HPA

TECH RETREAT 2024

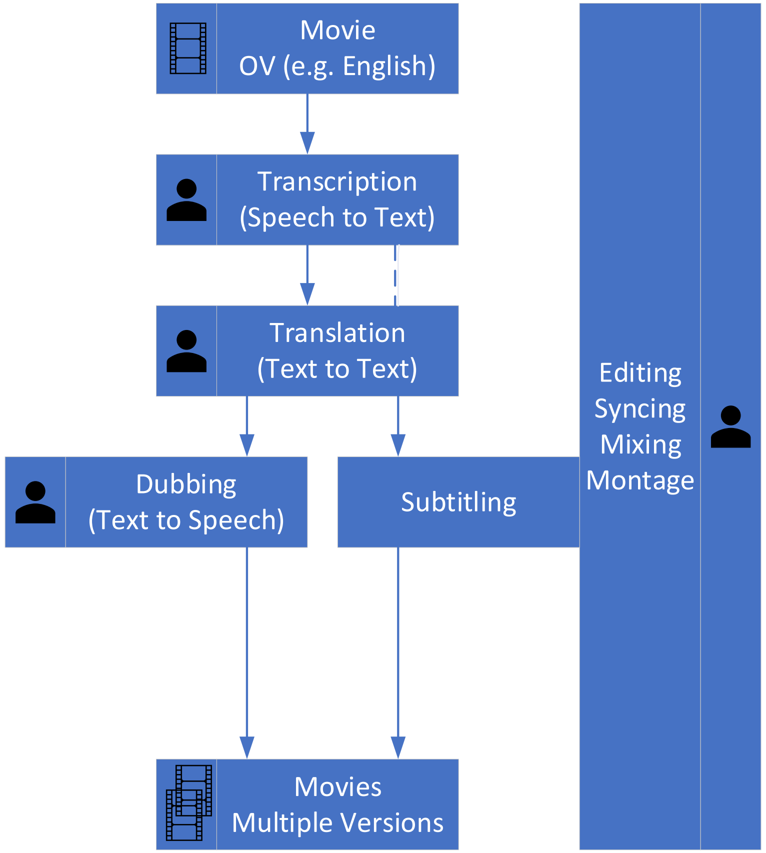
Multi-Language Generation of Movie Content

Siegfried Foessel, Fraunhofer IIS

Fraunhofer
IIS

Fraunhofer Institute for Integrated Circuits IIS

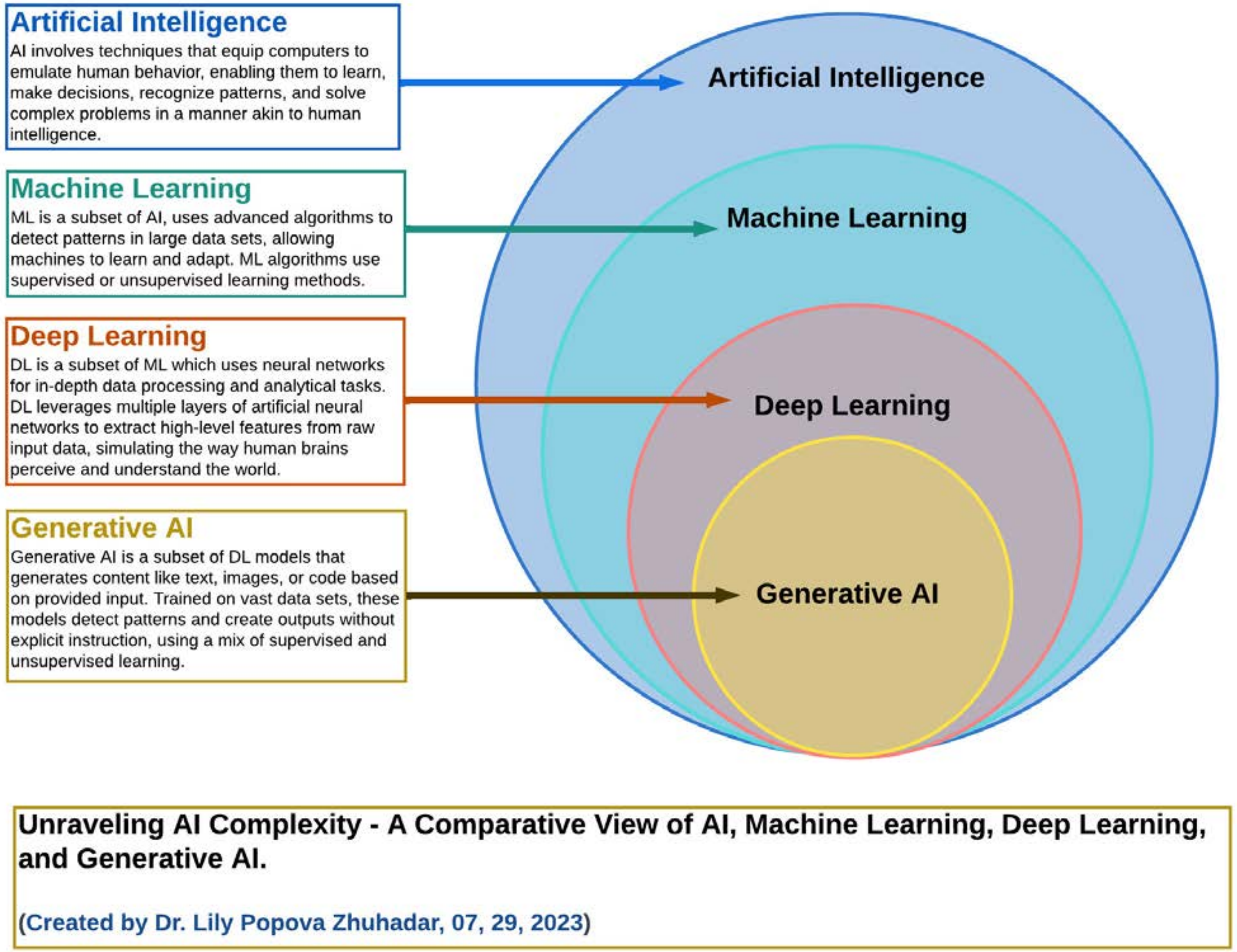
Previous workflow for creating movies in multiple languages



Research project started three years ago

- **Target**
  - Investigate the use of cloud infrastructure in the post production for collaborative work
  - Use of AI services for transcription and translation
  - Define workflows for quick integration of processing tasks

What does Artificial Intelligence mean?



Artificial Intelligence

AI involves techniques that equip computers to emulate human behavior, enabling them to learn, make decisions, recognize patterns, and solve complex problems in a manner akin to human intelligence.

Machine Learning

ML is a subset of AI, uses advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt. ML algorithms use supervised or unsupervised learning methods.

Deep Learning

DL is a subset of ML which uses neural networks for in-depth data processing and analytical tasks. DL leverages multiple layers of artificial neural networks to extract high-level features from raw input data, simulating the way human brains perceive and understand the world.

Generative AI

Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.

Unraveling AI Complexity - A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI.

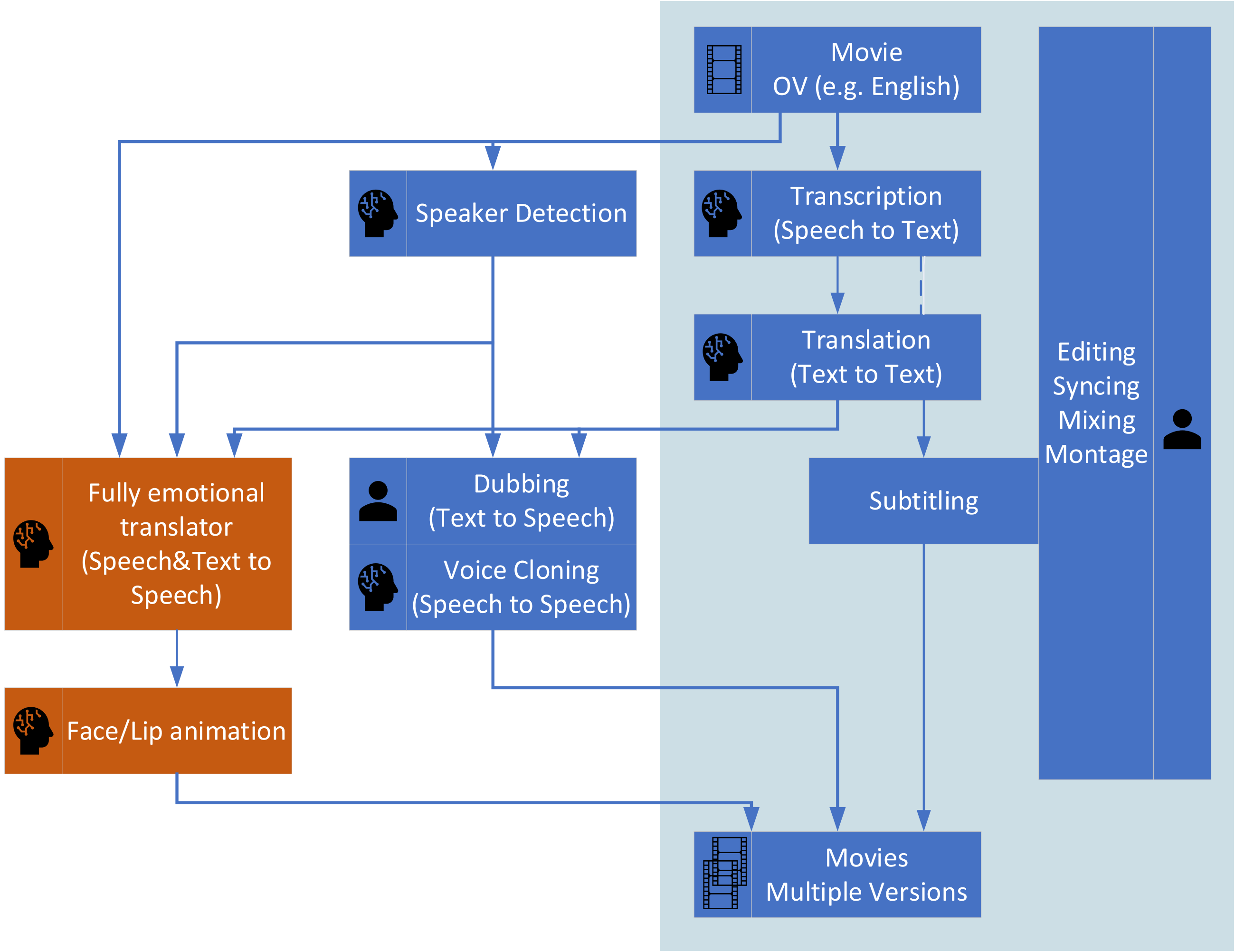
(Created by Dr. Lily Popova Zhuhadar, 07, 29, 2023)

Reference: https://commons.wikimedia.org/wiki/File:Unraveling_AI_Complexity_-_A_Comparative_View_of_AI,_Machine_Learning,_Deep_Learning,_and_Generative_AI.jpg

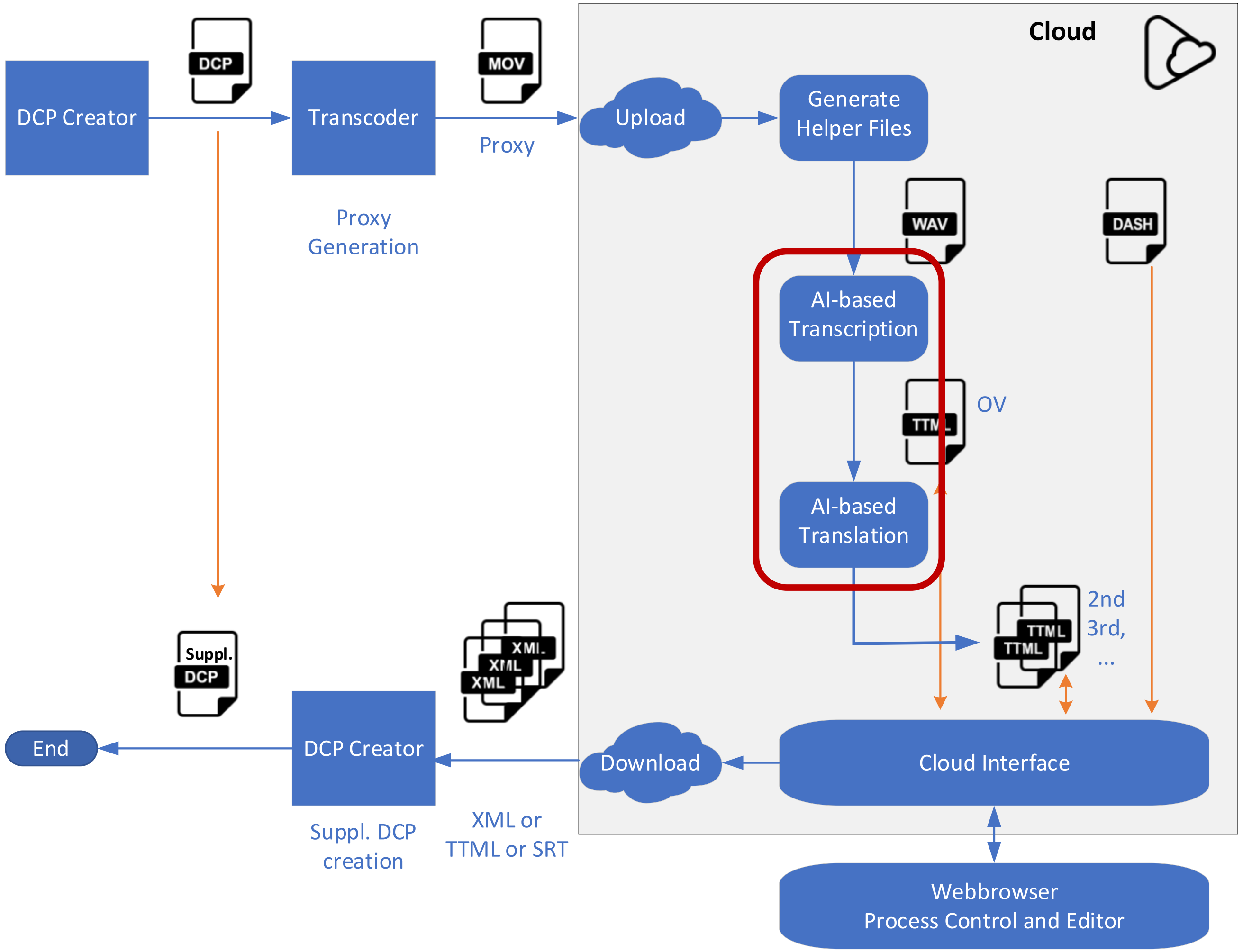
Natural Language Processing NLP

- **LLM Large Language Models**
  - OpenAI GPT-3/GPT-4, Bert, ChatGPT
- **STT Speech To Text / Speech Analysis**
  - Transkription, Subtitling, (sometimes Noise reduction included)
- **TTS Text To Speech / Speech Synthesis**
  - Speech Generation / Multi-Language Speaker
- **STS Speech To Speech**
  - Voice Cloning / Emotion Transfer

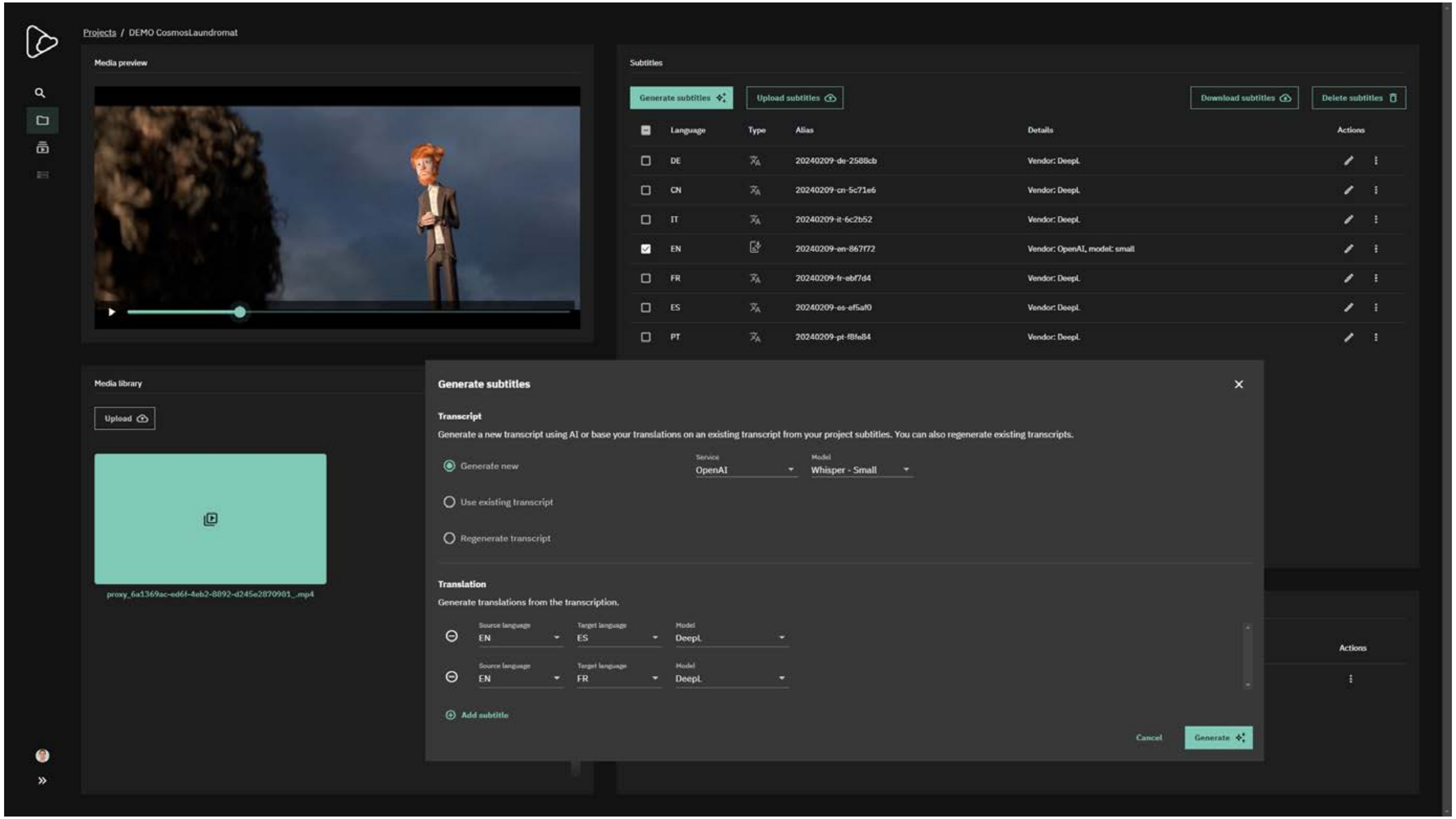
Future workflow for creating movies in multiple languages



Exemplary Workflow for DCP or IMF multi-language support



Exemplary Workflow for DCP or IMF multi-language support



☐	Language	Type	Alias	Details	Actions
☐	DE	文A	20240209-de-2588cb	Vendor: DeepL	✎ ⋮
☐	CN	文A	20240209-cn-5c71e6	Vendor: DeepL	✎ ⋮
☐	IT	文A	20240209-it-6c2b52	Vendor: DeepL	✎ ⋮
☑	EN		20240209-en-867f72	Vendor: OpenAI, model: small	✎ ⋮
☐	FR	文A	20240209-fr-ebf7d4	Vendor: DeepL	✎ ⋮
☐	ES	文A	20240209-es-ef5af0	Vendor: DeepL	✎ ⋮
☐	PT	文A	20240209-pt-f8fe84	Vendor: DeepL	✎ ⋮

Generate subtitles ✕

Transcript

Generate a new transcript using AI or base your translations on an existing transcript from your project subtitles. You can also regenerate existing transcripts.

● Generate new — Service: OpenAI — Model: Whisper - Small

○ Use existing transcript

○ Regenerate transcript

Translation

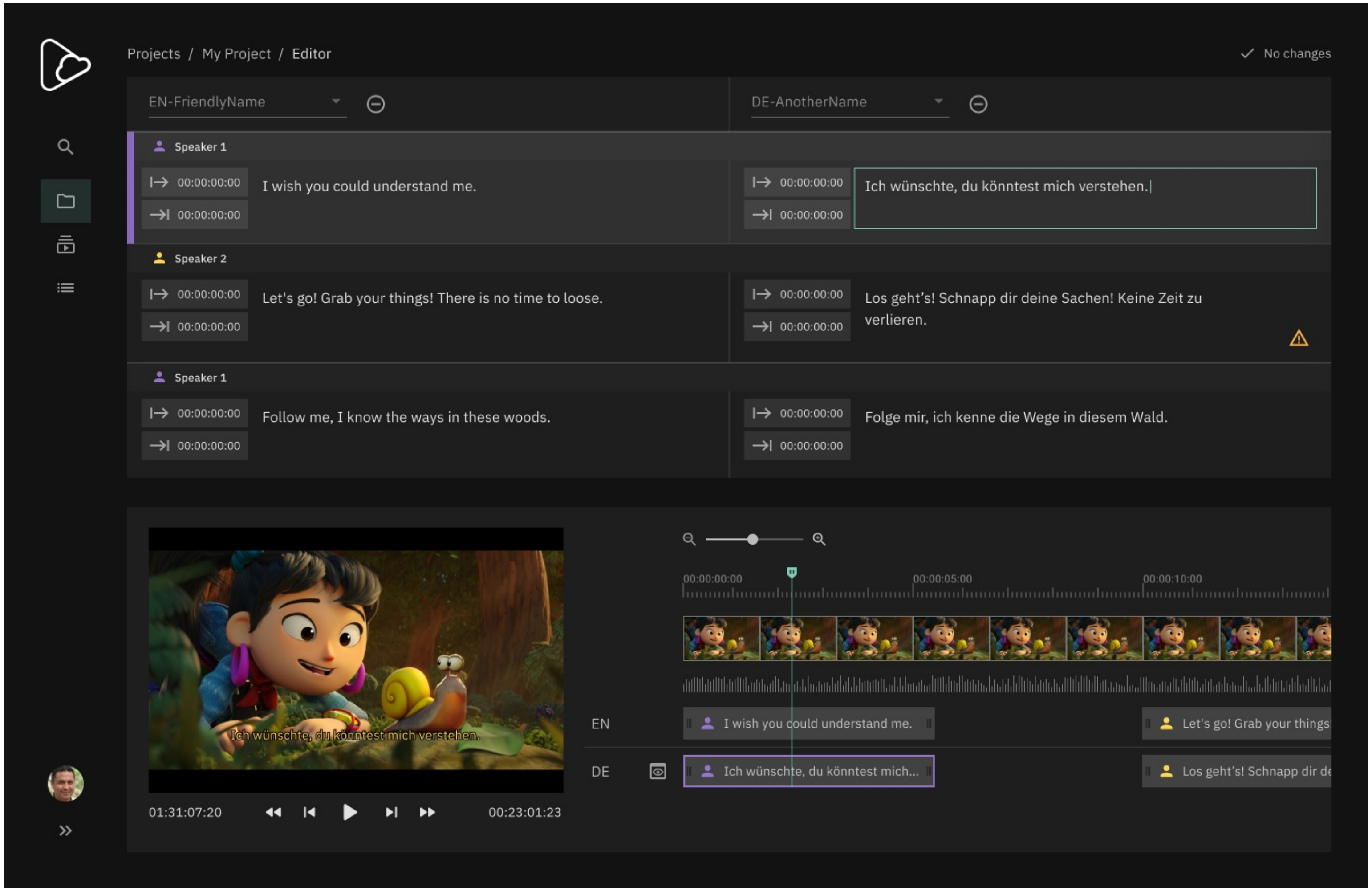
Generate translations from the transcription.

	Source language	Target language	Model
⊖	EN	ES	DeepL
⊖	EN	FR	DeepL

⊕ Add subtitle

Cancel Generate

Exemplary Workflow for DCP or IMF multi-language support



Projects / My Project / Editor

✓ No changes

EN-FriendlyName ⊖

DE-AnotherName ⊖

Speaker 1

↦ 00:00:00:00 →| 00:00:00:00	I wish you could understand me.	↦ 00:00:00:00 →| 00:00:00:00	Ich wünschte, du könntest mich verstehen.

Speaker 2

↦ 00:00:00:00 →| 00:00:00:00	Let's go! Grab your things! There is no time to loose.	↦ 00:00:00:00 →| 00:00:00:00	Los geht's! Schnapp dir deine Sachen! Keine Zeit zu verlieren. ⚠

Speaker 1

↦ 00:00:00:00 →| 00:00:00:00	Follow me, I know the ways in these woods.	↦ 00:00:00:00 →| 00:00:00:00	Folge mir, ich kenne die Wege in diesem Wald.

Ich wünschte, du könntest mich verstehen.

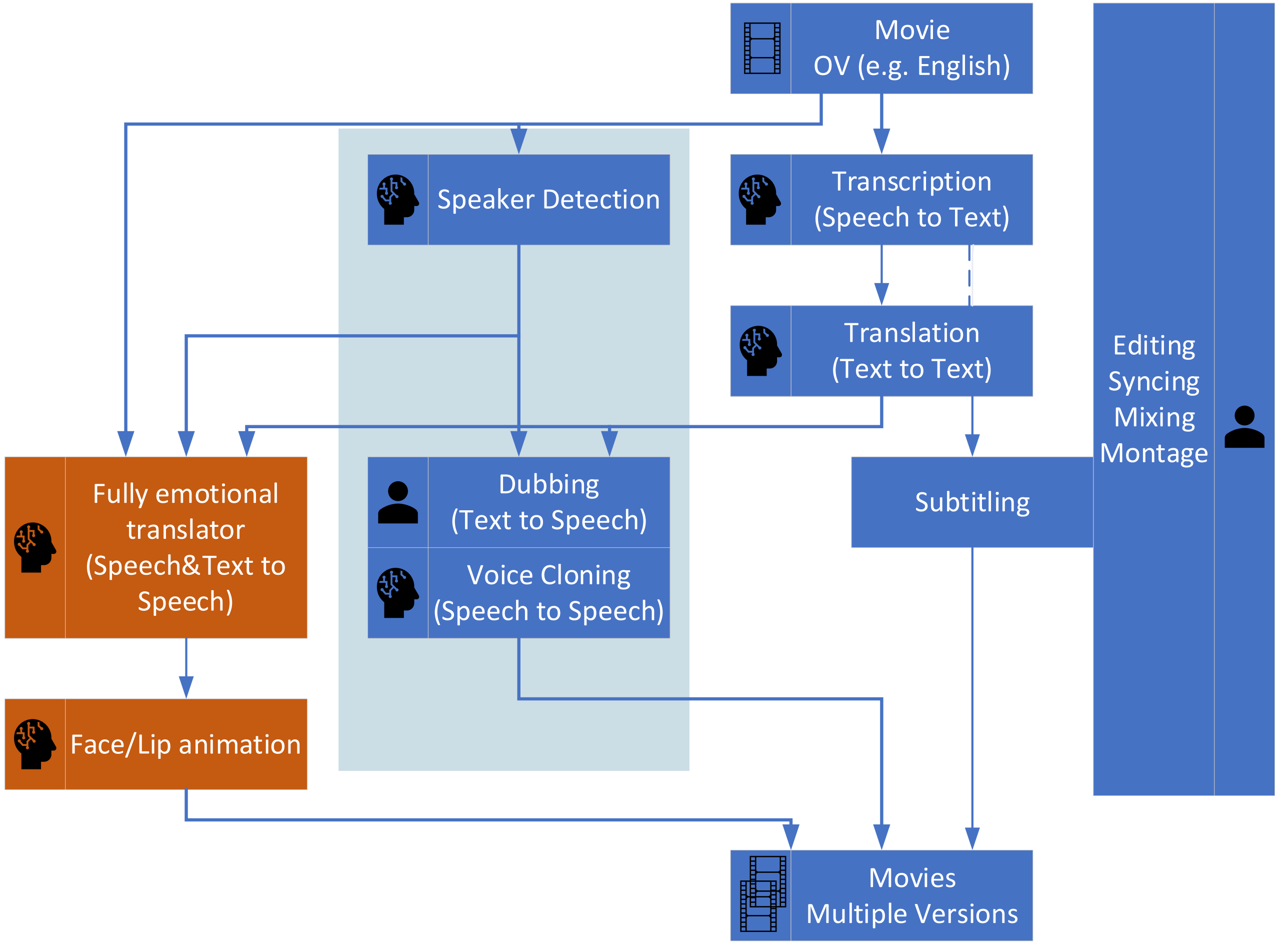
01:31:07:20 ◀◀ ⏮ ▶ ⏭ ▶▶ 00:23:01:23

00:00:00:00 00:00:05:00 00:00:10:00

EN: I wish you could understand me. | Let's go! Grab your things!

DE: Ich wünschte, du könntest mich... | Los geht's! Schnapp dir de

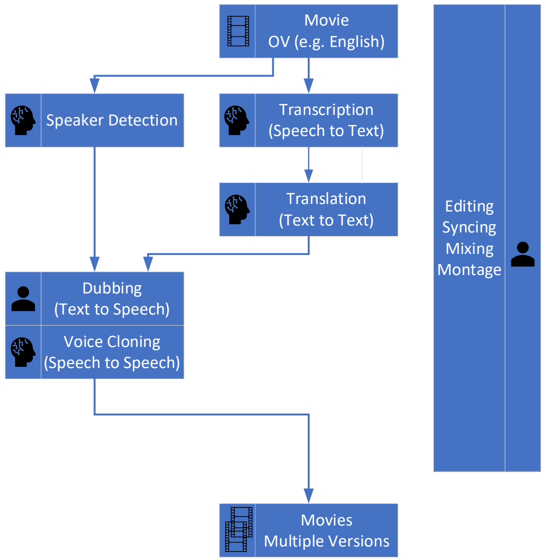
Future workflow for creating movies in multiple languages



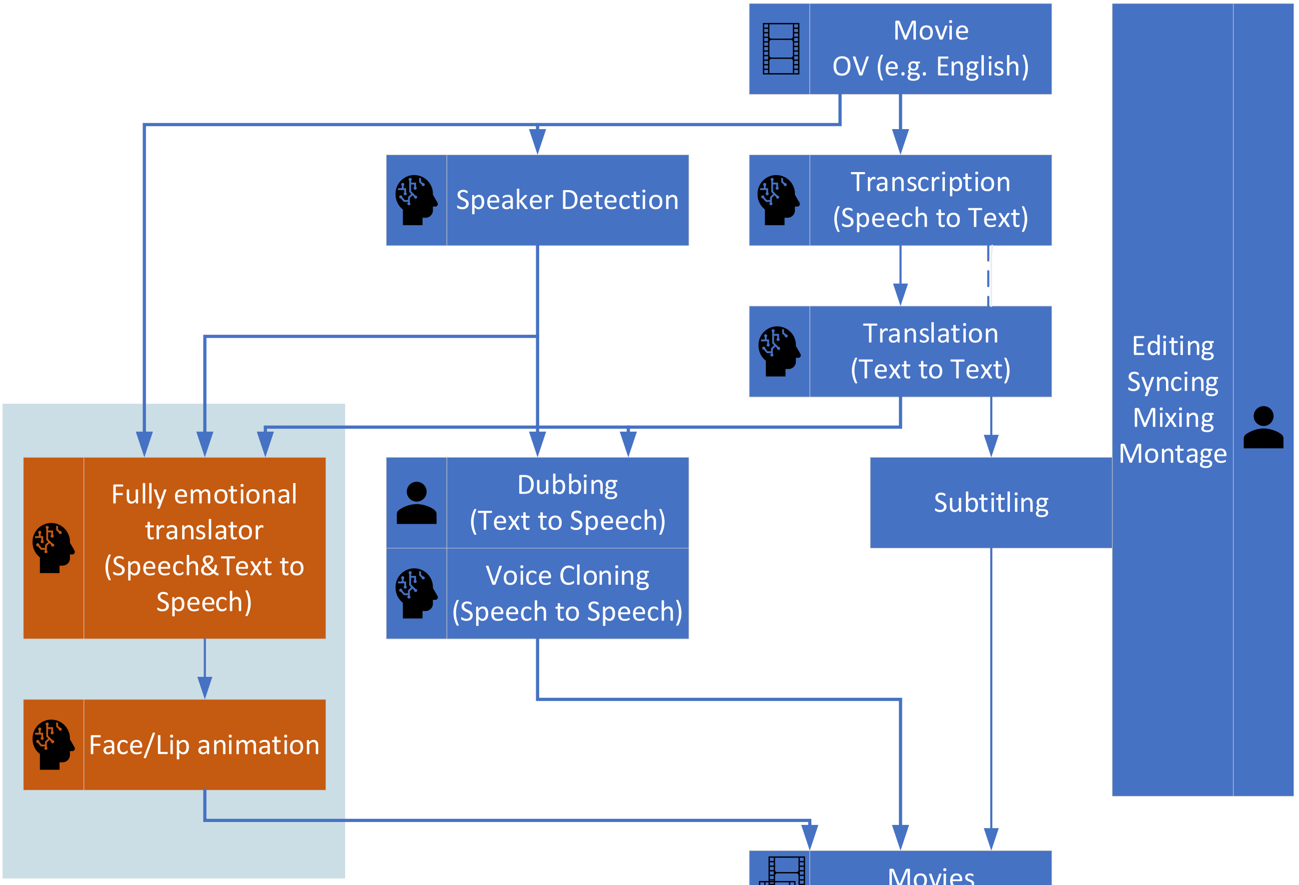
Creating cloned voices for dubbing

- **Speaker Detection**
  - Good speaker detection possible, if model is trained; however strong differences in models exist
- **Vocal Expression**
  - Vocal expression is an art, especially if you switch from one language to another
  - Human voice artists are a good solution
- **Voice Cloning**
  - Voice cloning is possible

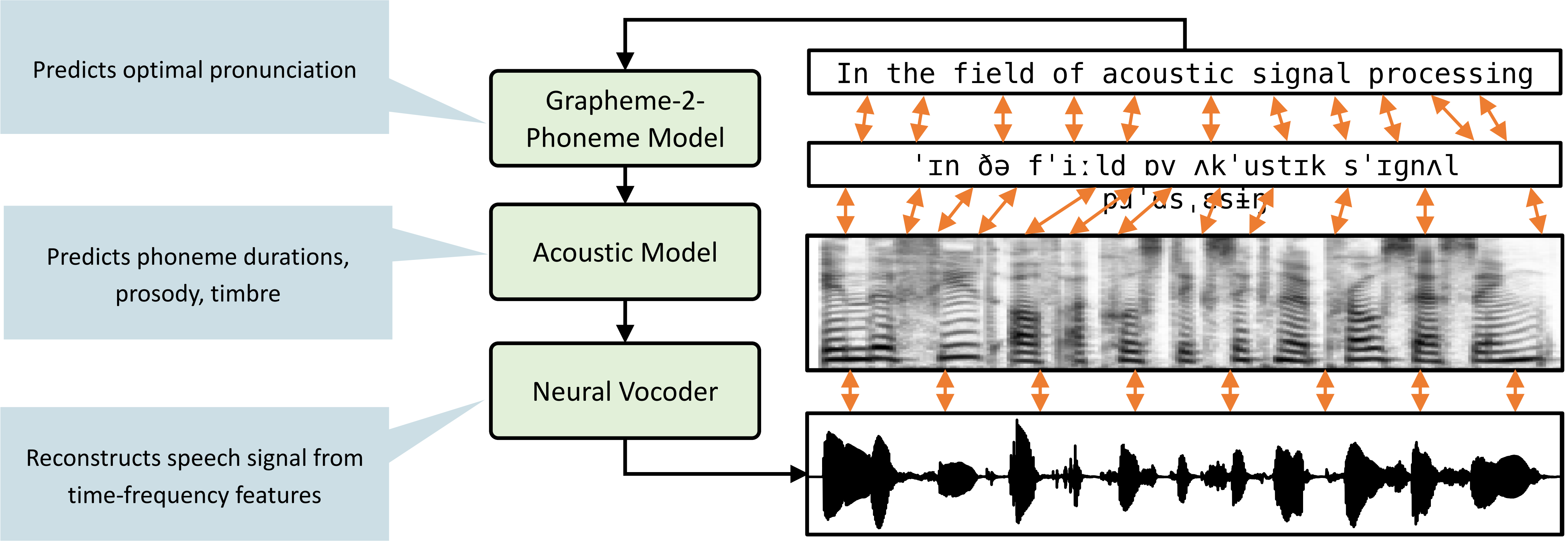
- **Famous Example in Germany**
  - Pumuckl (animated character in a german children live-action TV-series from the 80th)
  - Pumuckls voice was originally spoken by Hans Clarin,
    Part of the success of the series was his specific voice articulation and timbre
  - Recently new episodes were shot. But as Hans Clarin died already, another comedian – Maximilian Schafroth – performed the voice, but the timbre was taken from Hans Clarin. (Technology provided by Respeecher)



Future workflow for creating movies in multiple languages



Fraunhofer Allinga TTS Technology



Predicts optimal pronunciation

Predicts phoneme durations, prosody, timbre

Reconstructs speech signal from time-frequency features

Grapheme-2-Phoneme Model
Acoustic Model
Neural Vocoder
In the field of acoustic signal processing
ˈɪn ðə fˈiːld ɒv ʌkˈustɪk sˈɪgnʌl pɹˈɑsˌɛsɨŋ

Fraunhofer Allinga TTS Technology

Multilingual model fluently speaks German, English and French

Voice Name	**Original Language**	**Training Data**	**German:** In diesem Test können Sie meine Stimme auf Deutsch sprechen hören.	**English:** In this test, you can hear my voice speaking in English.	**French:** Dans ce test, vous pouvez entendre ma voix parler en français.	**Code-Switching:** Sehr berühmte Museen sind das Metropolitan Museum of Art in New York City, das Musée d'art moderne de la Ville de Paris und das Pergamonmuseum in Berlin.
Annie	High German	German + English	🔊	🔊	🔊	🔊
Benjamin	High German	German + English	🔊	🔊	🔊	🔊
Emilie	High German	German	🔊	🔊	🔊	🔊
Oskar	High German	German	🔊	🔊	🔊	🔊
Ethel	British English	English	🔊	🔊	🔊	🔊
George	British English	English	🔊	🔊	🔊	🔊
Pauline	French	French	🔊	🔊	🔊	🔊

Fraunhofer Allinga TTS Technology

Detailed, fine-granular control over prosody

- Examples: duration, pitch, volume
  - <speak>The following text is spoken <prosody **rate="150%"**> faster than normal. </prosody></speak>
  - <speak>The following text will be spoken <prosody **pitch=“-4st"**> four semitones lower </prosody> than normal.</speak>
  - <speak>A text can be spoken <prosody **volume="x-soft"**>very soft or </prosody> <prosody **volume="x-loud"**>very loud.</prosody> </speak>

- Examples: emphasis
  - <speak>She has <emphasis level=“moderate”>**bought**</emphasis> five apples in the store.</speak>
  - <speak>She has bought <emphasis level=“moderate”>**five**</emphasis> apples in the store.</speak>
  - <speak>She has bought five <emphasis level=“moderate”>**apples**</emphasis> in the store.</speak>

Fraunhofer Allinga TTS Technology

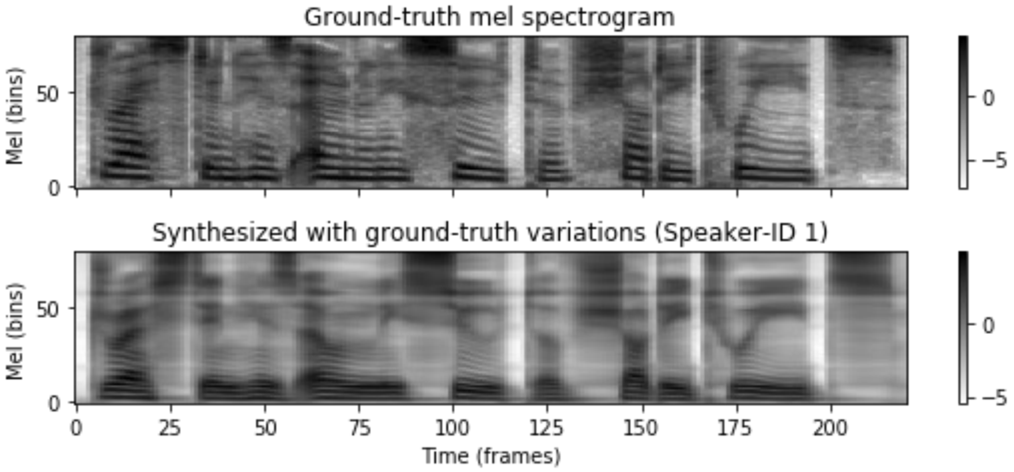
Control over emotional expressions

- **Neutral:** In the quiet office, the constant whir of computer fans and the muted voices of colleagues blended seamlessly, creating an atmosphere that remained comfortably neutral.
- **Happy:** Laughter bubbled up like a melody, and a tapestry of vibrant colors danced in the air, painting a picture of joy and happiness.
- **Angry:** His clenched fists trembled with rage as he confronted the relentless injustice, a firestorm of resentment burning in his eyes, fueled by the betrayal of those he once called allies.
- **Sad:** As the sun dipped below the horizon, casting shadows over the deserted playground, echoes of laughter once vibrant now lingered as silent ghosts in the chilly, empty air.

Fraunhofer Allinga TTS Technology

Voice Conversion (Speech-to-Speech)

- Convert any speech into another speaker‘s voice
- Content and prosody are transferred
- Voice timbre of target speaker is preserved

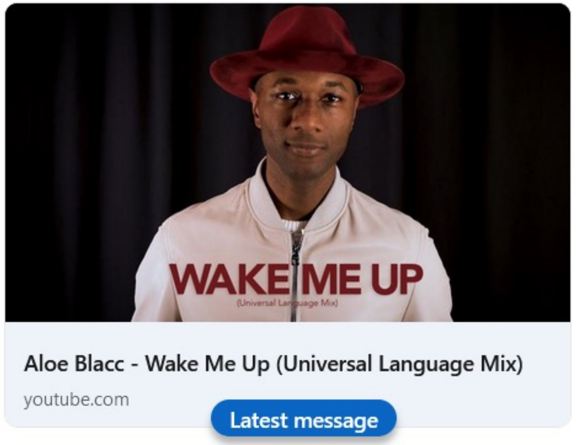


	Original Speaker	Target Speaker
Neutral:	🔊	🔊
Happy:	🔊	🔊
Angry:	🔊	🔊
Sad:	🔊	🔊

Fraunhofer
IIS

Example of speaking/singing in multiple languages

- Aloe Blacc voice and face/lip animation from English to Mandarin and Spanish
  https://www.youtube.com/watch?v=TzofRTcoPsU



Aloe Blacc - Wake Me Up (Universal Language Mix)

youtube.com

- Technology provided by Metaphysic.ai and Respeecher.com

Conclusion/Take aways

- The generation of subtitles in multiple languages is easy to realize today. AI is a good helper for automatic/semiautomatic transcription and translation. Human work is still usefull for quality control and finetuning.
- An original voice can speak in multiple languages. The right vocal expression in a different language is still a challenge. It can be modelled by tags, but the question is, if it is not easier to use an voice artist and clone the original voice to the foreign speaker if wished.
- On mid-term it seems possible that an actor can speak in multiple languages with his original intended expression. Even the lips can be animated to match the spoken foreign words.

Fraunhofer
IIS

Fraunhofer Institute for Integrated Circuits IIS

Contact

- Prof. Dr.-Ing. Siegfried Foessel
- Division AME
- Phone +49 9131 776-5140
- Fax +49 9131 776-5197
- siegfried.foessel@iis.fraunhofer.de

See Demo and more examples at Innovation Zone Booth 501!

HPA

TECH RETREAT 2024