**Analysis of lip-sync technologies and possible ways to improve them (?)**

**Authors:** Isaiko Svitlana, Pohorieltsev Pavlo

**Advisor:** Muntian Iryna

Professional College of Industrial Automation and Information Technologies of the Odessa National Academy of Food Technologies (Ukraine)

***Abstract.*** *Lip-sync technology advanced over the years. From hand drawn and synced images of first animated movies, to human live performances on stage, then on screen and finally – using advanced neural networks and variety of other techniques, to automate this process.*

*This work covers a bit of a history behind the lip-sync, dives into technologies of present and discuses about future possible growth ways.*

*The objective of this work is to provide reader with information about past and present techniques of lip-sync.*

***Keywords:*** *lip-sync, neural networks, wavelet transform, animation, translation, GAN, LSTM*

**І. INTRODUCTION**

Lip-sync можно описать по-разному: а) перенос артикуляции актёра озвучения на мультипликационного персонажа; анимация рта; б) синхронизация звукового и визуального ряда. В этой работе мы рассмотрим то, как развивался lip-sync, то к чему он пришёл, и в заключении – то, как он может развиться.

**II. LITERATURE ANALYSIS**

**2.1.** **Lip-sync appears.**

Идея lip-sync, или звукового дубляжа, возникла практически сразу с появлением звукового кино. Считается, что одним из первых фильмов, в котором песни записывались отдельно в студии, был легендарный американский мюзикл «Бродвейская мелодия» 1929 года.

Во время производства фильма руководство студии было недовольно тем, как звучит один из самых динамичных и ярких номеров мюзикла Painted Doll. Звукорежиссер Дуглас Ширер предложил решение, которое стало настоящим прорывом – записать песню отдельно. Актеры в свою очередь будут исполнять номер под уже готовый трек. Это помогло сохранить техническую сложность, визуальный эффект выступления и при этом не жертвовать его качеством.



Далее в 50-х годах с развитием кино, lip-sync стал ещё более популярным. Эту технику популярно было использовать также в разных телевизионных программах. Скептическое отношение к lip-sync со стороны зрителей потихоньку утихало и отношение улучшилось. Его использование стало обычной практикой.

Нормализовали lip-sync и «Скопитоны» — музыкальные аппараты, изобретенные во Франции, которые могли проигрывать видео. В 60-х годах их устанавливали в барах и кафе. Пользователь сам выбирал видео. Чаще всего это были выступления артистов, на которых они пели под фонограмму. Можно сказать, что это были первые прообразы клипов. В 1981 году появился телеканал с музыкальными клипами MTV. Частично благодаря этому lip-sync окончательно закрепился музыкальной среде. [1]



Использование lip-sync можно встретить и в анимации. На примере первого полнометражного мультфильма компании Дисней “Белоснежка и семь гномов” можно показать, что это был очень сложный и долгий процесс. Над проектом трудились на протяжении трех лет более 100 аниматоров, изготовлено было более миллиона рисунков, но в финальную версию мультфильма вошло около 250 тыс. [2]

Благодаря тому, что ничего не стоит на месте и всё развивается процесс анимации в данный момент более оптимизирован. В настоящее время, Дисней использует компьютерную анимацию, а первым новаторским мультфильмом был “ Рапунцель: Запутанная история”. Анимации которые развивались, начиная с миллиона рисунков пришли до компьютерной анимации, то есть в своем роде совершили прорыв.



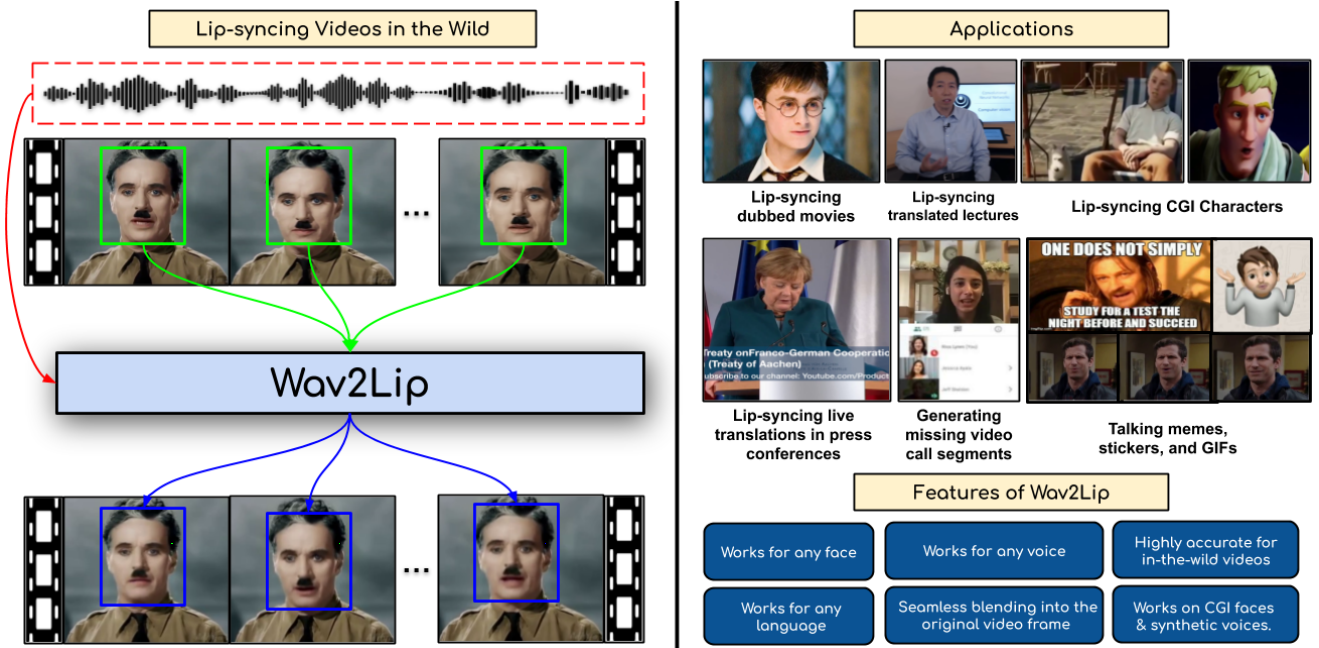
**2.2. Present of neural networks.**

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 Wav2Lip [3] 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 указаны в этой работе - наличие нескольких нейронных сетей для одной общей цели, но двух разных задач.

Так основной задачей является распознавание 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.

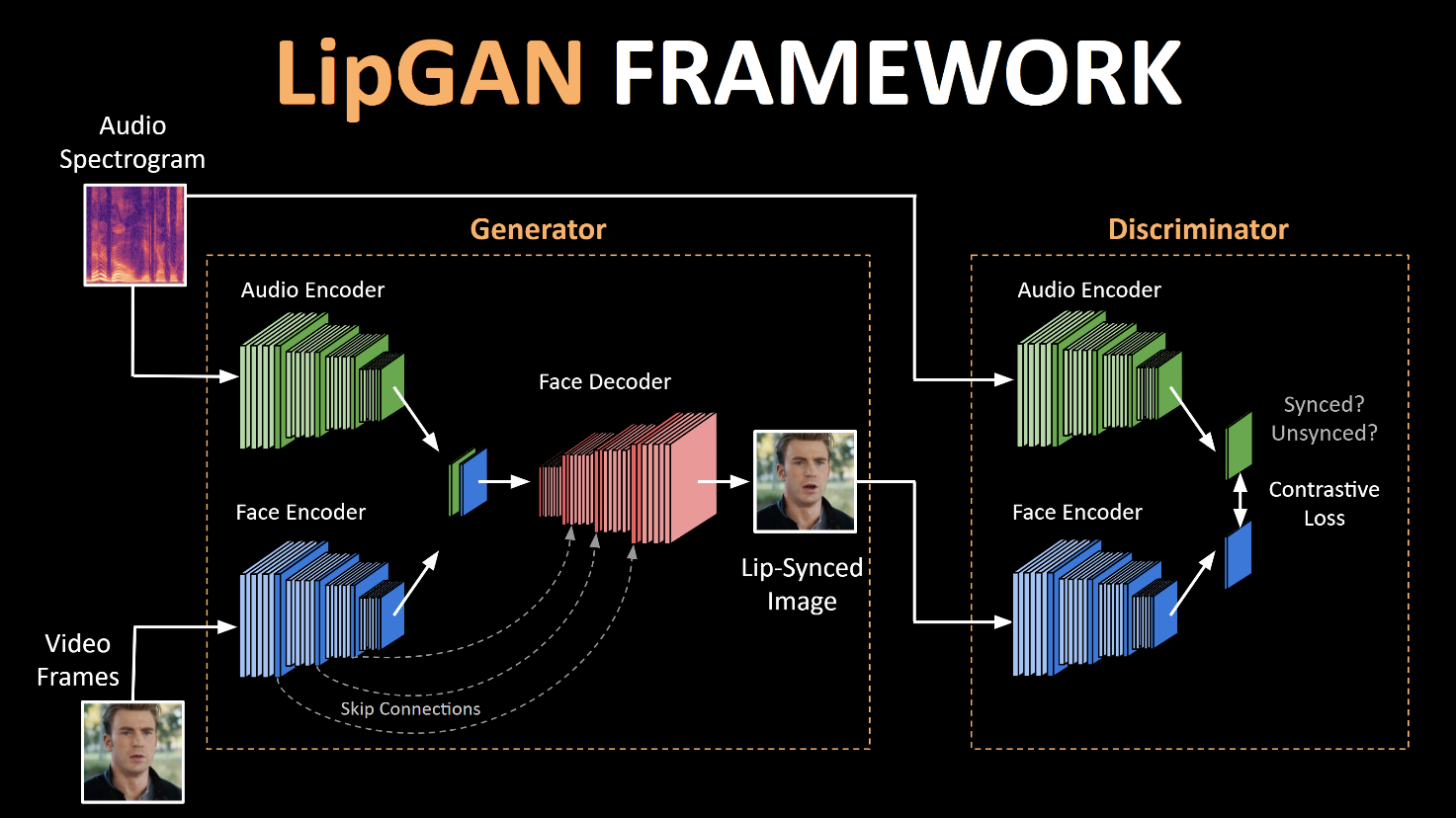
Эта же нейронная сеть показывает лучшие результаты на тестовых материалах, в разделах синхронизации и качества генерируемого материала. Результат становиться несколько хуже, если второй сети, ответственной за оценивание материала предоставить больше кадров для оценки. LipGAN использует всего один кадр, 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.

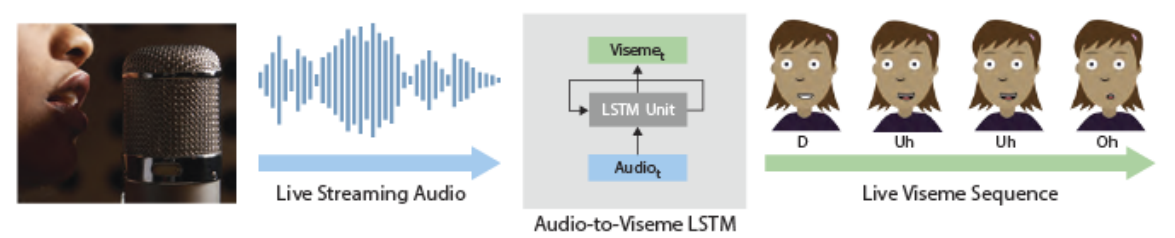
Однако не всегда нужно переводить какое-то произведение или выступление, например, если во время обучения появиться необходимость показать историческую личность с её цитатой ранее бы пришлось рисовать лицо, или и вовсе - прилепить рисованные выбивающиеся губы к портрету личности, что портит общее ощущение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

С использованием LipGAN [4] можно решить эту проблему, и ещё некоторые другие. Так нейронная сеть LipGAN анализирует видео поток, который может быть и вовсе - статичной картинкой и звуковую дорожку и на выходе даёт реалистичное видео с синхронизированными 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 CharacterLipSync [5], 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 делается 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 распознавания фонем может быть использование вейвлет-преобразования [6], с применением 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 коэффициента изменяет то, насколько малая разница должна быть между звучанием для разграничения фонем. Так при малых значениях коэффициента некоторые близко расположенные звуки могут повторять предыдущие, или образовывать слитные фонемы. Поэтому правильный подбор коэффициента важен для качественного определения текущего звука, что позволит сделать lip-sync качественнее.

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

Объект исследования: Методы lip-sync (-а), история lip-sync (-а)

Субъект исследования: Возможные улучшения текущих методов lip-sync (-а)

Методы исследования, использованные в данной работе: анализ научной литературы и публикаций, генерация возможных путей улучшения текущих методов.

**IV. RESULTS**

Используя указанные выше методы в сочетании с, например, интерфейсом безмолвного доступа можно увеличить точность распознавания 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 мировые гиганты технологий как Nvidia, Samsung и другие создавали продукты с применением этой технологии, есть вероятность её улучшения в ближайшем будущем.

# V. CONCLUSIONS

# В заключение можно сказать что lip-sync технологии развивались с середины прошлого века, и продолжают развиваться и по сей день. Используя современные методы анализа и синтеза можно создавать творческие работы быстрее и легче, чем когда-либо до этого. What a time to be alive!

**VІ. REFERENCES**

1. Julia Zaykova (2020, June 19). Read Your Lips: A Brief History of Lip-sync. Heroine. https://heroine.ru/chitaj-po-gubam-kratkaya-istoriya-lipsinka/

2.RBK.(2017, December 17). How Disney created the first feature-length cartoon [Video]. YouTube. <https://www.youtube.com/watch?v=-8LiVIufw4k>

3. 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>

4. 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

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

6. Medvedev M.S. (2004). Phone segmentation of a speech signal using a wavelet transform http://www.ict.nsc.ru/ws/YM2004/8614/Medvedev.html