Doctoral Research project proposal

KOUASSI JEAN-CLAUDE, Networks and Telecoms Engineer, Groupe EDHEC-Abidjan, Côte d'Ivoire
Machine Learning Engineer, Udacity

<u>THEME</u>: Automatic audio video translation system with deep learning, and human cognition.

Keywords: natural language processing, speech recognition, machine learning, deep learning

DESCRIPTION

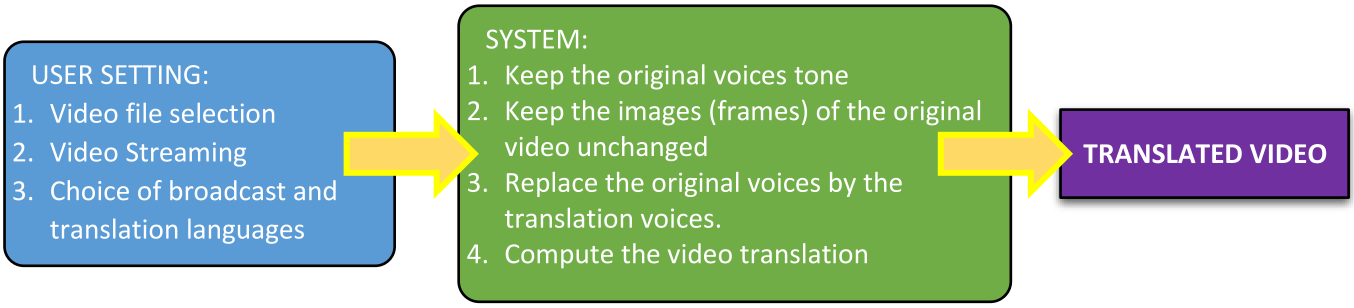
For several decades, with the advent of the Internet, the use of videos for the exchange of information has become widespread. We note its presence in very varied fields: with academicians, professionals and also for general use by the public. But this is not an accidental effect. Indeed, several in-depth studies have shown that the use of video brings more cognitive value to listeners, unlike simple reading.
"Its use as an illustration element, for example, makes it possible to give to learning a dimension closer to reality insofar as a complex situation to be described can be made more comprehensible by the presentation of unanimated images or animated sequences." (De Lièvre et al., 2000) (*see references*).
The universal nature of the information conveyed by this media has led to the invention of subtitling that allows a large number of Internet users of various languages to follow the videos broadcast in languages that are not theirs. Certainly, it allows reaching a wider audience including those who agree to read the subtitles. But it must be recognized that this method is fastidious and painful, especially for long series of videos up to several hours.
So, to increase the chances of assimilation of the user who reads a video broadcast originally in a language that is not his own, why not invent an automated translation system that would translate instantly, or almost, the played video, in his favorite language? This should immediately result in a clear improvement in the assimilation capacity of internet users who use subtitled videos.

Such a system should also be oriented towards human well-being, that is to say, it is desirable to take into account the different possibilities of its interaction with human cognition, with the goal to increase the assimilation capacities of individuals, depending on the application's final output.

RESEARCH OBJECTIVE

The main objective of this research work is to create a system that ensures the automated voice translation of a video into a desired language, other than that of the broadcast. It will be carried out using the latest advances in artificial intelligence, machine learning and deep learning.

Its general working scheme is the following:



USER SETTING:
1. Video file selection
2. Video Streaming
3. Choice of broadcast and translation languages

SYSTEM:
1. Keep the original voices tone
2. Keep the images (frames) of the original video unchanged
3. Replace the original voices by the translation voices.
4. Compute the video translation

TRANSLATED VIDEO

<u>Figure 1</u>: Functional description of the translation system

Thereafter, this project could be used for various sub-objectives to be determined. It can then be very useful for exchanges:

- Between academics
- During online courses (MOOC)
- The broadcasting of general information online
- Etc.

RESEARCH METHODOLOGY (PROPOSAL)

The overall procedure will be to make an initial inventory of the various technologies involved in this field of action. Then will follow the determination of which ones will be used according to the selected type of approach for the integration of the video: immersive (action on the external environment of the video) or intrusive (modification of the file content).

For the first six (06) months: Collection and analysis of the various technologies involved in this field of action.

During one (01) year: Design and restrained functionality tests.

For the following six (06) months: Tests including individuals feedbacks and improvement phases.

The last year: Scientific publications on the performed work and writing of the thesis.

CONTRIBUTION

This project is a contribution in the field of Human-Computer Interactions (HCI). Until now almost all video translation systems are limited to a simple subtitling of their content, or to separate translation video files, one video file per language.
This work will be an opening towards another possibility which proposes to be more effective, firstly in increasing the assimilation capacities of individuals (cognitive value) without too much effort, and on the other hand, in increasing the audience. This method makes the video accessible, a priori, to any other person whatever its language. And in addition, it is a technique that will be continuously improved thereafter.

REFERENCES

1. **Advanced Video Coding Systems**, Wen Gao & Siwei Ma, ISBN 978-3-319-14243-2 (eBook), 2014.
2. **Why use video during training (in french)**
3. **TricorNet: A Hybrid Temporal Convolutional and Recurrent Network for Video Action Segmentation,** Li Ding & Chenliang Xu, 2017
4. **An Improved Video Analysis using Context based Extension of LSH**, Angana Chakraborty & Sanghamitra Bandyopadhyay, 2017
5. **AENet: Learning Deep Audio Features for Video Analysis**, Naoya Takahashi, Michael Gygli, Luc Van Gool, 2017
6. **Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation**, Melvin Johnson, Mike Schuster et al., 2016
7. **Wavenet: a Generative Model for raw Audio**, Aaron van den Oord, Sander Dieleman et al., 2016
8. **Sequence to Sequence – Video to Text**, Subhashini Venugopalan, Marcus Rohrbach et al., 2015
9. **Translating Videos to Natural Language Using Deep Recurrent Neural Networks**, Subhashini Venugopalan, Huijuan Xu et al., 2015
10. **Movie/Script: Alignment and Parsing of Video and Text Transcription**, Timothee Cour, Chris Jordan et al., 2008