JALI-Driven Expressive Facial Animation and Multilingual Speech in Cyberpunk 2077

Pif Edwards
JALI Research

Chris Landreth
JALI Research

Mateusz Popławski
CD PROJEKT RED

Robert Malinowski
CD PROJEKT RED

Sarah Watling
JALI Research

Eugene Fiume
JALI Research

Karan Singh
JALI Research

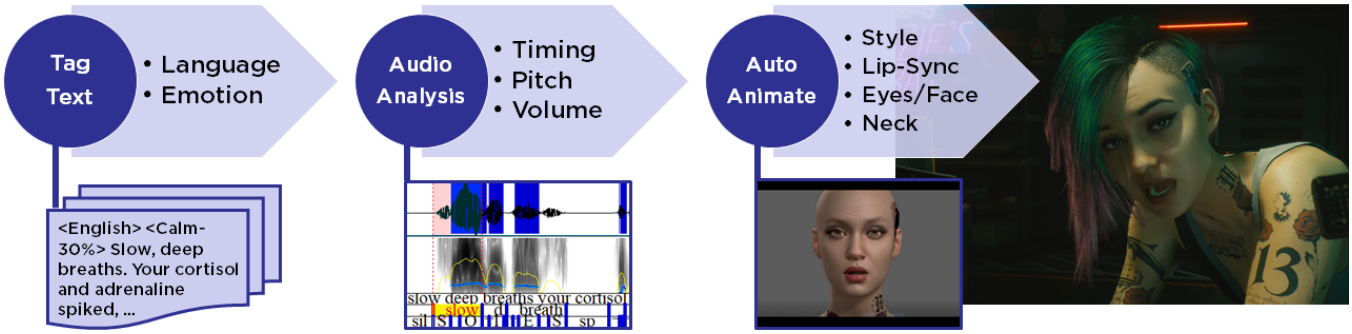

Figure 1: Cyberpunk 2077 workflow: An audio performance and tagged animation transcript is combined with hand-crafted animation to produce expressive, multi-lingual speech at an unprecedented scale.

ABSTRACT

Cyberpunk 2077 is a highly anticipated massive open-world video game, with a complex, branching narrative. This talk details new research and innovative workflow contributions, developed by JALI, toward the generation of an unprecedented number of hours of realistic, expressive speech animation in ten languages, often with multiple languages interleaved within individual sentences. The speech animation workflow is largely automatic but remains under animator control, using a combination of audio and tagged text transcripts. We use insights from anatomy, perception, and the psycho-linguistic literature to develop independent and combined language models that drive procedural animation of the mouth and paralingual (speech supportive non-verbal expression) motion of the neck, brows and eyes. Directorial tags in the speech transcript further enable the integration of performance capture driven facial emotion. The entire workflow is animator-centric, allowing efficient key-frame customization and editing of the resulting facial animation on any typical FACS-like face rig. The talk will focus equally on technical contributions and its integration and creative use within the animation pipeline of the highly anticipated AAA game title: Cyberpunk 2077.

SIGGRAPH '20 Talks, August 17, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7971-7/20/08.
https://doi.org/10.1145/3388767.3407339

ACM Reference Format:
Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Malinowski, Sarah Watling, Eugene Fiume, and Karan Singh. 2020. JALI-Driven Expressive Facial Animation and Multilingual Speech in Cyberpunk 2077. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH '20 Talks), August 17, 2020.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3388767.3407339

1 INTRODUCTION AND RELATED WORK

Animating an engaging anthropomorphic face is the most challenging aspect of character animation. Humans are highly adapted to recognize and understand faces. Imperfect emulation of the subtlest facial nuance on an animated character can alarmingly drop the character into the *Uncanny Valley*, an emotional abyss where the audience loses trust and empathy with the character.

Facial animation today is produced in one of three ways.

- *Professionally key-framed animation* is high quality but laborious, often hindered by a language barrier for animators in a multi-lingual setting.
- *Performance Capture* is high quality but not editable (use or re-capture), challenging for occluded parts like the tongue, and limited by access to professional capture setups.
- *Procedural models* are often plagued by a cartoony one-to-one phoneme-viseme mapping, and poor co-articulation and paralinguals.

High-end film and game studios have traditionally relied on their most talented face actors and polyglot animators for this critical and labor-intensive task, neither of which have scaled beyond a few hundred lines of spoken content. In contrast, Cyberpunk 2077's ambitious vision of a complex narrative, driven by tens of thousands of

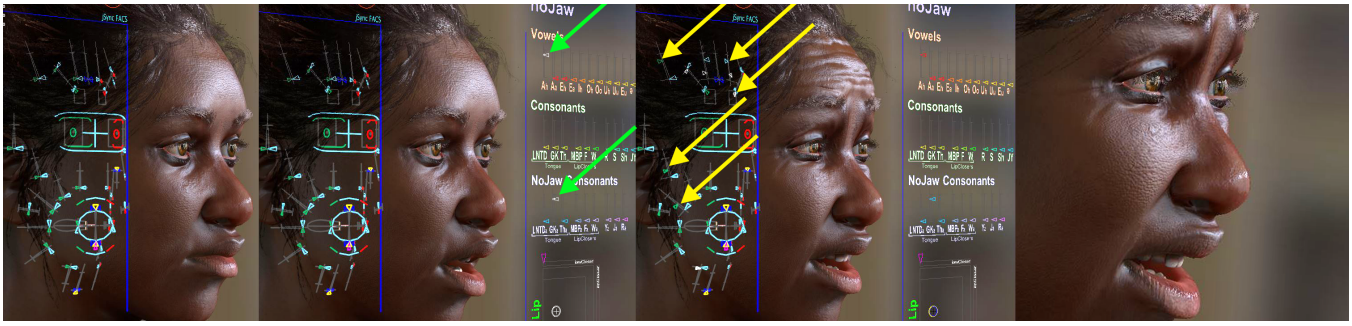

Figure 2: JALI language models drive FACS-like rigs (a); to produce sparse and compact speech animation curves (green arrows) (b); augmented with paralingual eye, brow, neck motion (yellow arrows) (c); and keyframe or performance captured emotion (d).

lines of character speech in multiple languages, required a complete re-design of a typical AAA game's facial animation workflow.

We judiciously combine aspects of recent research on procedural animation for lip–synchronization [Edwards et al. 2016], with new research on modeling paralinguals, and integration with character emotion and idle behaviour that is handcrafted or gleaned from performance capture. Animators use speech transcript tags to annotate and control the timing and intensity of facial emotion, non-verbal sounds, and language, during a vocal performance (see Figure 1). Tags can also be used to direct the emotional stance of the character to augment the expressive signal that is extracted from the audio soundtrack. Our workflow generalizes beyond Cyberpunk 2077 to multiple applications including film, games, and real-time text-to-speech.

2 LANGUAGE MODELS

The dialogue for the entire gameplay in Cyberpunk 2077 has been reproduced in ten languages: English, Spanish, French, Polish, Russian, Italian, Brazilian Portuguese, Mandarin, Japanese and German. We have developed new acoustic and Grapheme-to-Phoneme models by training 50 to 400 hours of transcribed speech in these languages. The acoustic models were trained using tools that leverage Kaldi [Povey et al. 2011]. Our language models adjust to specific speakers, and our workflow is able to handle mixed-language vernacular.

3 SPEECH AND PARALINGUALS

An audio-aligned transcript is used to generate phonetic timings that produce sparse, compact, easy-to-edit animation curves (see Figure 2). Language-specific phonemes and co-articulation enable mixed language speech in a range of speech styles inferred from the audio signal. We further use insights from the perceptual and psycholinguistic literature to define correlations between speech and paralingual animation including blinks, saccades and brow motion. For example, blinks are based on several influences: audio analysis of voice recording, grammatical clues in the transcript, and length of time since the last blink. These features are used to determine a probablistic motivation for blinks (or their absence) as being expressive or communicative: for facial maintenance such as dryness of eyes; or cognitive and indicative of internal thought. The nature of the blink determines its duration and timing [Parke and Waters 2008]. The animation of brows similarly relates to emotional cues in pitch, volume, and higher-order speech inflection [Ekman 2004].

4 EMOTION AND IDLE INTEGRATION

An emotional and idling repertoire of a character is represented as a collection of facial motion clips of tagged intensity. We modulate and integrate these clips with speech and paralingual animation using transcript tags and phonetic timing. Our approach ensures that the integrated motion does not conflict with the precise facial poses needed to articulate some phonemes. For example, overlaying a smile does not override a pucker to pronounce the phoneme "oo". Likewise, the integration and interaction of facial performance capture with JALI preserves correct jaw/lip anatomy.

5 CONCLUSION AND FUTURE WORK

The accompanying video is an early glimpse into the game. Detailed character footage and spoken narrative from the game will be revealed at the SIGGRAPH 2020 talk, coincident to the release of the game. Our current implementation does have limitations that we will also present in the talk such as working with atypical voices, exaggerated accents, and operatic singing. We believe Cyberpunk 2077 will set a new benchmark for large-scale, multi-lingual, expressive character speech in video games.

REFERENCES

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 (July 2016), 127:1–127:11.

Paul Ekman. 2004. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation*. Springer, 39–50.

Frederic I Parke and Keith Waters. 2008. Facial Animation. In *Computer Facial Animation*. A K Peters Ltd, 1–277.

Daniel Povey et al. 2011. The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Dec. 2011).