Content and Style Aware Audio-Driven Facial Animation

Qingju Liu[1], Hyeongwoo Kim[2], Gaurav Bharaj[1]
[1]Flawless AI, UK [2]Imperial College London, UK

Motivation: An Audio-Driven Facial Animation (ADFA) method which enables content and style reenactment.

Challenges: To model the diverse styles disentangled from content with limited data.

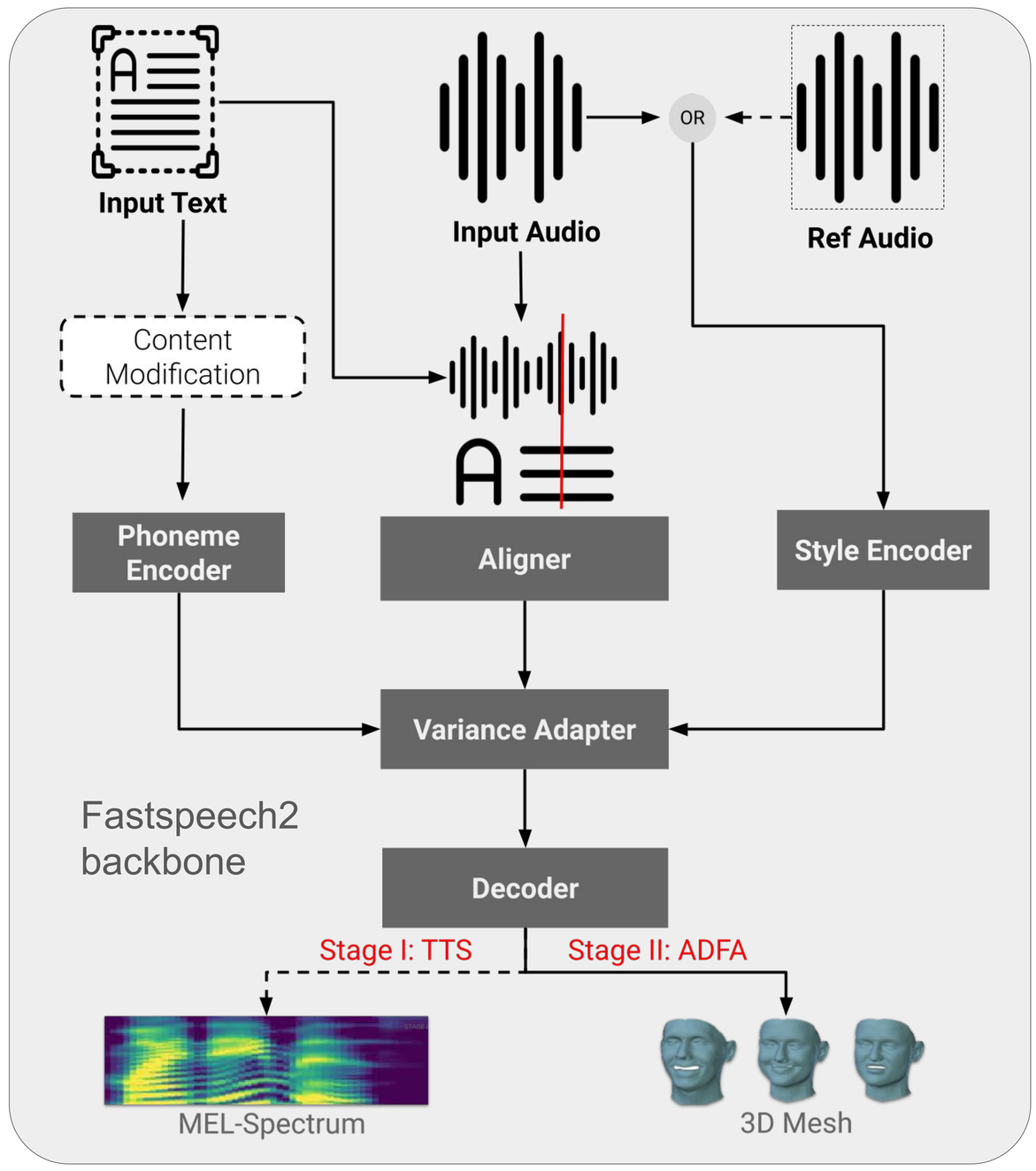
Novelties:

Knowledge transfer from TTS (speech synthesis) to ADFA with two-stage training.

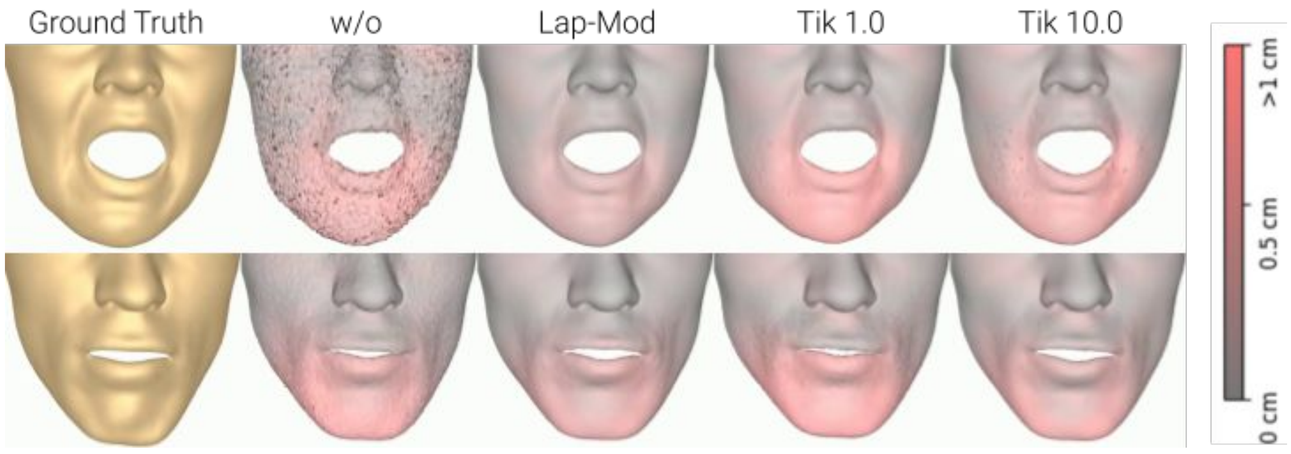
Styles evolve from audio styles (how does it sound) to visual style (how do articulations look like).

Modified Laplacian loss to reduce computational complexity.

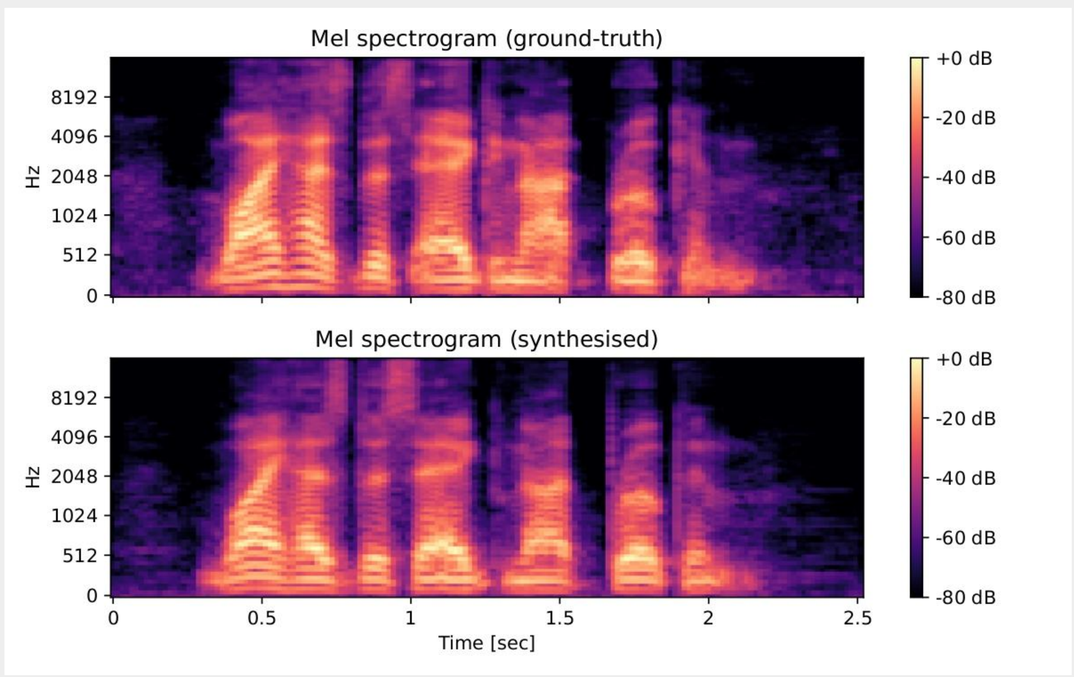
Allowing both audio style transfer and content modification.



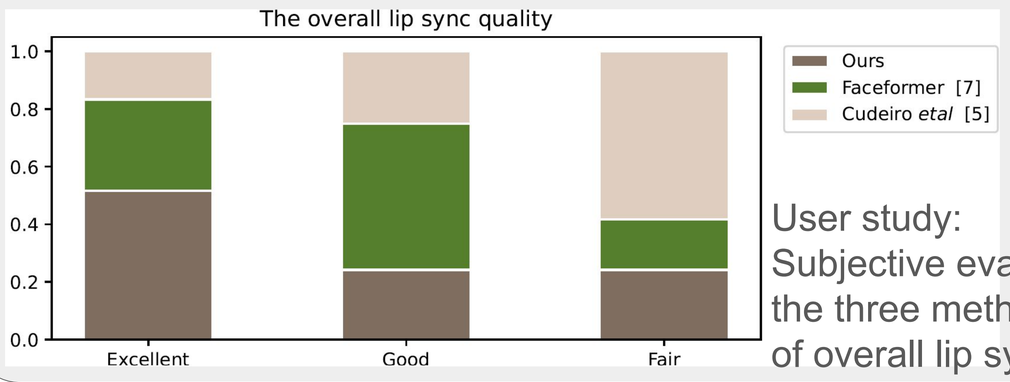
$$\mathcal{L}_{stageII} = \mathcal{L}_{geometry} + \mathcal{L}_{temporal} + \lambda \mathcal{L}_{lap-mod}$$



Stage I training: Allows to synthesize natural speech—TTS.



Stage II training: We can generate facial animation, modify style and content.



Content reenactment example

User study: Subjective evaluations of the three methods in terms of overall lip sync quality.

Dataset:

Stage I: ESD + LJSpeech

Stage II: VOCAset, BIWI

Comparisons of bilabial mouth closure for "our ex*p*eri*m*ent's *p*ositive outco*m*e"