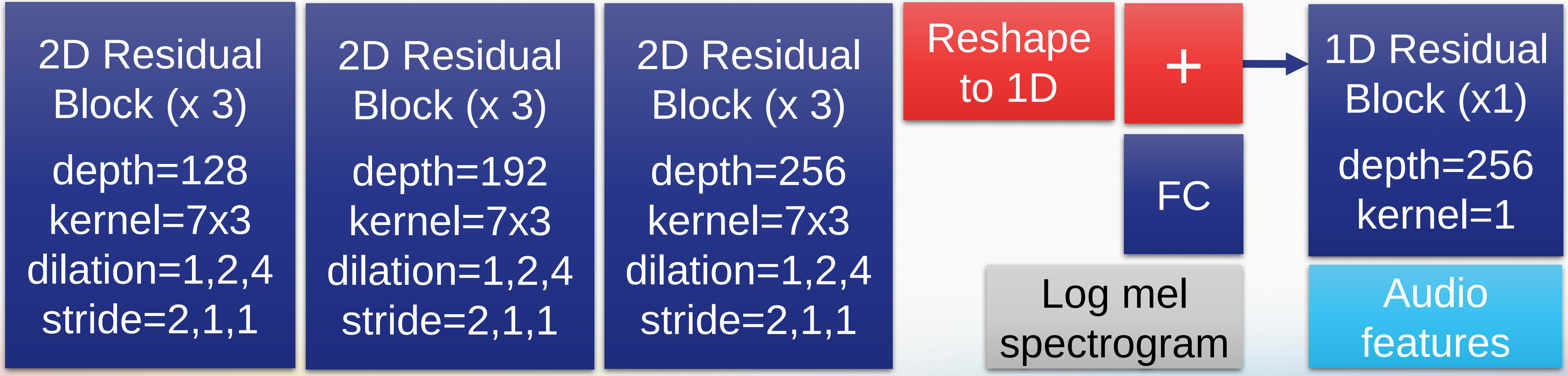
CUSTOM RESIDUAL BLOCKS

- Extension of ResNet block [1].
- **GN**: Group Normalization [2].
- **SE**: Squeeze and Excitation [3].
- **Dotted**: ResNet shortcut connection.



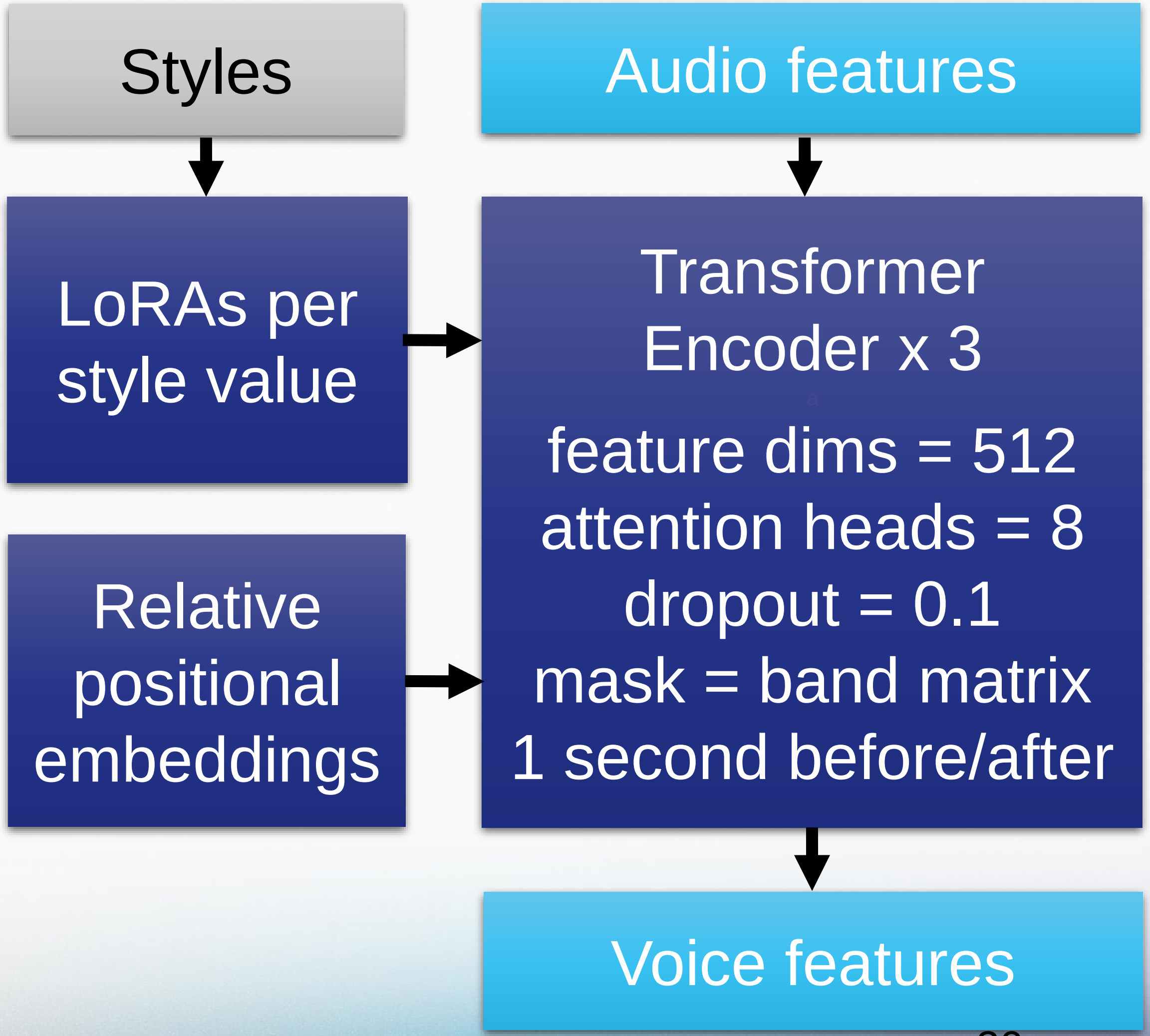
COMPUTING AUDIO FEATURES

- 3 groups of 3 consecutive residual blocks with different time dilations.
- In each group: 1st block halves height, 2nd and 3rd apply time dilation.
- Absolute pitch features are combined using one more residual block.

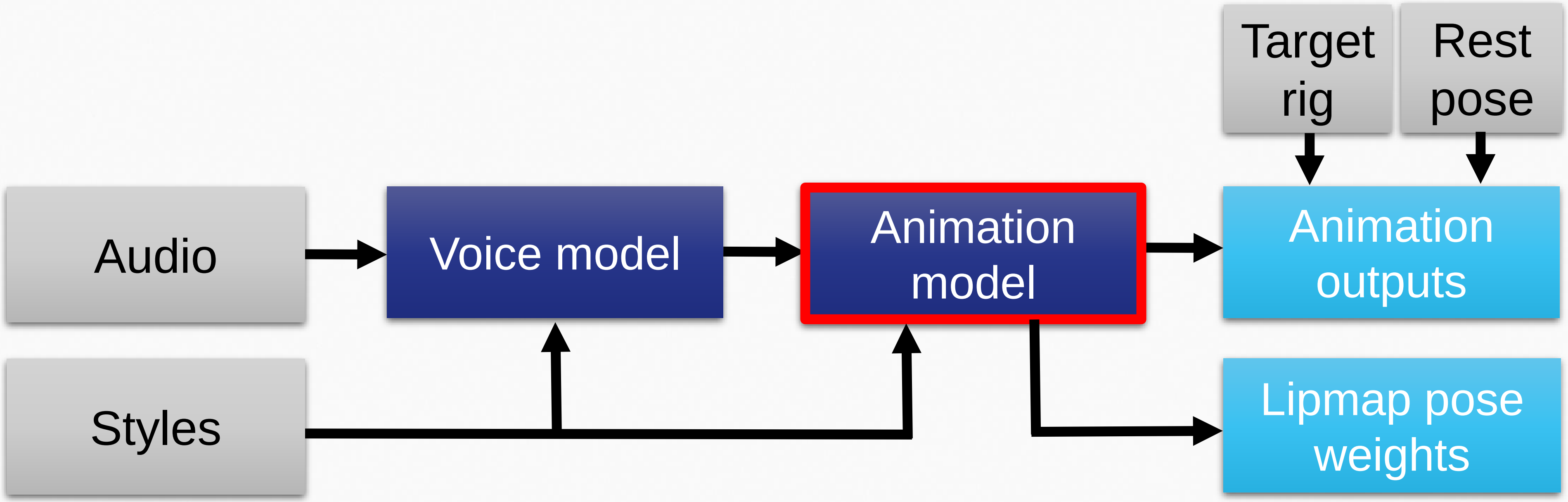


VOICE FEATURES

- 3x Transformer Encoder blocks [4].
  - Attend 1 second before / after.
  - Relative positional embeddings [5].
- Learn styles using LoRAs [6].
  - LoRAs for each style value.
  - Rank 8, added to qkv matrix.
  - Shared by all transformer blocks.

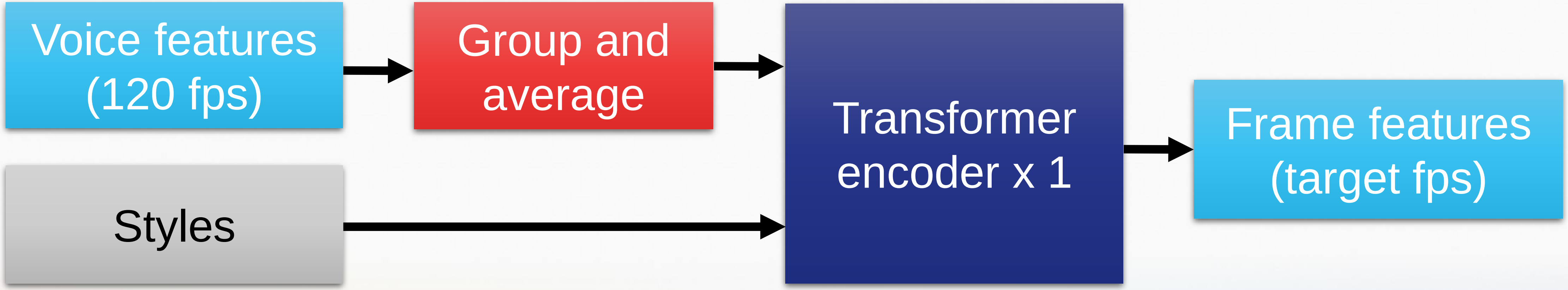


ANIMATION MODEL



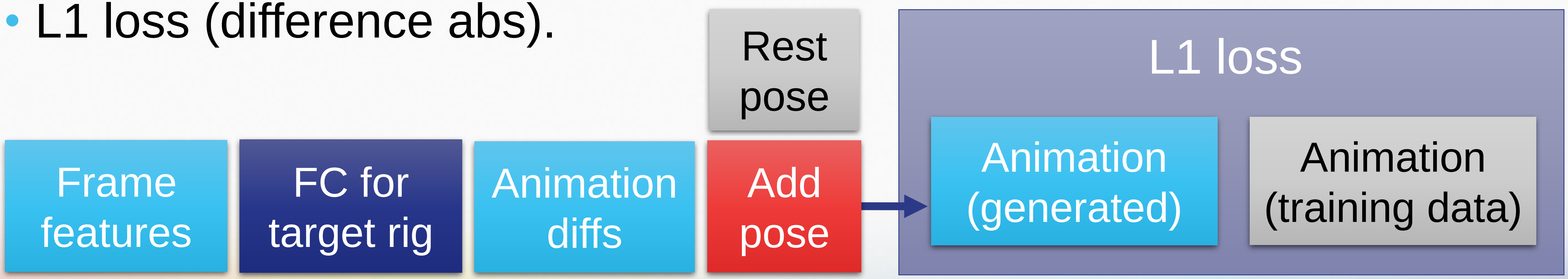
FRAME FEATURES

- Voice features use 120 fps.
- Group and average voice features to 30 or 60 fps as needed.
- Apply one more transformer encoder block as before.



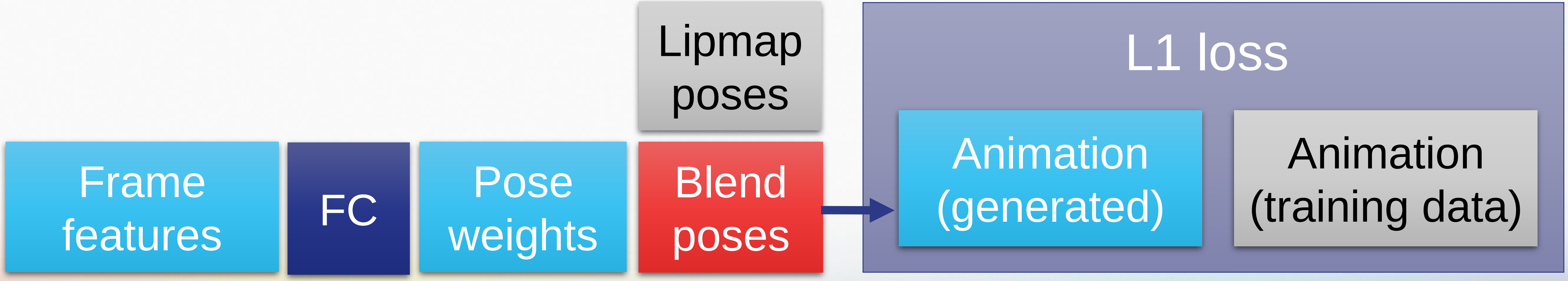
ANIMATION OUTPUTS

- Transform frame features into per-bone animation transform diffs.
- Format is the same as in training data (could also be rig parameters).
- Separate Fully Connected layers for each output rig / facial skeleton.
- Add character rest pose. Per frame outputs, no keyframes.
- L1 loss (difference abs).



LIPMAP POSE WEIGHTS

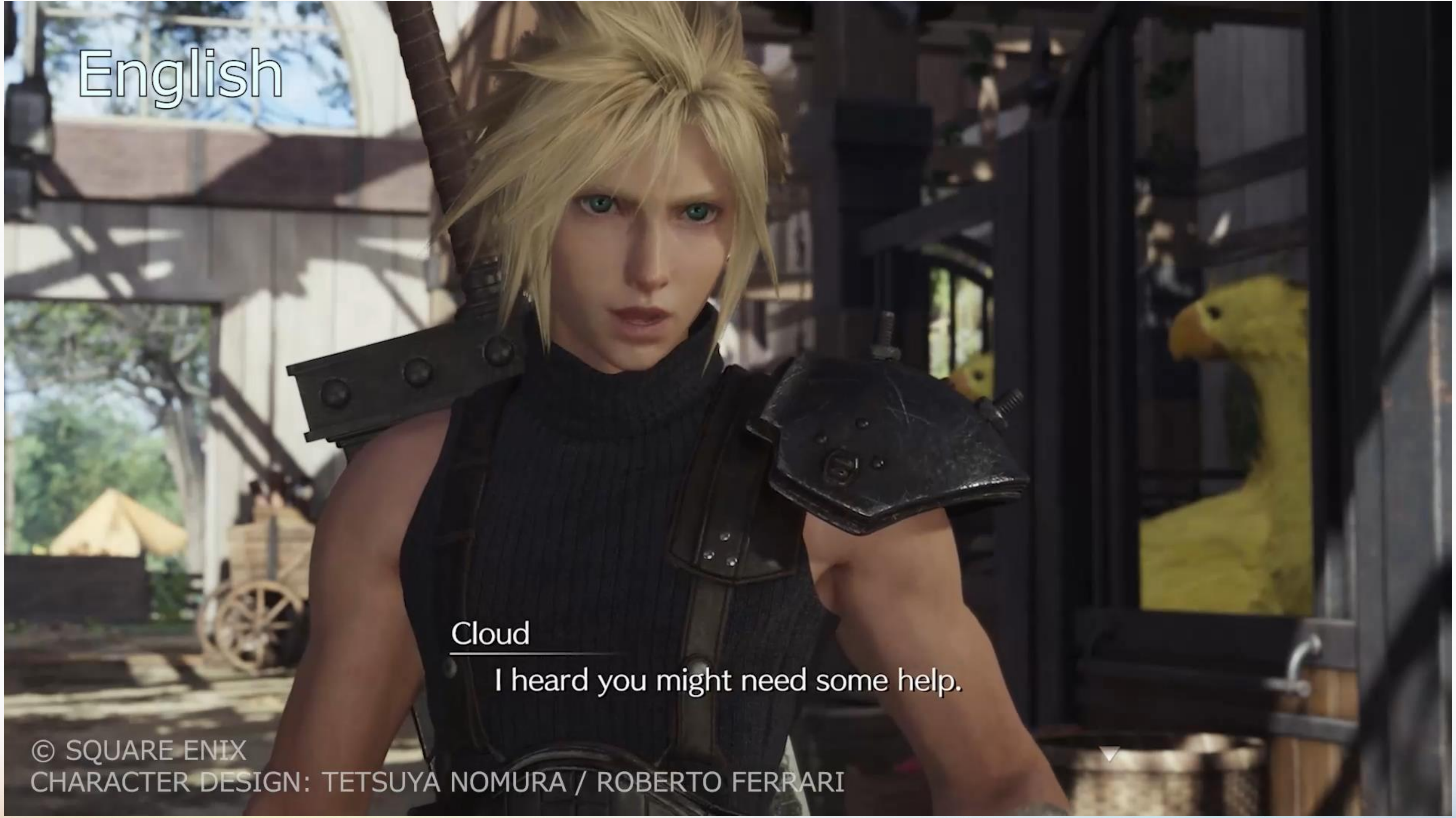
- Transform frame features into lipmap pose weights (rig independent).
- Lipmap poses are fixed, pose weights are generated.
- Blending poses is a differentiable operation, can backpropagate.
- Same loss as animation outputs.



ADDITIONAL LOSS IMPROVEMENTS

- Compensate how numerically relevant errors are for each value.
  - For example, a 0.5 error in a translation is likely small, but not in a scale factor.
  - Precompute the min-max ranges of each value in training data.
  - Rescale errors to make them relative to the sizes of their min-max ranges.
- Improve closing the mouth in generated animations.
  - Exploit the fact that rest poses have their mouths closed.
  - Scale up errors the closer training data values are to their rest pose.

RESULTS IN THE GAME (MULTIPLE LANGUAGES)



Cloud
I heard you might need some help.

© SQUARE ENIX
CHARACTER DESIGN: TETSUYA NOMURA / ROBERTO FERRARI

CONCLUSION

In FINAL FANTASY VII REBIRTH,

our machine learning-based framework Lip-Sync ML

allowed animators to efficiently create high-quality lip-sync animations.

This enabled characters to feel more natural when speaking,

making dialogue scenes feel more immersive.

Thank you for your attention.

Masato Nakada
nakmasat@square-enix.com

Leandro Graciá Gil
graclean@square-enix.com

For slides

(http://www.jp.square-enix.com/tech/publications.html)

REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016.
  Identity Mappings in Deep Residual Networks. https://arxiv.org/pdf/1603.05027
- [2] Yuxin Wu and Kaiming He. 2018.
  Group Normalization. https://arxiv.org/pdf/1803.08494
- [3] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2019.
  Squeeze-And-Excitation Networks. https://arxiv.org/pdf/1709.01507
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,
  Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2023.
  Attention Is All You Need. https://arxiv.org/pdf/1706.03762
- [5] 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: Generating Music With Long-Term Structure. https://arxiv.org/pdf/1809.04281
- [6] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021.
  LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/pdf/2106.09685