Enabling Multilingual Communication: Automated Lip-synchronization Dubbing for Albanian Videos

ALDO DIKU

UNIVERSITY OF NEW YORK TIRANA

JULY 2025

This is to certify that I have read this project and that, in my opinion, it is fully adequate, in scope and quality, as a thesis for the degree of Bachelor of Arts in Computer Science.

(Title and Name)
(Project Advisor) _____ _____

This is to confirm that this thesis complies with all the standards set by the Department of Computer Science of University of New York Tirana.

Date: Seal/Signature:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

First Name, Last Name:

Signature:

ABSTRACT

ENABLING MULTILINGUAL COMMUNICATION: AUTOMATED LIP-SYNCHRONIZATION DUBBING FOR ALBANIAN VIDEOS

Diku, Aldo.
BA. in Computer Science
Thesis Advisor: Prof. Miralda Çuka
July 2025, 183 pages

Using available methods and tools (which i will mention when i gather all the tools and methods) to create lip-synced dubbed videos from an albanian video input and outputting a video of the same speaker in another language with the lips moving according to that languages movements.

Keywords: Lip-sync, dubbing, albanian language

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Miralda Çuka, for their invaluable guidance, unwavering support, and insightful feedback throughout this project. Their expertise and encouragement were instrumental in shaping this work.

Finally, I am grateful for the support of my family and friends. Something something a little bit longer and better made.

Contents

3 Introduction

3.1 Background of the study

Option 1
The increasing globalisation of online content consumption presents a significant challange for video creators aiming to reach diverse audiences, thereby creating the need to produce multilingual versions of their work. This often involves an increased workload, time and other resources, a perfect example of this is a German content creator in YouTube called der8auer (link?). Der8auer makes the same content in two languages, in english and in german. This workflow includes a lot steps, filming, editing, storing and uploading the content are all things that take time. This project addresses the challange of automatic translation and lip synchronization of video content into multiple languages. The main focus of the project will be the Albanian language as an underrepresented and low resource language. To bridge a gap in current research and practical applications of automated video translation and lip synchronization.
Option 2
Being able to make only one video and have it in multiple languages will help creators but also the target audience as well. Cutting down on the time and money it takes to make two or more videos with the same subject but in different languages. One such case is a youtube creator named der8auer from Germany, for the same topic he creates two videos, one in english and one in german. Sometimes his videos are quite long and having to do them again in another language will tire you out. Having a model that can take your video, translate it into another language and synchronize your lips to the movement of the language of your choosing would be a tool in your arsenal.
This project focuses on lip synchronization video dubbing on Albanian language as a low resource and underrepresented language, addressing a significant gap in current research and practical applications.
Some of the usages of this lip-sync dubbing technology are: video dubbing and translation, real-time Face-to-Face translation(this would be into the future) and multilingual communication, gaming and virtual environments (some games come in different languages but the characters are coded to move their lips only according to one language, although some companies have done work to change this depending on the language the users select, cyberpunk 2077 used JALI ai for animation lip-sync) reduced cost and labor and time in this case, entertainment and content creation, speech recognition and lip-reading. Film, entertainment and media production. Education and training videos, especially in cases where there is a diverse workforce.

3.2 How lip synchronization works

Phonemes are the distinct units of sound in speech, or the individual sounds that make up speech and visemes are visual representations of phonemes. Visemes serve to map sounds (phonemes) to corresponding mouth positions or shapes. The goal is to go from the audible (phonemes) to the visual (visemes). Visemes group similar-looking mouth shapes, which reduces the complexity required for animation. Mapping phonemes down to a smaller number of visemes gives artists fewer expressions to pose.
The history of lip-syncing comes from animation and in particular two-dimentional cartoons

where artists had to make facial movement according to the sounds of speech [[2]].

3.3 Methodology

A key limmitation encountered was the lack of suitable Albanian video datasets, specifically open-license content with direct-to-camera speech, reflecting the broader challange of working with low-resource languages in multimedia applications. To address the first issue, custom video contennt was created featuring self-recorded footage in a one by one aspect ratio and lowest camera setting to save . The first step for this project was the data creation, seeing as there is not a lot of free and publicly available free-use content I can use we had to create the data myself.

Videos were recorded at 1440x1440 resolution (1:1 aspect ratio) at 30 frames per second, with durations ranging from 10 seconds up till 5 minutes. The study at (citation needed for the optimal length of time of the videos) found that the optimal length time was for videos with one word in them. To adapt to this, a script was made to create small chunks of videos from a lengthy video to be closer to that optimal format. Videos could have been created by following and reading a script thereby avoiding the transcipiton step, however we want to create a workflow that will take any video of an albanian speaker and turn it into a dub lip synched video.

From the videos the next step is audio extraction which will be used for transcription and for the base of the voice cloning tool before inputing it into the final lip synchronization model. For the audio extraction ffmpeg library was used, and audio was extracted in a lossless codec, 16000hz sample rate and signed 16-bit Pulse Code Modulation (PCM), per Google Speech-to-text API documentation [1].

The audio gets sent through the transcription service, and gets saved in a text file along with the name of the transcripted audio file. When all audio files have been transcripted, they are translated using (PUT TOOL USED HERE) in (PUT NUMBER OF LANGUAGES USED HERE) languages. (SHOULD MENTION BEFORE WHAT I TRIED FLORJAN ASR FOR EXAMPLE, PESHPERIMA V2, TRAINING OUR OWN MODEL WITH DATASET FROM FLORJAN ETC.) (MAYBE NOT IN THIS SECTION OF THE PAGE????) (Speech-to-text API from Google Cloud services offered a wide range of languages and Albanian was a supported language, utilizing their Chirp 2 model, the transcriptions were very accurate, even more so as google gives you a confidence score about the sentence or even individual words.)

After translating, the translated text and the original audio are used to make the voice clone. We need the original audio of the transcript because of emotions displayed in the audio can be used by the voice cloning tool.

Cloned voice and the original video from where the cloned voice was based on are then used to train the lip synchronization model.

Several models (citations needed) were available to try for albanian transcription although with different accuracy. Automatic Speech Recognition (ASR) models are measured using Word Error Rate (WER) where a score of 0% means perfect translation and 100% is the worst case, working on the albanian language I focused only on those models that had support for it. Whisper by OpenAI was one such model, availabe on HuggingFace. To make such model have a better performance (need some results of the model transcribing an audio file) we need to fine-tune it. (THIS SECTION CAN MOVE TO THE "WHAT I TRIED TO DO AND FAILED" SECTION)

Here lies another issue with working with a low resource language, not only are there low resources for video, but also for audio which we could use to fine-tune the model. There were two models made especially with albanian language in mind, albanian-asr and peshperima-v2. Because of the low ammount of resources albanian-asr was only able to get to a 46.3% accuracy, which does not help at all. Peshperima-large-v2 was also not successful when testing. Whisper had a mulltiple sized models which required fine-tuning before being able to use it for transcription. The only trainable model size for a reasonable ammount of time was the whisper-small. The low ammount of data for the audio made the whisper-small not do much better than albanian-asr.
Another option was Wav2Vec and its different variations made from Facebook/Meta, this one also failed like the previous models because of lacking datasets. No model tested was able to perform in a state where they could have been used in a real world application, for that reason we needed to switch to a finished model that did not need fine-tuning.
Google cloud service offered speech-to-text API options and for many different languages. One of their models was offered for albanian language and had great performance, giving a confidence score about the transcription which would prove to be very usefull since you can use it as a metric to decide if you want to accept the transcription or not.
The next step is the preparation of the dataset collected, which was in 1-5 minute video format. From this we extracted audio from the video and tried finding moments of silence inbetween the speaking in order to cut the video into chunks to then later on feed it into the Convolutional Neural Network (CNN). One of the reasons of why this is a hard problem is that you cannot just get the unique sounds and join them together to form a word, each unique sound changes depending on what is the letter or sound before it. Here is where the deep learning/machine learning helps as it studites large amounts of video and audio to notice these features. **One of the challanges of creating a lip synchronization in videos is the need for the lip movements to accurately align with a specified target speech segment, especially in multilingual and unconstrained environment. Visual data falling out of sunc with updated audio and inaccurate lip movements in target videos.**
CNN and Generative Adversial Networks (GANs) to create the lip movements that sync up, this combined with a Discriminator Network which would try and detect the GANs fake lip movements and real ones, pushing each other to get better.

3.4 Work Steps Overview

From the video we need to get the transcription in albanian, for this step several models were tried and tested although only the google cloud API were the best performing one. The next and easiest step to implement is translating the transcription to the target language, translation has come a long way and there are many tools which can achieve high accuracy, we decided to use googe translate. The following task is to take the translated transcript and use a voice cloning tool/model to make the voice dubbing. **There are several models availabe which have not been tested by me yet.** After this we have to use object tracking to track the mouth area in the video to then use the deep learning model for lip frame creation. **This depends heavily on the model the will be applied.**

3.5 Audio

The first step of the project is extracting audio from the videos, this was done using ffmpeg. Going by the suggestions of googles speech-to-text API best practices, the audio was sampled at 16000 Hz, converted into a singned-integer bitrate, and in a lossless format.
A simple test was done to check which specs were best suited for the videos we were creating for the dataset. These were all tested using googles speech-to-text API giving a confidence score for each of the tests. If we are going to do tests.
Need to add the test results here.
Might need to split audio file into two mono files or downmix into one mono file.
One group of audio files were made in a lossless format, signed 16bit PCM (linear16), 16000hz sampling rate. While the other folder, was made with the same codec but with 48000hz sampling rate.

3.6 Notes

The google api workflow was difficult to implement for it to work correctly and there were more lines of code to implement it than the other models. The accuracy of the API was not the best, but it was the only one that gave a confidence score for each of the tests. A new model made by Kushtrim Vioska, based on openai's whisper-large-v3-turbo and trained on 200 hours of albanian audio was able to transcribe the audio with a great accuracy, unfortuantely the model at this time does not have the weights open source and was only able to be used through a gradio app which slowed down the process of transcribing the audio significantly and the resources of the huggingface model are volitale because of the amount of requests it gets.
Going through the next step of the workflow we move from the transciption to the translation of the text, which can be done with high accuracy using a lot of different models which **need to be tested and researched more thoroughly**.
Using the translated text into the language of our choice we can then try and use a voice cloning model to make the voice dubbing. A lot of models were available for this task, Nari Labs released Dia, in 22nd April 2025, coqui-ai, which a lot of other models are based on. A paper in 5th of May 2025 was released by **Yemin Shi et al. Voila: Voice-Language Foundation Models for Real-Time Autonomous Interaction and Voice Role-Play** which created a foundation model for realt-time autonomous interaction and voice role-play. https://arxiv.org/abs/2505.02707.
It can be posisble to make a comparison system between the google API and the Kushtrim Vioska model to see how much do they match. Not sure what we can achieve with this. Google API does give us a confidence score for each of the tests. We can probably use this to see how much the model is confident in the transcription and then use that to our advantage. If we also use the word confidence we can then compare to see if Kushtrims model made a different transcription, whenever the confidence is low we can use Kushtrims model to make a transcription.
Additionaly, [4] used "vVISWA" dataset which hade isolated words or independent speech, which worked great in combining the audio and video in the trainning set, and according to them reducing overfitting due to its inherent data augmentation effect.

3.7 Studying the literature

There are a lot of models and papers which have been made for the task of lip-synchronization, but most of them are focused on english and other high resource languages. Pawar, D. et al. (2024) used Generative Adversarial Networks (GANs), using both audio and visual features techiques like MFCCs and VGG-M-based CNNs to achieve accurate lip-synchronization. [4] shows some of the models using GANs as their architecture. [3] used a combination of a deep neural network with one dimension and a Long Short-Term Memory (LSTM) to generate a face model from speech input.
Traditionally, methods required highly controlled and precisely aligned datasets, making them less flexible and scalable for diverse inputs. Newer approaches aim to reduce this dependency.

3.8 Questions raised

List of questions:

- How do we decide if we are going to use a transcription or not?
- How do we know if the transcription is correct programatically?
- How should the data be prepared for the lip-synchronization model?
- Video format, 30 frames per second or 25, quality of the video, resolution, etc.
- Audio format, mp3, lossless, bitrate, etc.
- How does the duration of video chunks used during preprocessing impact the performance of the final machine learning model, and what chunk length yields the best trade-off between information retention and computational efficiency?
- How are voice cloning models evaluated, as it feels like it should be done with a human evaluation and that is faulty most of the time?
- Which would be the best model for the lip-synchronization task based on the data we have?
- Best way to handle differences between original audio timing and the voice cloned sample timing. (slow down the voice? speedup)

AS for the optimal length of the video chunks, the study done by [4] faced challanges with lengthy continuous speech, inlcuding time-consuming merging issues and potential overfitting. This was addressed by using the "vVISWA" dataset which contains isolated words or indepentent speech. This improved the performance of the model a lot and reduced overfitting.

List of challanges:

- Low resources for the datasets in albanian, in video and audio.
- No open source models for the transcription of the audio from albanian.
- Voice cloning models being evaluated using a human evaluation and that is biased.

- Storage of data, both in the cloud and locally.
- Computing power.
- Training time.
- If the spoken albanian is with a heavy dialect, an Albanian ASR trained with large number dialects is needed
- If the cutting of videos is not done correctly, it is possible to cut a word in half
- Translating the transcribed albanian dialect. How possible it is? NLP script transformation + translate*
- Voice cloning emotions.
- Video chunking first then transcription, or transcription (have to keep timings) then chunking - might be more possible and more cost effective than the NLP to find the problems in the language/chunked words but not for dialects.
- Could probably use the timings from google speech-to-text api to make the cuts since it offers that. Making the better option to be transcript first then cut/chunk second.
- Multiple Ethical challanges
- Are NLP-s better than LLM-s in these cases/uses.?

List of improvements:

- Better dataset, quantity and quality, diversity (age, gender, moustache, skin color, lip color- lipstick, ethnicity, also in this case, dialect)
- Separation of dataset into single word video for better model performance.
- Albanian ASR open-source model, google speech-to-text API is paid service, and Kushtrims uses huggingface spaces which is slow
- Comparison between Google Speech-to-text API and Kushtrim ASR
- Addition on an NLP to handle transcription errors
- NLP for dialect speech, to transform it into something that can be translated, maybe LLM is better in this case.
- Better voice cloning. More emotional expression
- Better face reproduction from the emotions of the voice clone.
- Real time lip-synch dubbing
- Separating the dataset into two groups the new dataset with videos close to one word per video and see the results of that and the original dataset with longer videos.
- NLP can also be utilized in the audio length check, if we have too much of a difference in audio length, we could apply NLP-s to change the sentence into something shorter or longer but keeping the sentence about the same.

References

[1] Google Cloud. *Speech to Text API*.

[2] Arien Kock Jonathan Gratch. An evaluation of automatic lip-syncing methods for game environments. Technical report, University of Twente, Department of Computer Science, University of SOurthern California, Institute for Creative Technologies, 2022.

[3] Xiaohong Li, Xiang Wang, Kai Wang, and Shiguo Lian. A novel speech-driven lip-sync model with cnn and lstm. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, page 1–6. IEEE, October 2021.

[4] Diksha Pawar, Prashant Borde, and Pravin Yannawar. Generating dynamic lip-syncing using target audio in a multimedia environment. *Natural Language Processing Journal*, 8:100084, 2024.