**TITLE**

**Author(-s):** First name Last name

**Advisor(-s):** First name Last name

University (Country)

***Abstract.*** *Summary of the work must have no more than 1500 characters. Text must be typed in italics with a single interval after the name of the university.*

***Keywords:*** *up to 10 keywords. Text must be typed in italics.*

**І. INTRODUCTION**

Text…

**II. LITERATURE ANALYSIS**

This section can be split into several subsections (if necessary). Titles of subsections must be typed in bold. The subsections must be numbered consecutively within the section. Subsection text must be typed after a single interval.

References are placed in square brackets in the order of mention in the text, e.g. [1], [2–7].

**2.1. History and something, something**

Text…

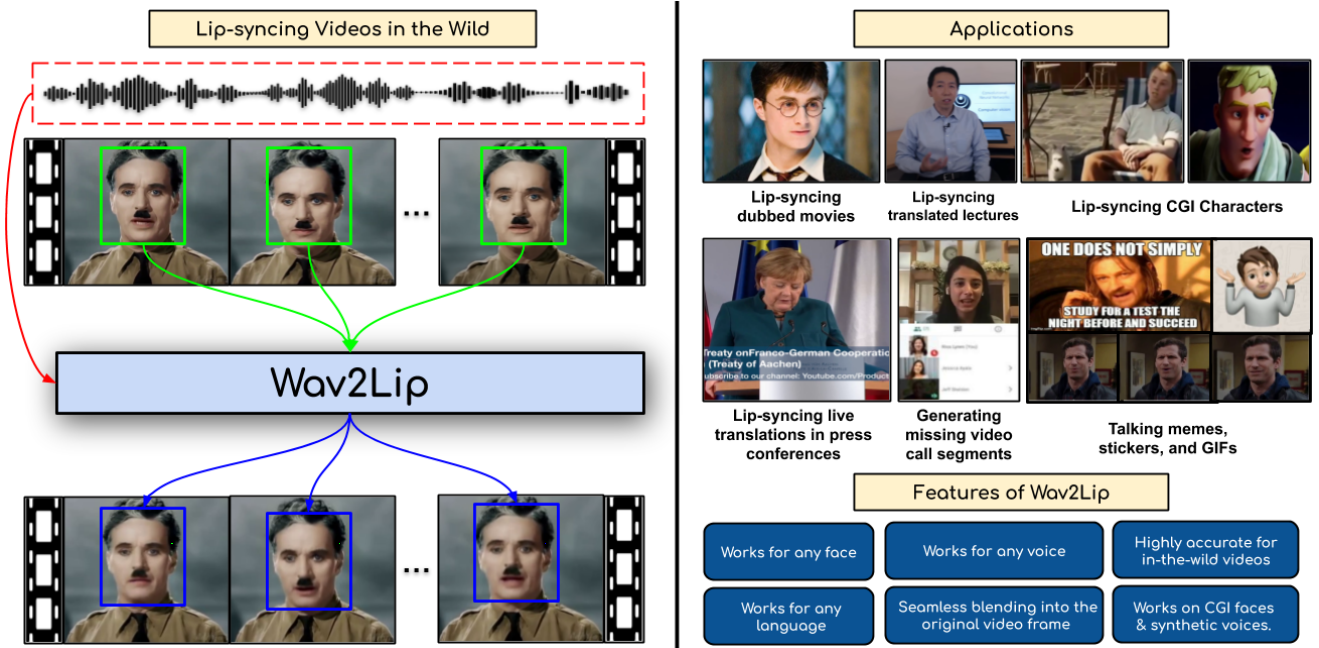
**2.2. Neural networks.**

Как можно было заметить ранее - lip sync - сложная и кропотливая работа актёров и аниматоров, однако и такой материал легко испортить ошибками как со стороны актёров, так и со стороны людей за озвучкой.

Поэтому возникла потребность в качественном lip-sync-е с минимальными затратами для создателя, и при этом - высоким качеством материала.

Как и во многих других отраслях логичным становиться использование нейронных сетей для решения таких сложных заданий.

Так для решения проблем переводов может использоваться работа Prajwal K.R., Rudrabha Mukhopadhyay, Namboodiri V. P., Jawahar C.V. (2020) Wav2Lip[1?] подобная нейронная сеть, что анализирует как аудио, так и видео сингнал находит лицо говорящего человека и анализируя звук находит произносимые звуки, что имеют яркую артикуляцию.



Основным отличием этой нейронной сети от всех последующих сетей, что указанны в этой работе - наличие нескольких нейронных сетей для одной общей цели, но двух разных задач.

Так основной задачей являеться распознование фонем и генерация лица, однако так же присутствует вторая сеть, которая анализирует и оценивает качество сгенерированного материала, как на наличие артефактов изображения, так и артефактов генерации речевого аппарата.

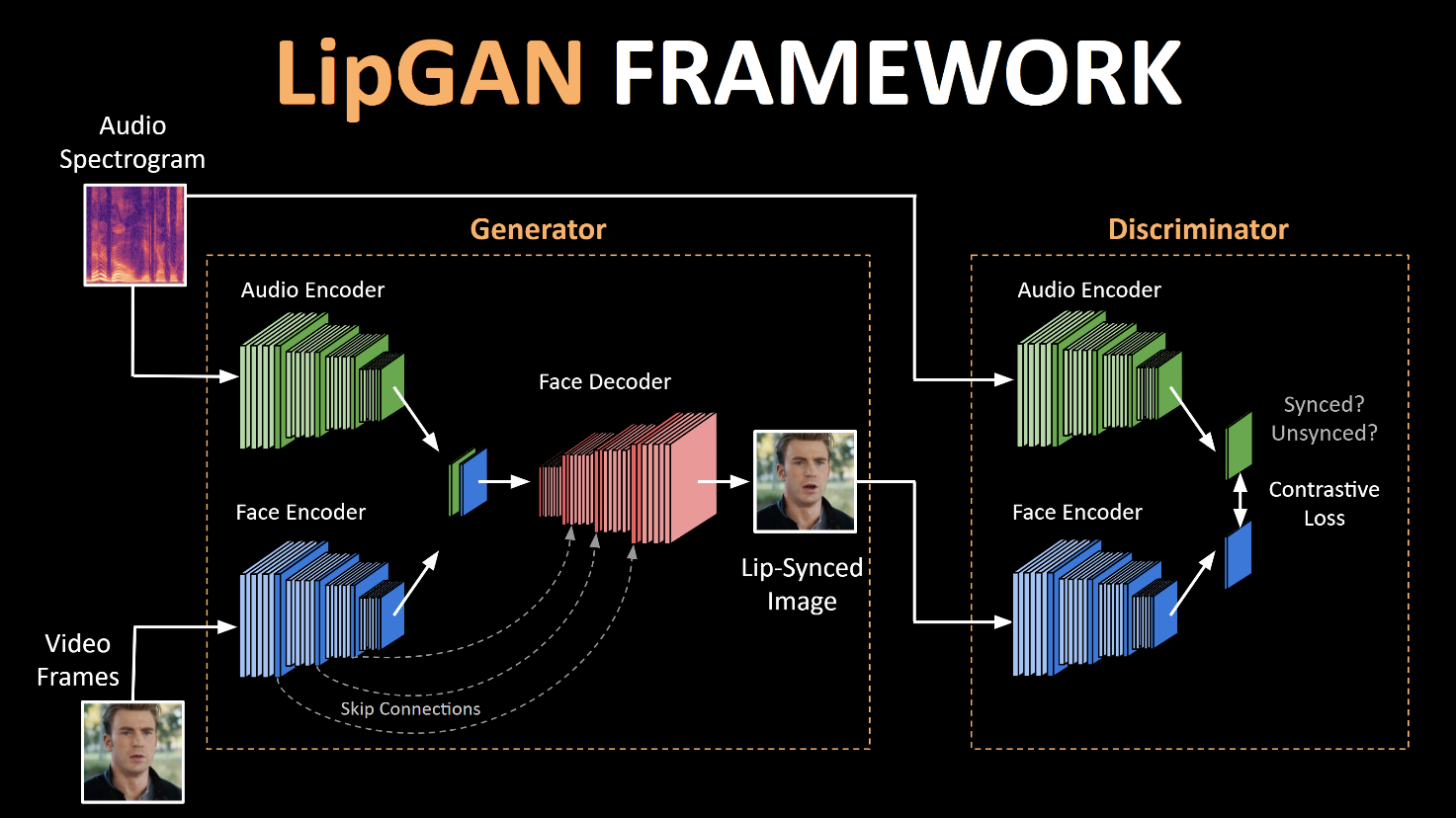
Эта же нейронная сеть показывает лучшие результаты на тестовых материалах, в разделах синхронизации и качества генерируемого материала. Результат становиться несколько хуже, если второй сети, ответственной за оценивание материала предоставить больше кадров для оценки. LipGAN использует всего один кадр, и пусть это может сделать более синхронизированное видео, на практике качество артикуляции играет большее значение для людей, если принять во внимание то, что погрешности во времени не так уж и велики. Как показано в таблице 2.2.1 если увеличить кол-во кадров для анализа с 1 до 5, точность падает с 79 % до почти 92%, а остальные метрики становяться лучше примерно на треть.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Fine-Tuned? | Off-sync Acc | LSE-D | LSE-C |
| LipGAN T=1  T= 1 | ✓  × | 55.6%  79.3% | 10.33  8.583 | 3.19  4.845 |
| T = 3  T= 3 | ✓  × | 72.3%  87.4% | 10.14  7.230 | 3.214  6.533 |
| T = 5  T= 5 | ✓  × | 73.6%  91.6% | 9.953  6.386 | 3.508  7.789 |

Так же понятно, такая нейронная сеть применима только к тем работам, где говорящее лицо - человек.

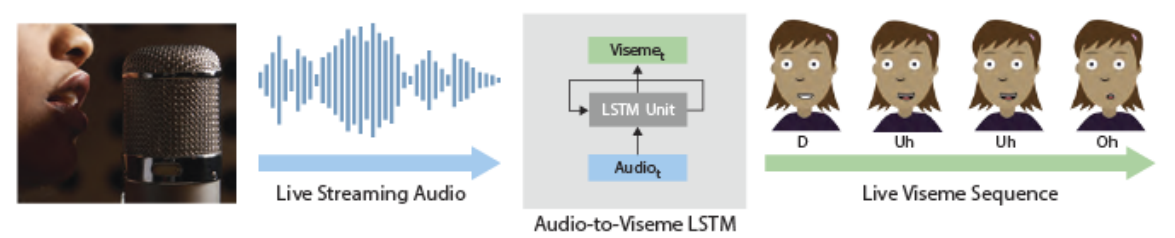
Однако не всегда нужно переводить какое-то произведение или выступление, например, если во время обучения появиться необходимость показать историческую личность с её цитатой ранее бы пришлось рисовать лицо, или и вовсе - прилепить рисованые выбивающиеся губы к портрету личности, что портит общее ощущение важности цитаты и придаёт комичности.

С использованием Prajwal K.R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, Jawahar C.V. (2019) LipGAN[2?] можно решить эту проблему, и ещё некоторые другие. Так нейронная сеть LipGAN анализирует видео поток, который может быть и вовсе - статичной картинкой и звукуовую дорожку и на выходе даёт реалистичное видео с синхронизироваными движениями рта и звука. Так например можно "Оживить" Елизавету I и вручить ей её цитату на уроках истории.



Эта же технология по заявлению авторов позволит разработчикам игр создавать правдоподобные анимации лица для персонажей своих игр с множеством локализаций.

Однако, если вы автор собственных анимаций или стример что не хочет показывать своё лицо, то вам может помочь работа Prajwal K.R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, Jawahar C.V. (2019) CharacterLipSync[3?], эта нейронная сеть использует заранее нарисованные эмоции и звуки для создания видео потока в реальном времени. Может использоваться как с трекингом по лицу, так и без него. Как и во всех нейронных сетях выше - производиться анализ аудио-дорожки, однако видео сигнал здесь полностью генерируем без использования трекинга (что часто делаеться во время стримов), таким образом нейронная сеть не только анализирует аудио и подбирает необходимый сейчас рот для персонажа, но ещё и создаёт переход между каждой сменой, что вручную делать намного дольше, даже если бы все рты были расставлены сразу.



Одной из техник распознования фонем может быть использование вейвлет-преобразования, с применеием нейронных сетей для подбора коэффициентов. Используя данные преобразования, можно найти границы фонем. Фонема - минимальная смыслоразличительная единица языка. Из библиотеки звуков можно подобрать фонему, а это так же - хорошая работа для нейронных сетей. Подбор порогового коэфициента изменяет то, насколько малая разница должна быть между звучанием для разграничения фонем. Так при малых значениях коэфициента некоторые близко расположенные звуки могут повторять предыдущие, или образовывать слитные фонемы. Поэтому правильный подбор коэфициента важен для качественного определения текущего звука, что позволит сделать липсинк качественнее.

**III. OBJECT, SUBJECT, AND METHODS OF RESEARCH**

Text…

**IV. RESULTS**

This section can be split into several subsections (if necessary). Tables and figures must be referenced in the text and formatted as follows:

a

b

c

Fig. 1. Title of the figure…:

a – first subfigure, b – second subfigure, c – third subfigure

Table 1. Title of the table…

|  |  |  |  |
| --- | --- | --- | --- |
| **Header A** | **Header 1** | **Header 2** | **Header 3** |
| **Header B** | Data | Data | Data |
| **Header C** | Data | Data | Data |

# V. CONCLUSIONS

Используя указанные выше методы в сочетании с, например, интерфейсом безмолвного доступа можно увеличить точность распознования фонем, что позволит увеличить как качество их размещения за счёт того что нейронная сеть будет тратить меньше времени на их поиск, так и качество визуальное за счёт освободившегося времени. Так же без использования ИБД, но пользуясь некоторыми особенностями организма, что были исследованы во время создания таких интерфейсов. Имея картинку должного качества можно анализировать смещения горла во время разговора, или, если есть такая возможность - анализировать перемещения языка и изменения формы рта. Последние два совета куда лучше применимы к людям озвучивающим записи, однако комбинация с ИБД позволит улучшить точность для всех сторон записи.

Также можно указать то, что при наличии более совершенного алгоритма машинного зрения и более точной калибровки оценивания для сети Wav2Lip, можно добиться лучших результатов по качеству генерации и более точную синхронизацию.

И поскольку в последнее время мирвые гиганты технологий как Nvidia, Samsung и другие создавали продукты с применением этой технологии, есть вероятность её улучшения в ближайшем будущем.

**VІ. REFERENCES**

1. Prajwal K.R., Rudrabha Mukhopadhyay, Namboodiri V. P., Jawahar C.V. (2020). A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. https://arxiv.org/pdf/2008.10010.pdf

2. Prajwal K.R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, Jawahar C.V. (2019) Towards Automatic Face-to-Face Translation https://dl.acm.org/doi/10.1145/3343031.3351066

3. Deepali Aneja, Wilmot Li (2019). Real-Time Lip Sync for Live 2D Animationhttps://arxiv.org/pdf/2008.10010.pdf

Sources must be cited in the standard APA style:

https://apastyle.apa.org/style-grammar-guidelines/references/examples