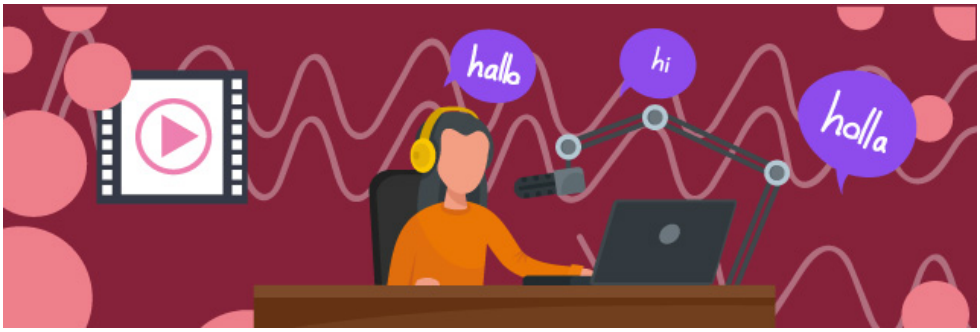


dubble

ESE 4510: Senior Design Final Report

Dubble: Automating the Dubbing Industry

April 20, 2022

Team 2:

Vineeth Veeramachaneni (EE ‘23, veevinn@seas.upenn.edu)

Kaitlynn Soo (SSE ‘23, kasoo@seas.upenn.edu)

Maya Guru (CIS ‘23, mayaguru@seas.upenn.edu)

Sahit Penmatcha (CIS ‘23, sahitpen@seas.upenn.edu)

Cathy Chen (CIS ‘23, cathych1@seas.upenn.edu)

Advisors:

Chris Callision-Burch (CIS, ccb@seas.upenn.edu)

Liam Dugan (ldugan@seas.upenn.edu)

Electrical and Systems Engineering & Computer and Information Science Inter-Departmental Senior Design project

Table of Contents

II. Executive Summary

Our project aims to automate the dubbing industry using machine learning and deepfake technology to replicate actors' voices in different languages. The goal is to improve the efficiency, cost-effectiveness, and quality of dubbed videos, with a focus on the film and television industry and the possibility to expand to other forms of voice and video content. Our target market is streaming services and large production studios such as Netflix and Disney.

On the technical side, we have currently developed a Speech-to-Text (STT) and Text-to-Speech (TTS) model to lay the foundations for our technology. To do so, we used a gaming desktop capable of handling the computational complexity and storage constraints necessary to train and run our model. Our final prototype outputs a video file in a target language given an input video file through a fully automated pipeline. The development of our Speech-to-Speech (STS) model has also begun, with the data collection and preprocessing completed. The eventual goal is for the Speech-to-Speech (STS) mode to solve many of the issues and loss of voice characteristics that currently occur with the STT and TTS models. Additionally, we have considered potential regulatory issues regarding the use of deepfake technology and have taken those into consideration when designing our model architecture.

III. Overview of Project

Our project aims to automate the dubbing industry using machine learning and deepfake technology to replicate actors' voices in different languages. By focusing on the film and television industry, we aim to solve three key problems with the current dubbing process: time, money, and quality. Currently, studios and voice actors spend a significant amount of time manually dubbing videos, with studios dedicating around 25% of their post-production time to the process. This is a costly and time-consuming endeavor that our technology aims to streamline and automate. In addition to time, the current dubbing process is also expensive, with studios spending around $100,000 per 90-minute video per language to dub videos. Our technology has the potential to significantly reduce these costs by automating the process and reducing the need for expensive voice actors and coordination. Finally, the quality of currently dubbed videos is often poor, with many viewers preferring subtitles over dubs because of the unnatural and dissimilar voices of the voice actors. By using deepfake technology, we aim to improve the quality of dubbed videos and make them sound more natural and similar to the original actors' voices. Our target market is streaming services and big studios such as Netflix and Disney, which are continuously globalizing their platforms through international shows and movies. Our technology has the potential to improve the dubbing process and make it more efficient, cost-effective, and of higher quality.

IV. Technical Description

Introduction

Dubble is an automated dubbing tool designed to make the process of dubbing films in multiple languages more efficient and accessible. By leveraging state-of-the-art technologies, Dubble aims to create a seamless experience for film producers, content creators, corporate localization teams and many other potential clients of ours. This design document outlines the core components, architecture, and technologies used in the Dubble automated dubbing tool.

Current System Overview

Dubble follows a six-step process:

1. Background audio separation using Spleeter
2. Speaker diarization and separation using pyannote by Hugging Face
3. Transcription and translation using Whisper and Google Translate
4. Audio generation preserving the original actor's voice and intonations using Coqui
5. Lip-syncing using Wav2Lip
6. Overlaying the translated and generated audio onto the lip-synced video

System Components

1. Audio Separation

Spleeter, an open-source library developed by Deezer, is used for the audio separation process. It uses machine learning to separate audio tracks into different stems, such as vocals and instrumentals. This process helps isolate the dialogue from the rest of the audio components. We save the audio file to recombine with the translated output later on, and we continue to manipulate the dialogue.

2. Background Voice Separation

After the background audio separation using Spleeter in Step 1, the clean audio track will be processed using pyannote by Hugging Face for speaker diarization in Step 2. Pyannote will analyze the audio track, detect speaker boundaries, and label each segment with a unique speaker ID. The output of pyannote will be a JSON or TextGrid file containing the start and end times of each speaker segment, along with the corresponding speaker ID. We are able to save specific voice embeddings for each speaker as we generate mel spectrograms for each piece of dialogue, and look for patterns to determine distinguished speaker voice characteristics.

3. Transcription and Translation

The transcription of the cleaned vocal track will be performed using the Whisper Automatic Speech Recognition (ASR) system. Whisper is a state-of-the-art transcription model, trained on 687,000 hours of audio data and text transcripts from 99 different languages. Whisper transcribes dialogue by timestep, which can be reconciled with the speaker timesteps from the previous step. Once transcribed, the text will be translated into the desired target language using Google Translate. This combination of tools will provide accurate transcriptions and translations for the subsequent steps.

```
[00:00.000 --> 00:02.000]  Aquí tienes tu café.
[00:02.000 --> 00:05.000]  ¿Y la noche buena qué? ¿Al pueblo como siempre?
[00:05.000 --> 00:06.000]  No creo.
[00:06.000 --> 00:08.000]  Vamos, que sigues enfada con tu hermano.
[00:08.000 --> 00:09.000]  Mira.
[00:09.000 --> 00:11.000]  Vean sobre unos décimos que no voy a devolver.
[00:11.000 --> 00:13.000]  Pugamos una medias.
[00:16.000 --> 00:18.000]  Imagínate que toca.
[00:18.000 --> 00:20.000]  Venga, pero me lo quedo entero.
```

Figure 1: Transcribed text (in original language) with time-stamps

```
[00:00.000 --> 00:02.000]  Here's your coffee.
[00:02.000 --> 00:03.000]  And the good night here.
[00:03.000 --> 00:05.000]  To the town as always.
[00:05.000 --> 00:06.000]  I don't think so.
[00:06.000 --> 00:08.000]  Come on, you're still angry with your brother.
[00:08.000 --> 00:09.000]  Look,
[00:09.000 --> 00:11.000]  look at us, we said I'm not going back.
[00:11.000 --> 00:13.000]  Let's take a half.
[00:15.000 --> 00:17.000]  Imagine it's your turn.
[00:17.000 --> 00:19.000]  Come on, but I'll stay all night.
```

Figure 2: Translation of original language text to target language text

4. Audio Generation

Coqui TTS, a state-of-the-art text-to-speech system, will be used to generate the dubbed audio. The system will preserve the original actor's voice and intonations, ensuring that the dubbed audio maintains the emotional impact and authenticity of the original performance. This is done by recreating the output dialogue into audio and transforming it to include the properties of the original mel spectrogram we created before. We use our speaker embeddings and time steps from our previous steps, along with the output text dialogue, to recreate an audio track in the new language with the original speaker’s voice.

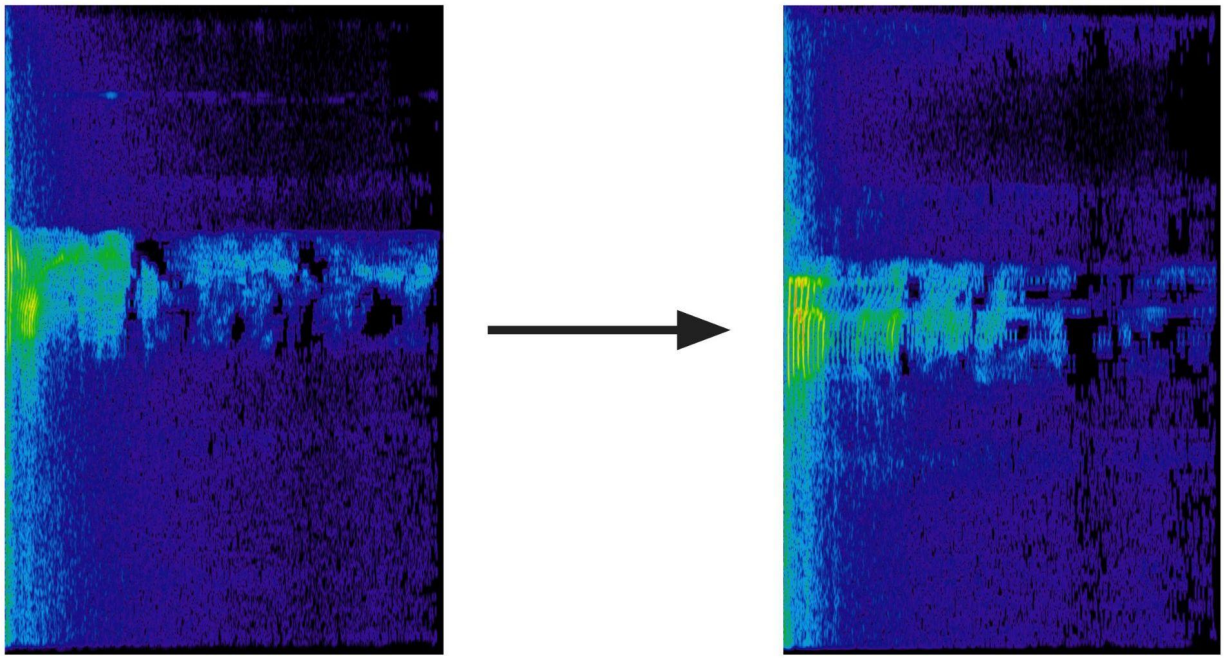


Figure 3: Mel Spectrogram input of “你好” (hello) in Chinese to output “hello” in English

5. Lip-syncing

Wav2Lip, a deep learning based lip-syncing model trained on face images, will be used to generate lip-synced video. Wav2Lip generates new frames with modified lip movements that correspond to changes in the audio spectrogram. This step ensures that the translated and generated audio aligns with the actors' lip movements, creating a natural and visually coherent final product.

6. Audio and Video Overlay

The translated and generated audio will be overlaid onto the lip-synced video, completing the automated dubbing process. The resulting video will have the actors' original performances synchronized with the dubbed audio in the target language.

System Architecture

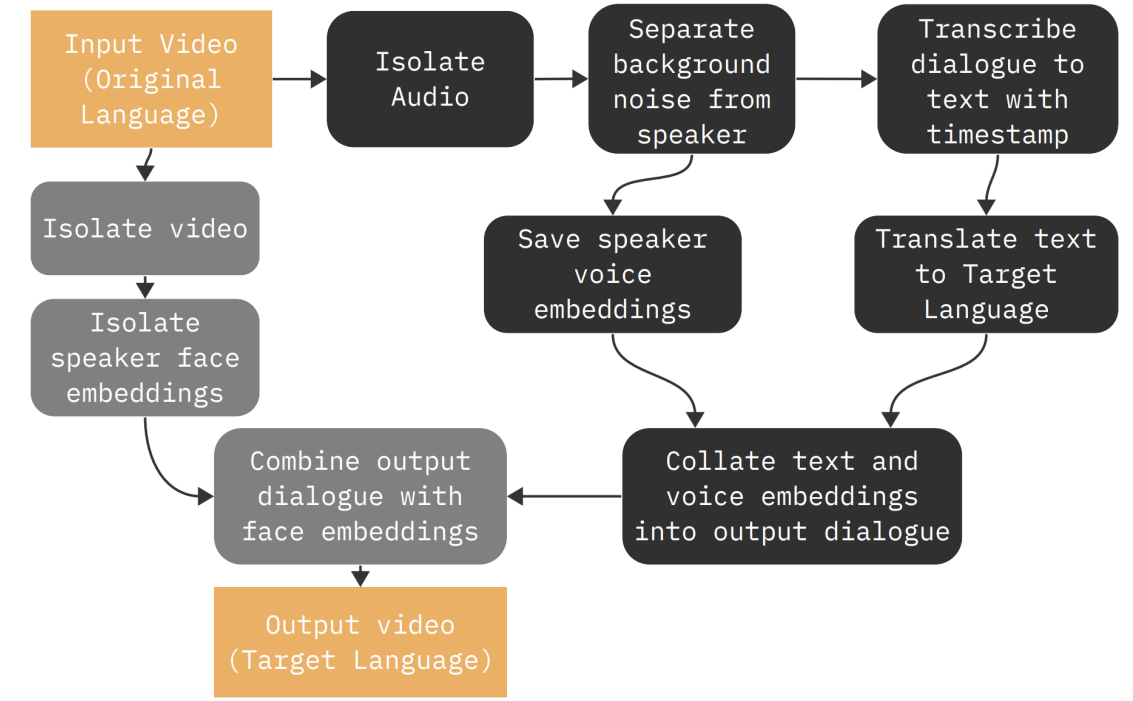


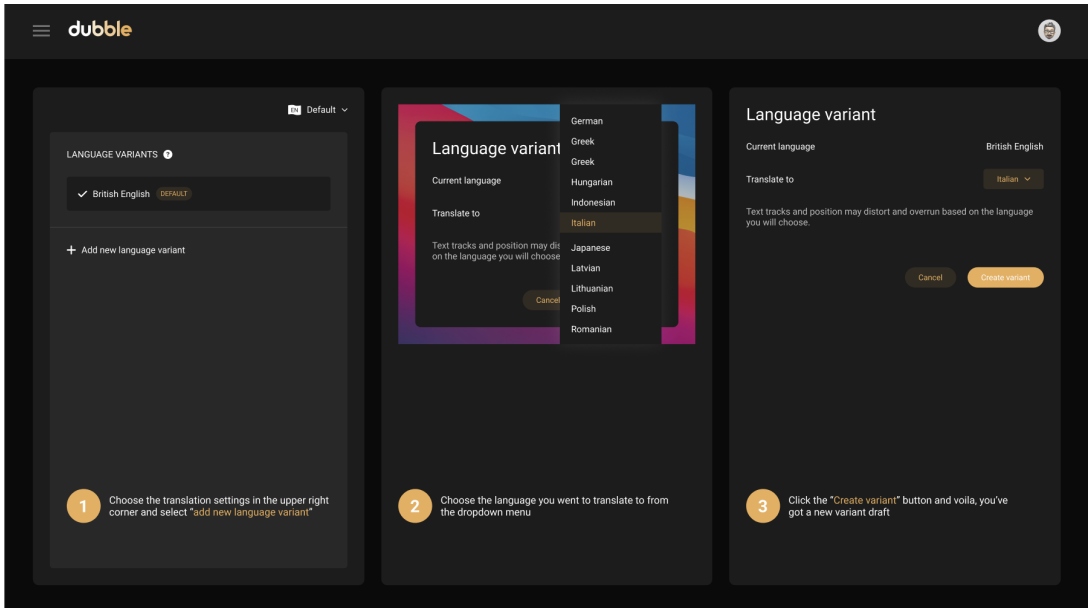
Figure 3: Speech-to-Text, Text Translation, Text-to-Speech Model Architecture

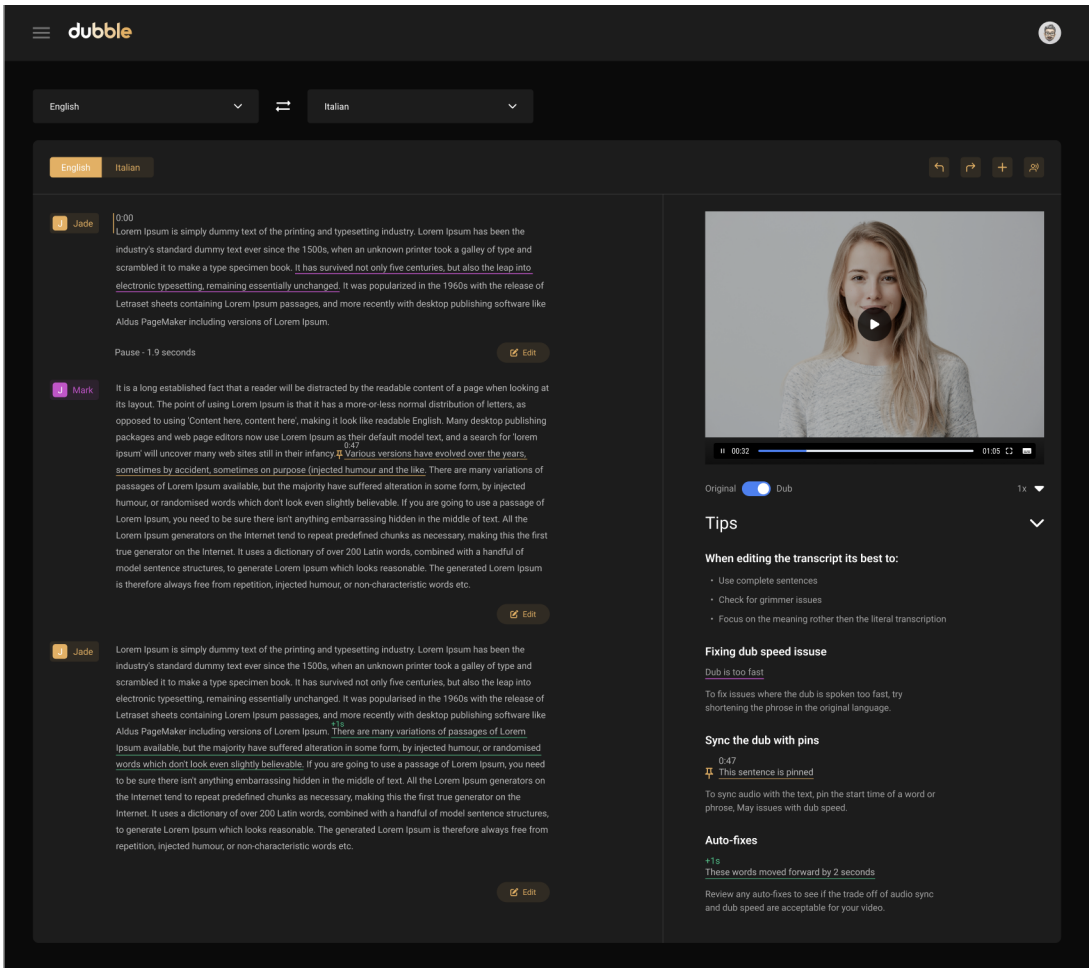
Modularity

Dubble is designed with a modular architecture, allowing for individual components to be updated or replaced with minimal impact on the overall system. This flexibility ensures that the tool remains adaptable to advancements in AI and machine learning technologies. Furthermore, the modularity allows for human input at each stage so the final input can be as accurate as possible.

Customizability

Our UI is designed to allow maximum customization for our users. They can select their desired target language as well as elect to manually verify and edit translated scripts if they would like. Furthermore, since the only input to our model is a video, this service can be used in any personal or commercial use case, with no tweaking required.


dubble
Default
LANGUAGE VARIANTS
British English
DEFAULT
Add new language variant
1
Choose the translation settings in the upper right corner and select "add new language variant"
Language variant
Current language
Translate to
Text tracks and position may dis
on the language you will choose
Cancel
German
Greek
Greek
Hungarian
Indonesian
Italian
Japanese
Latvian
Lithuanian
Polish
Romanian
2
Choose the language you went to translate to from the dropdown menu
Language variant
Current language
British English
Translate to
Italian
Text tracks and position may distort and overrun based on the language you will choose.
Cancel
Create variant
3
Click the "Create variant" button and voila, you've got a new variant draft

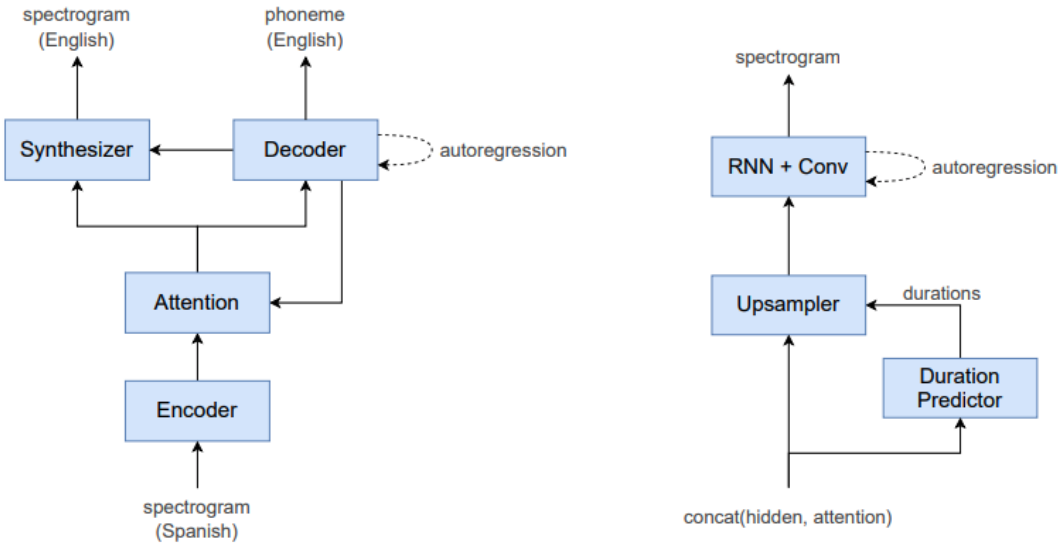

dubble
English
Italian
English
Italian
Jade
0:00
Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularized in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.
Pause - 1.9 seconds
Edit
Mark
It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a more-or-less normal distribution of letters, as opposed to using 'Content here, content here', making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for 'lorem ipsum' will uncover many web sites still in their infancy. 0:47 Various versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like. There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.
Edit
Jade
Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum. +1s There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.
Edit
00:32
01:05
Original
Dub
1x
Tips
When editing the transcript its best to:
Use complete sentences
Check for grimmer issues
Focus on the meaning rother then the literal transcription
Fixing dub speed issuse
Dub is too fast
To fix issues where the dub is spoken too fast, try shortening the phrose in the original language.
Sync the dub with pins
0:47
This sentence is pinned
To sync audio with the text, pin the start time of a word or phrose, May issues with dub speed.
Auto-fixes
+1s
These words moved forward by 2 seconds
Review any auto-fixes to see if the trade off of audio sync and dub speed are acceptable for your video.

Next Steps - Speech to Speech Direct Translation

Novel Datasets

To further enhance the performance and accuracy of the Dubble automated dubbing tool, we have acquired a novel dataset by downloading dubbed videos available for free on YouTube. This dataset will serve as an invaluable resource for training and fine-tuning our models. By analyzing a diverse range of professionally dubbed videos in diverse and dramatic genres, our system will learn to better replicate the nuances and subtleties of human dubbing artists. Additionally, this data-driven approach will help identify common challenges and pitfalls in the dubbing process, informing our ongoing efforts to improve the tool's overall effectiveness.

Speech to Speech Model



(a) Overview of Translatotron 2. *(b)* Synthesizer of Translatotron 2.

Figure 1: A Translatotron 2 model that translates Spanish speech into English speech.

Paper: arXiv

Leveraging the principles from the "Translatotron 2" paper, we are developing a speech-to-speech direct translation model. The core idea would be to adapt the Transformer architecture to handle speech data as input and output, bypassing the need for intermediate text representation. Here are the steps to incorporate the paper to Dubble using our novel datasets:

1. Preprocessing: Convert raw speech data described above into pairs of audio recordings, phonemes from transcripts. These features will serve as input for the Transformer model, capturing essential information about the speech signal.

2. Modular Architecture: Adapt the modular architecture of Translatotron 2 to separate translation and voice conversion tasks in Dubble.
3. Phonetic Bottleneck: Capture the phonetic content of the speech while discarding speaker-specific and context specific information. This will facilitate better translation in the dubbed output. Utilize one self-attention mechanism to capture dependencies between source language phonemes and target language phonemes. This allows the model to learn complex patterns and relationships within the speech data for a more accurate translation result.slid
4. Paralinguistic and non-linguistic characteristics: Synthesize the expected phonemes with original phonemes to preserve speaker turns and other characteristics for more natural sounding output.
5. Post Processing: Convert the generated target speech spectrogram back into a raw audio waveform. This can be done using techniques like Griffin-Lim algorithm, WaveNet, or other vocoder technologies.

Social and Global Constraints and Considerations

Speaker Privacy

In the development of the Dubble automated dubbing tool, we have taken into account the potential misuse of the technology, specifically in relation to deepfake abuse. By designing our speech-to-speech direct translation model without encoding speaker identity and employing a speech-to-text system that restricts users from editing transcriptions to anything, we aim to minimize the risk of the tool being exploited to generate malicious or deceptive content. This approach ensures that the system cannot be manipulated to create misleading videos of prominent figures uttering statements they never made. As a result, our design prioritizes ethical considerations and promotes responsible usage, reflecting our commitment to addressing the broader social implications of speech translation and dubbing technologies.

From a technological standpoint, speaker privacy is essential to ensure that user data is not exposed to unauthorized parties. This includes ensuring that data is encrypted when being sent over the network, and that only authorized personnel can access the data. Additionally, the system should be designed to prevent cross-site scripting attacks and other malicious activities. Furthermore, the technology should have robust authentication measures in place to ensure that only legitimate users have access to the system. Finally, the system should have regular security audits to ensure that it is up to date with the latest security measures.

Language Divide

As the world becomes more globally connected and information access increases, the language divide becomes more apparent as well. Without the ability to translate text, audio, or video, information will still render useless. Dubble aims to bridge this critical gap in providing equitable access to users and consumers, regardless of the languages they understand. These tools can be particularly useful in places like education, healthcare, and business by making video information and resources more accessible to people of different languages. Dubble can also help bridge cultural divides by translating applications in entertainment or even video games to help people appreciate and understand different cultures in their native tongue.

Testing and Evaluation

A comprehensive testing plan will be developed to assess the performance and accuracy of the Dubble automated dubbing tool. Evaluation metrics will include audio quality, translation accuracy, lip-syncing accuracy, and overall user satisfaction. Accuracy of transcription and translation is currently done through manual human verification by native speakers of the languages we have implemented, due to the limitation of available translated text scripts. There is currently no automated tool that can verify the accuracy of transcription or translation on the market currently. Continuous improvements will be made based on user feedback and ongoing advancements in the underlying technologies.

V. Self-learning

V-A. Areas of Self-learning

To successfully complete the project, there are a few main components that we had to teach ourselves. These components included:

- Machine learning frameworks, such as TensorFlow and PyTorch, which provide a range of tools and libraries to help us develop and train machine learning models.
- Speech and text processing libraries, such as HuggingFace, Coqui, Whisper, and Google Translate, which can help extract and analyze speech features and text from our data.
- General data processing tools such as Pandas and Scikit-Learn, which can help us clean, format, and transform our extracted data to make it more suitable for training a machine learning model.
- Model evaluation and optimization tools, such as Keras and TensorBoard, which can help us assess the performance of our model and identify areas for improvement.

In addition to these tools, we also taught ourselves more about machine learning algorithms and techniques, such as supervised learning, unsupervised learning, and deep

learning, to better understand how to train and optimize a model. We have also spoken with our advisor, Dr. Callision-Burch, and his Ph.D. student Liam Dugan, on how to work with audio data and natural-language processing.

V-B. Classes and Knowledge

The following classes have contributed to our group's knowledge and ability to complete our project:

- CIS 5200: Machine Learning
- CIS 5450: Big Data Analysis
- CIS 5210: Artificial Intelligence
- CIS 8000: Ph.D. Special Topics
- ESE 3050: Foundations of Data Science
- ESE 4020: Statistics for Data Science

Specific knowledge from these courses, including Python, machine learning principles, general data science experience, natural language processing (NLP) knowledge, and experience using spectrograms, contributed to our group's ability. Knowledge gaps in the NLP space and model selection and architecture for speech and text recognition were supplemented by individual research and conversations and advising by our project advisors.

V-C. Feedback

One of the most valuable pieces of feedback our team has received is to take great care in the data we use for our machine learning model. Our advisors suggested that we start with a language that one of our team members is fluent in so that we can manually check how well the model is performing. This will help us ensure the quality of our model and make sure that it is able to accurately translate and dub between the selected languages.

To address this feedback, we focused on finding high-quality Chinese TV shows that have English dubs, as one of our team members is fluent in both languages. This allowed us to accumulate a large amount of audio data in both languages, while also adhering to copyright laws. We were able to collect over 1,000 hours of data for free, which will be essential for training our model and creating a proof of concept.

Overall, this advice has been incredibly valuable for our team, as it has helped us focus on the quality of our data and ensure that our model is able to accurately translate between the selected languages. We are confident that this will help us create a successful proof-of-concept and pave the way for the future development of our model.

At this time, we haven't had any feedback that we disagree with yet. All of the feedback from our advisor and peers that we have received has been constructive and helpful.

VI. Ethical and Professional Responsibilities

We explored the idea of accents and languages in recreating actors' voices while also translating their dialogue. We wanted to make sure that the content we produced is not discriminatory to any culture while also enhancing the production of the movie itself. As for proprietary data, we have gotten in contact with professors at Penn Law that specialize in IP law to discuss fair use copyright laws pertaining to the data we will be using. We also ensured that data used to train our model was collected within fair use copyright laws. We will also keep our API private, in line with industry standards, to ensure that the deepfake technology is not used in deceptive or harmful ways.

VII. Meetings

Our team has been meeting weekly in person during the CIS 8000 class time on Fridays, which is a graduate-level CIS research course taught by our advisors. The meeting time and location have been working well thus far, as we are able to meet simultaneously as a team and with our advisor and Ph.D. advisor as well. Through the weekly check-ins, we are able to get consistent feedback, advice, and direction in our project. The weekly meetings have been productive thus far and we use the time to split up weekly assignments for each team member, which varies from week to week (e.g. gathering data, testing out a new API, etc.).

Our advisor set up a Slack channel for our project, which we have also been utilizing to communicate and to give team updates in between in-person meetings. Our team also communicates internally over text throughout the week. We also plan to meet with contacts at Penn Law next semester to discuss possible privacy and authorization constraints, if any, that may arise through dubbing television shows and films.

VIII. Reflection of Project Schedule and level of achievement

We came into the first semester of the project with the idea that we would split it into two phases. The first phase would include creating a transcription and translation model and pre-training on existing state-of-the-art models while collecting a large and useful speech-to-speech dataset. The second phase would involve creating the direct translation model using the gathered data while also using our transcription and translation model to serve as reinforcement.

Our milestones for the fall semester can be seen on this chart:

Table 1. Fall semester milestones

Timeline	Action	Team members
(9/22)	Create Project Proposal and Framework of STT and TTS Model	Maya, Sahit, Cathy, Kaitlynn, Vinny
(10/21)	Create STT Model (Figure out how to isolate background noise, how to work with pre-trained models)	Sahit, Cathy
(11/1)	Create TTS Model, Combine with STT Model for Demo (figure out how to preserve voice characteristics of original speaker)	Maya, Sahit, Cathy
(11/27)	Fine Tune Parameters of STT and TTS models for more accurate results	Kaitlynn, Vinny
(12/8)	Introduce Speaker Classification for audio with multiple speakers	Maya, Sahit, Vinny
(12/12)	Finish Collecting Chinese/English Audio Dataset	Kaitlynn, Cathy

We were able to create a successful speech-to-text and text-to-speech model in the Fall, which can take in an input video and create an output video with target language audio. We trained our model with fine-tuned parameters and multiple speakers. In addition to the model, we collected 1000 hours of dub data from Youtube, and plan to use it to train our speech-to-speech model. We were able to find audio and English and Mandarin, for which our group has native speakers, which will be crucial to verify the success of our created dub.

Our spring semester milestones include preprocessing our collected data in a format that is suitable for our model. We worked on figuring out a model architecture for our speech-to-speech direct translation model, as research is still ongoing in this domain. Furthermore, we needed to create a model that will function efficiently on the computer and GPU we provided, since we do not have unlimited resources to work with. Lastly, we wanted to be able to integrate our current model as an underlying checkpoint for our new model to make sure that it learns translation and vocal characteristics separately (which can create a more generalizable model). There are many approaches we can take for this, but the one we are considering currently involves both processes running simultaneously, with a few breaks in the direct translation neural network to check if the output matches that of the transcription and translation model, just to ensure that our direct translation model is operating in the right order.

A condensed list of our Spring Semester milestones can be found in this chart:

Table 2. Spring semester milestones

Timeline	Action	Team members
(1/11)	Unsupervised clustering to differentiate speakers and preserve voice characteristics	Maya, Sahit, Cathy
(1/11)	Look for additional data sources for STS dataset	Sahit, Cathy, Kaitlynn, Vinny
(2/1)	Data pre-processing for STS model input	Maya, Sahit, Cathy
(2/15)	Selection of satisfactory STS model architecture	Kaitlynn, Vinny
(3/20)	Integrating existing model with STS model for robustness	Maya, Sahit, Cathy, Kaitlynn, Vinny
(4/1)	Training/Test Period	Maya, Sahit, Cathy

As for achievement, we were able to accomplish all our goals from the Fall milestones and most of the Spring milestones. We were able to accomplish speaker diarization, model selection, and data pre-processing. and develop an end-to-end dubbing platform. We have yet to integrate the existing model with the STS model, as model selection and pre-processing were more time-intensive than expected, but we hope to continue to work with our advisors on this part post-project.

IX. Discussion of teamwork

IX-A. Team Coordination

Our internal task distribution is done weekly through our weekly check-ins and based on each person's ability and pertinent skills. For example, the CIS group members (Cathy, Sahit, and Maya) have worked more on the model itself, while the ESE and SSE group members (Vinny and Kaitlynn) have worked on the data set collection aspect, as well as overall coordination and tasks unnecessary for the senior design course. Though each member has different roles, we also make sure everyone understands the inner workings and technicalities of our project.

IX-B. Team Challenge

A challenge we encountered was determining how to create and lay the groundwork for a Speech-to-Speech model, which is our ultimate end goal. The main roadblock was the lack of public data available to train such a model. However, after discussion with our advisors and hours of research, we decided to create our own proprietary dataset using publicly available Chinese-English videos (as to not infringe on copyright laws). We then wrote scripts and compiled playlists of videos in this dataset (over 1000 hours) and are now ready to begin feature engineering and training a Speech-to-Speech model next semester.

IX-C. Inter-Departmental Team Contributions

Our project is inter-departmental through the CIS department and ESE department. However, all members have experience and knowledge in data science, machine learning, and experience coding. All members worked on researching the problem and finding possible solutions to automate dubbing. As mentioned above, the CIS group members (Cathy, Sahit, and Maya) worked on the Speech-to-Text (STT) and Text-to-Speech (TTS) models, as well as combining the different APIs and open-source services to create our final output. The ESE/SSE group members (Vinny and Kaitlynn) worked together on laying the groundwork for the Speech-to-Speech model by compiling a dataset, as well as overall project coordination and research into model architecture, legal/privacy constraints, and industry/ethical standards. Though each member has different roles, our team also made sure everyone understands the inner workings and technicalities of our project.

X. Budget and justification

Our project is a software application that will require significant computing power and storage. Initially, we planned to use a shared Amazon Web Services (AWS) account, which was estimated to cost around $1,000 for data processing, storage, and model training on EC2. The breakdown of our initial budget for the project's cost is as follows:

- Storage: we estimate that we will need approximately 1,000 GB of storage to store 20,000 hours of audio data, which will cost around $23 per month. Additionally, we will need to store text data and other miscellaneous data, which will likely be significantly less than the audio data and will likely cost around $30 per month.

- Computing: the bulk of our costs will come from using Amazon Elastic Compute Cloud (EC2) instances. For 20 hours of training on 4 powerful instances (with the specifications specified below), we estimate the cost to be around $60. We are unsure how long we will need to run the instances, but we do not anticipate the total training time to exceed 1,000 hours, which would cost around $750.

While the cost of using AWS is similar to the cost of purchasing and maintaining a dedicated computer, the advantage of AWS is that we can easily scale up and down the number of instances depending on our needs, allowing us to use more powerful and costly instances when necessary, and smaller and cheaper instances when time is not an issue.

However, we received some hesitations from our senior design advisor regarding the use of a Penn-backed AWS account due to the history of students exceeding their budgets. As an alternative, they proposed providing us with a powerful computer with ample storage, which ended up costing around $1,800, with $1,600 for a gaming PC and $200 for an additional 2TB SSD card. Outside of computing costs, we do not anticipate any other expenses for the project.

XI. Discussion and Conclusion

To conclude, our team has successfully created a Speech-to-Text (STT) and Text-to-Speech (TTS) model that can take in an input video and create an output video with target language audio. The team has also collected 1,000 hours of dub data from publicly available sources and plans to use it to train a Speech-to-Speech model next semester.

Challenges we faced include continuing to improve our model and output to account for more natural-sounding speech incorporating voice characteristics. However, we believe many of the challenges with our existing models will be solved with the Speech-to-Speech (STS) model to be implemented next semester. The project's ongoing goals and milestones include using the preprocessed data to train for a speech-to-speech direct translation model. The team is also considering integrating the current model as an underlying checkpoint for the new model to improve its generalizability.

Overall, our team has gained an immense amount of knowledge in the NLP and speech/voice modeling space, thanks to our project advisors and individual research. Our team has also learned how to effectively organize, delegate, and accomplish tasks while balancing our course load, which will be valuable as we continue our project.

XII. Business Analysis (M&T)

XII - A. Executive Summary

Our project aims to automate the dubbing industry using machine learning and deepfake technology to replicate actors' voices in different languages. The goal is to improve the efficiency, cost-effectiveness, and quality of dubbed videos, with a focus first on corporate training and then pivot to the film and television industry and the possibility to expand to other forms of voice and video content.

On the technical side, we have currently developed a Speech-to-Text (STT) and Text-to-Speech (TTS) model to lay the foundations for our technology. To do so, we have acquired a gaming desktop capable of handling the computational complexity and storage constraints necessary to train and run our model. However, as we scale our project, we may pivot to using Amazon Web Services (AWS). Our final prototype aims to output a video file in a target language given an input video file, but we currently need to implement multiple stages of user input and editing to ensure high-quality output. We believe that our Speech-to-Speech (STS) model, paired with a novel training dataset, will solve many of the issues and loss of voice

characteristics that currently occur with the STT and TTS models. Additionally, we are also considering potential regulatory issues regarding the use of deepfake technology and have taken those into consideration when designing our model architecture.

II. Solution

From a user perspective, our solution would include an user uploading a video to our service and selecting a language to translate to. Depending on how technical the concepts discussed in the video are, we offer two versions of our product: basic, and advanced. Both will take the input video and language and generate an output video with audio in the target language. This video can then be exported by the user to any platform of their choosing. There will be opportunities at each step of the process (transcription, translation, video recreation) for user input to ensure the software is performing to the user's standard.

For corporations (in corporate training or the entertainment industry), we would offer a more personalized solution that would involve the corporation sending their catalog or video file. After analyzing the video(s), we would provide a pricing quote based on the complexity of translation. Once a price is agreed upon, we would work internally through the same processes as the individual user perspective, with extra verifications to ensure the quality of the output videos are sufficient. We would then send the output videos back to the corporation.

III. Value Proposition

Currently available dubbing services and software only achieve one part of the audio translation process, whether that be purely transcription, purely translation, or purely pronunciation. Additionally, competitors who are supporting the full value chain either do not offer high quality output audio or require human input during the process. These interruptions take time and cost money, and other than resulting in a usable output, do not help movie studios from a financial perspective. Dubble plans to automate transcription, translation, and pronunciation to the point where all that is needed is an input video, and Dubble will generate an output video in a new language. In order to do so, we are creating a novel dataset of movie and TV show dubs, which will not only be useful for our application but will also continue to fuel the research into speech translation, with the eventual goal of real-time translation in reach. As a result, we see Dubble as having endless growth opportunities in the audio translation space, moving from corporate training, to large-scale entertainment, to short-form and real-time content. Existing technologies do not provide this type of adaptability and scalability.

IV. Stakeholders

In the corporate training market, our key stakeholders are company HR departments and employees, who will be directing and viewing our videos. Currently, HR departments either create new videos for each language or internally translate their existing content to show to international employees. We would also need to work with the production companies of these training videos to make sure that they are willing to have their actors' likeness and voice being duplicated through our deepfake process.

In the traditional film and television dubbing industry, primary stakeholders include movie studios, voice actors/actresses, and dubbing studios. During post-production, studios will usually work with a professional translation company and a dubbing studio to prepare and record a script in a new language. This must be repeated language needed, which can often involve many moving parts. Our technology aims to eliminate the need for dubbing studios, voice actors, and translation companies; directly automating the translation process and overlaying the original actors' voices on the new script. The dubbing studio then works with the movie studios to voice and sound-engineer the original movie to overlay the new dubbed voices into the movie or show.

V. Market Opportunity & Customer Segments

We are working with two main customer segments: individuals and corporations. We would like to offer a lower cost, easily usable service for individuals to use for their own videos, facilitated through our online user interface. Our offerings for corporations are targeted towards corporate training and entertainment, and are tailored to each enterprise specifically.

We plan to initially target corporations through their often mandatory corporate training, which typically occurs through pre-recorded videos. These training modules are often global, especially with larger corporations, and language dubs will help companies provide more engaging and effective content with little investment of their own.

We then hope to also target entertainment studios once our model becomes more nuanced, due to the many specific requirements in both translation and voice encodings that the entertainment industry requires. These would include the movie, television, video game, and social media sectors. These studios spend lots of time and money employing dub studios and voice actors, so we believe our product can alleviate some of these issues. The segment is very concentrated with a few large players, as smaller movie studios may not have the budget or reach to justify dubbing purposes.

VI. Market Size

The global corporate training was valued at $332.93 billion in 2019, with an expectation to grow at a CAGR of 8.0% and reach $487.30 billion by 2030, primarily due to the increase of asynchronous, do-it-yourself modules largely comprising of pre-recorded videos. Our solution fits directly into this growth, and we believe we can impact a significant portion of this market growth through our service.

The global market for films is experiencing remarkable growth due to the increasing demand for films with native language translations. The global film and video market reached a value of nearly $234.9 billion in 2020 and is expected to reach $410.6 billion in 2030. In 2021, the global film dubbing market was valued at US$2.43 billion. The total addressable film and television dubbing market is also expected to grow at a CAGR of 5.60% from 2022-2030 and is projected to reach US$3.61 billion by 2030.

Additionally, the application of artificial intelligence (AI) in film dubbing is expected to create a lucrative opportunity for market expansion over the forecast period, as it is likely to further drive the demand for foreign films across the globe.

VII. Competition

Currently, six players hold 60% of the global automated dubbing market, which include VideoDubber, Straive, AppTek, Papercup, Vidby, and My Dubbing. The large majority of these companies are targeted at short-form content such as news programs, talk shows, documentaries, and lifestyle shows. New entrants into the market are experimenting with the use of AI for voice encoding purposes, including Respeecher, Flawless.ai, and DeepDub; however, these companies and their technologies are still at an early stage and consist of a combination of automation combined with manual voice-engineering work.

Our improvements to existing technology and practices in the field aim to reduce cost and time for movie studios while also improving the quality of dubs for viewers, and provide a more streamlined and accessible alternative to existing players in the market.

VIII. Revenue Estimates

We offer pricing for our services based on whether it is directed at individuals or enterprises.

Through our website, we offer basic and advanced automated options. For our basic service, we offer transcription, translation, and video recreation at $0.45 per minute of video per language of translation. We determined this price based on market rates for transcription and translation services, which ranged anywhere from $0.20 - $0.50 per minute (rev.com, otter.ai, etc). Our advanced service will be $4.50 per minute per language, as it involves human intervention to verify translations. This price was determined based on market rates for human translation and transcription. As we grow, we will consider pricing models which incorporate unlimited subscriptions for this service.

For enterprises, we would work with each company individually to determine the complexity of translation and transcription to deliver particular pricing quotes for each order. We estimate these prices will range from $1000 - $8000 per hour per language depending on the content of the video. We gathered these estimates from current industry dubbing rates and our cost breakdown.

IX. Cost Estimates

We would like to break our costs down into different segments. The first segment includes the continuous development of our transcription and translation model, such that our product continues to get better as we move forward. These costs would include data acquisition, computing costs, and data storage costs. We would also consider costs associated with the selling of our service as well, such as hired translators and verfiers. Lastly, we would need to consider administrative costs, such as salaries, marketing, and other business-related expenses.

To create a strong and successful model, we must first acquire more data from existing sources. While we are already creating a novel dataset, we would estimate to spend $100,000-500,000 on data from movie studios or streaming platforms that can help enhance the product. We would also likely enter partnerships with these companies to offer discounts for using our services in the future. We would likely use Amazon Web Services (AWS) to account for the high computing power and storage needed. We estimate that we will need approximately 1,000 GB of storage to store 20,000 hours of audio data, which will cost around $23 per month, which is subject to change as we scale. Additionally, we will need to store text data and other miscellaneous data, which will likely be significantly less than the audio data and will likely cost around $30 per month for the initial 20,000 hours. The bulk of our computing costs will come from using Amazon Elastic Compute Cloud (EC2) instances. For 20 hours of training on 4 instances, we estimate the cost to be around $60. We are unsure how long we will need to run the instances, but we do not anticipate the total training time to exceed 1,000 hours, which would cost around $750. In total, compute and storage costs would be $53 per month with a fixed cost of $750.

For our online individual service, we also expect cloud computing costs to cap at $0.30 per hour to run inference on each submitted video. Furthermore, costs would stem from hiring human translators to verify our automated translations to ensure high translation accuracy. We estimate that the cost for a translator to verify our translated text will be a maximum of $70 per hour of video.

For our corporate service, we estimate costs for translators and verifiers to be larger as they play more of a role in audio and timing evaluation. We estimate these costs will combine to $200 per hour per language. We will also require a large marketing budget to acquire customers in these industries. We expect this marketing budget to scale to $500,000 per year.

Lasty, we foresee additional administrative costs scaling up to $1,000,000 per year as we continue to grow. These costs would include salaries, business operations, insurance, and taxes among other expenses.

X. Technical Update

We are continuing to work on our Speech-to-Text/Text-to-Speech model, enhancing the voice characteristic aspect to make sure the output audio sounds as close to the original speaker as possible. We are also actively working on accurate translation between two languages and characterize different speakers. We are also continuing work on our novel Speech-to-Speech model, which will hopefully rectify the errors found in the current version of the model, such as preserving intonation and the timing variations of sentences in multiple languages. We also gathered a novel dataset of audio dubs which we will use to train the model, and we are preprocessing this data to be inputted in the model.

XIII. Appendix

This is a demo of EY's data storage and leak corporate training video dubbed into French. This demo incorporates multiple speakers, background noises, non-human actors and lip synching.

This is our MVP demo of Morgan Freeman in the movie *Lucy*. This demo incorporates multiple language functionality.

The following are Figma designs we created for Dubble's user interface. These designs are the foundation of the interface we are building.

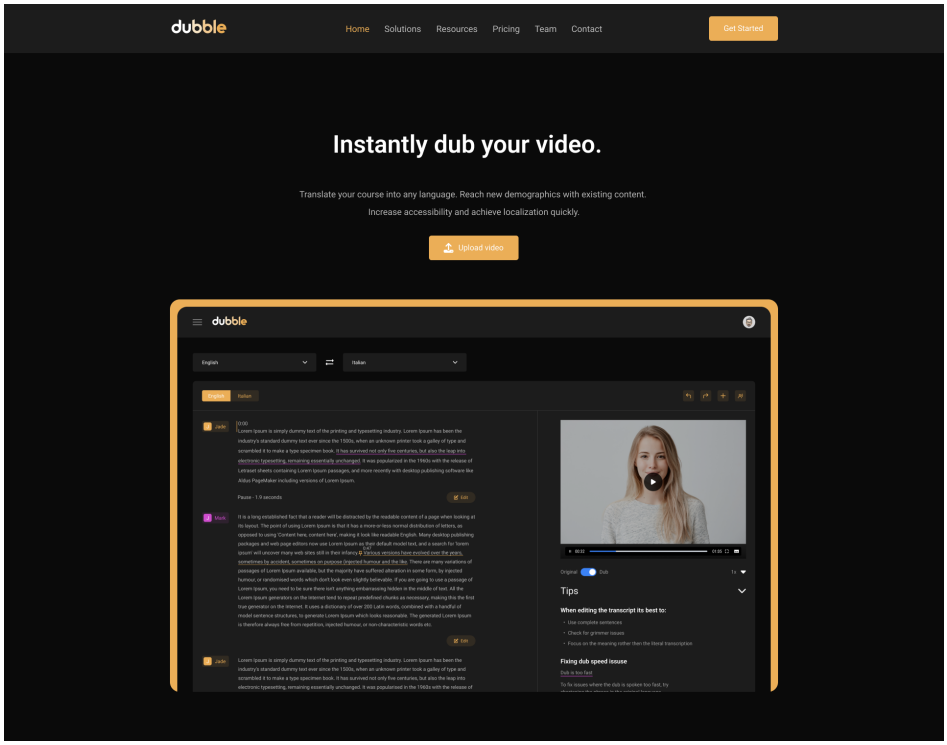


Figure 1. Landing Page

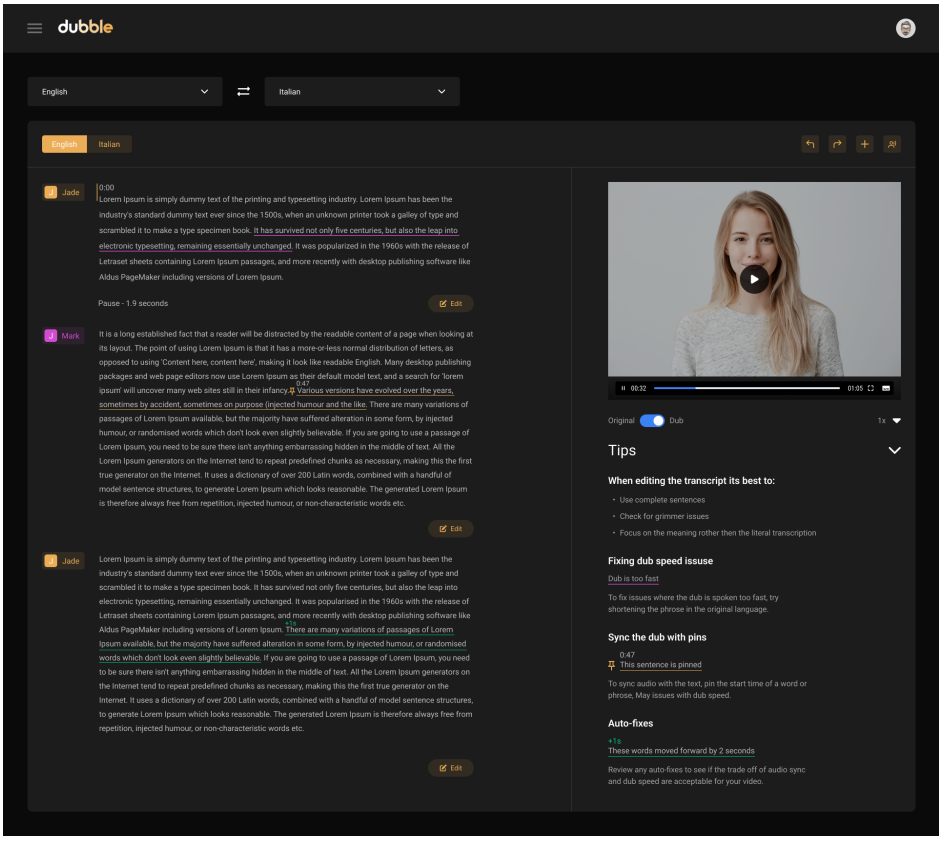


Figure 2. Transcription + Translation Page

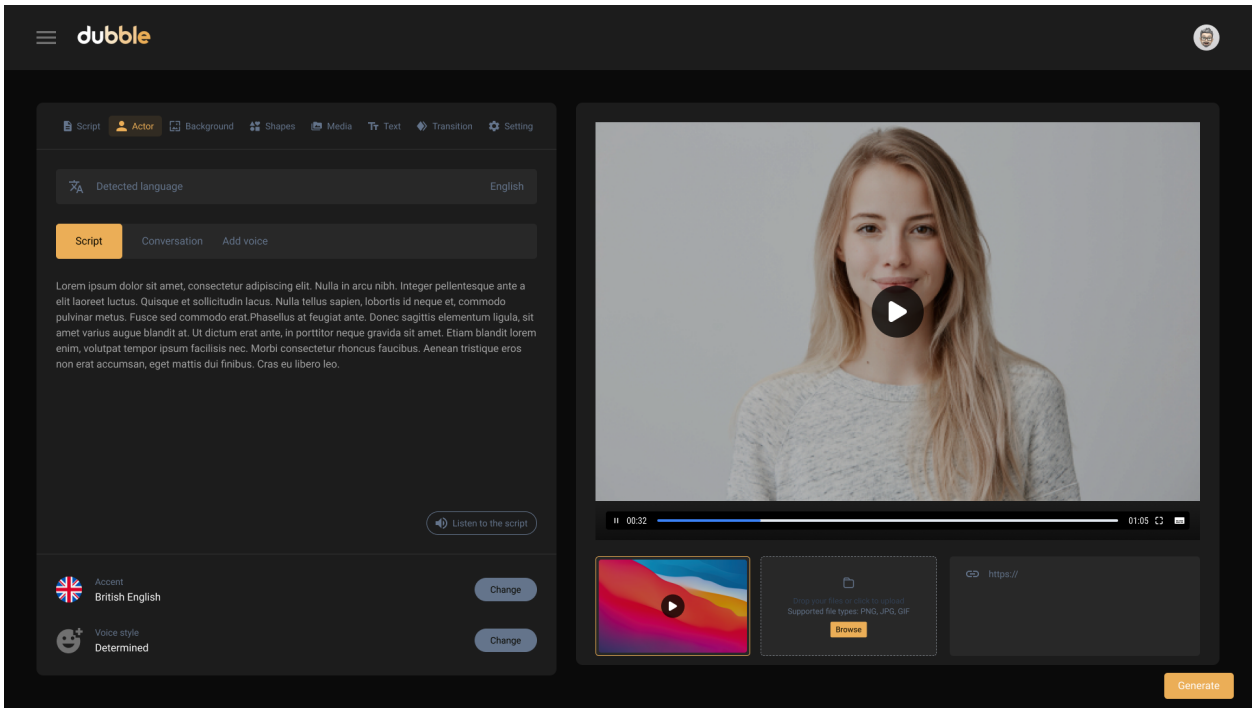


Figure 3. Voice Selection Page

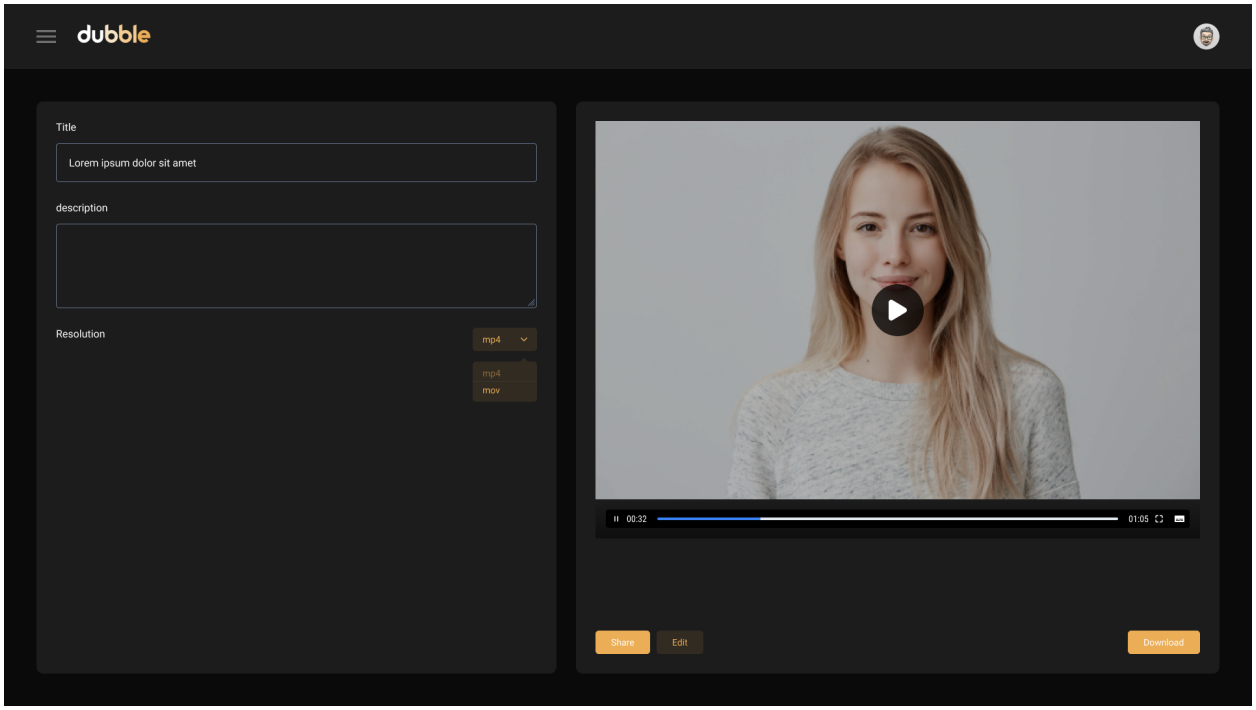


Figure 4. Export Page

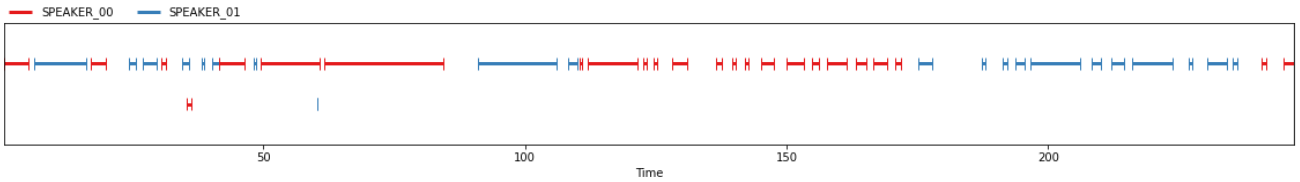


Figure 5. Speaker Diarization