Text driven virtual speakers

Vladimir Obradović
Sentience Lab
Casares Costa, Spain
vladimir@sentiencelab.com

Ilija Rajak
School of Mathematics and Computers
Novi Sad, Serbia
skolarajak@gmail.com

Milan Sečujski
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
milan.secujski@uns.ac.rs

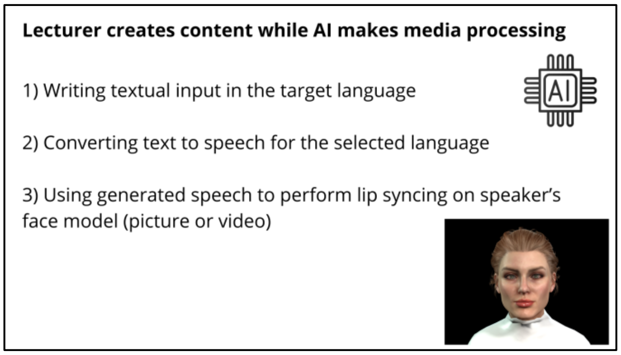
Vlado Delić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
vlado.delic@uns.ac.rs

***Abstract*—Online courses have had exponential growth during COVID-19 pandemic, and video lectures are also important for lifelong learning. However, lecturers experience a number of challenges in creating video lectures, related to both speech recording (microphone and noise; diction, articulation and intonation) and video recording (camera and light; consistency in appearance). It is particularly difficult to modify and update recorded content. The paper presents a solution for these problems based on the application of artificial intelligence in creating virtual speakers based on TTS synthesis and Wav2Lip GAN trained on a custom data set. A pilot project which included the evaluation and testing of the developed system by dozens of teachers will be presented in detail. The use of TTS overcomes the problems in achieving speaker consistency by providing high quality speech in different languages, while the attention and motivation of students is improved by using animated virtual speakers.**

Keywords*—*artificial intelligence, text-to-speech synthesis, virtual speaker, video lectures

I. Introduction

Online and e-Learning have been experiencing an exponential growth trend since 2016 [1]. Due to massive and global migration to online way of work, where collaboration is heavily depending on video conference activities, one form emerged as a primary way of communication and information sharing, which is a shared screen with a slide or a document in the center and a talking head of a speaker in the corner, Fig. 1. This format as a visual extension over a simple slide and a voice-over is very interesting to explore in domain of education. By stimulating both auditory and visual senses, the attention and focus time on the topic is significantly increased [2]. However, lecturers have more challenges in creating video lectures. In ideal case, the content creator would be only focused on educational content design and creation, while everything else could be accomplished by a machine, through automation based on artificial intelligence (AI), as proposed in the paper.



Lecturer creates content while AI makes media processing

1) Writing textual input in the target language

2) Converting text to speech for the selected language

3) Using generated speech to perform lip syncing on speaker's face model (picture or video)

Fig. 1. Video generation process by utilizing AI in the last two steps.

After a brief introduction, the paper is organised as follows. Section II elaborates challenges in creating video lectures and proposes a two-step solution based on artificial intelligence. The proposed solutions for both audio and video synthesis are presented in more detail in Section III, including some legal, moral, ethical considerations. Section IV presents the first results of a pilot project that engaged dozens of teachers in the testing of the developed system. Section V contains conclusions and directions for future work.

II. Video lectures creation challenges and a two-step solution based on AI

A. Educational content creator not fully focused to content creation

While creating educational content, its creator is focused on many other activities besides designing and creating the content itself. This is especially true in case of multimedia content. In case of audio or video content, there are many activities that need to be performed in both pre- and post-processing phase, such as setting up and configuring both hardware and software tools for audio/video capturing, preparing physical environment, as well as video/audio editing, postprocessing audio/video by removing audio noise or color grading, etc.

B. Audio content creation challenges

The major challenge for audio content creation is in maintaining the consistency of intonation, articulation and speaking style long period. For that reason, mistakes are made and the content has to be re-recorded or re-worked. What is particularly demanding, expensive and can be complex, is modifying, updating, editing and maintaining such content over the time. A need to update an audio recording may arise after several months and such a task can be quite demanding.

C. Video content creation challenges

Video content creation shares the same challenges that arise in audio content creation and adds several new ones. There is still need for a consistency, but in this case it is extended to personal appearance, fashion style and dress code, as well as stage set up and lighting conditions. Camera does not care how the presenter feels at the moment, whether he or she has slept well, whether he/she is in a good mood or has had a rough day or night. Due to all those reasons, modifying and updating video content over a longer time period is extremely difficult.

D. Multilingual support challenges

Recording content in multiple languages usually requires multiple speakers who can fluently speak target languages. This can represent a challenge in a situation where a particular teacher or educator should deliver the message. Subtitling the video can be an option, but a much better result can be achieved if he/she is a native speaker of all target languages. Otherwise, AI can help as in the proposed solution based on a text driven virtual speaker.

This research was partly supported by the Science Fund of the Republic of Serbia, #6524560, at the AI project S-ADAPT (2020-2022).

E. Not easy to perform content and semantic search of A/V content

Performing a content or semantic search over audio or video material in its native form can be error-prone and it is rather expensive on large scale. Plain text search represents a much simpler and more straightforward task.

F. Hypothesis: AI can be used for efficient text-to-video conversion, so educational content creator is focused on a content and not on an audio or video media production

If we assume that we start from a content in a textual form and we use text-to-speech synthesis (TTS) to create audio from text, and then we use audio to drive accurate lip movements (lip-syncing, Fig. 2), then we can address all of the abovementioned issues. Since our information is "en-coded" in text, we can then easily translate that text, search, modify and update the content over the time. Complete audio and video content can always be derived from the textual.

III. Proposed solution for video lectures creation

A solution for lecturers enabling them to create their video lectures based on existing slides is proposed in this section. It is based on applications of AI in two steps, as shown in Fig. 2.

A. Speech synthesis (the first step)

The proposed system needs textual input as a content associated to each slide. Quality of synthetized speech depend on TTS available for a given language. Modern TTS systems should be able to produce intelligible and natural-sounding speech as well as speech in multiple voices, styles, and preferably in multiple languages as well. Multilingual and cross-lingual speech synthesis has been developed for the last two decades [3-5].

The pilot project presented in this paper is based on TTS for Serbian developed within a cooperation between the University of Novi Sad and AlfaNum Ltd. The latest versions of that TTS are presented in [6-9], based on deep neural networks (DNN). Apart from a high quality of synthesized speech, AlfaNum TTS has the capability to adapt its voice from a multispeaker space to a dedicated speaker based on a small amount of his/her voice [10]. The capability to simulate each voice enables the creation of any virtual speaker, while changes in the speaking style make speech more interesting. However, it should be noted that these opportunities open some ethical, moral and legal implications.

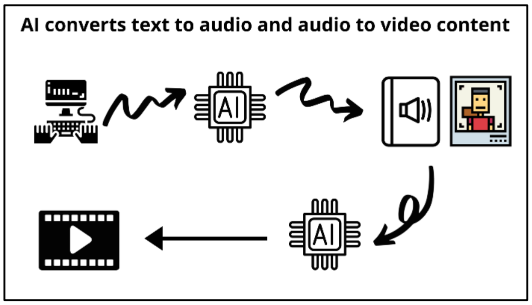


Fig. 2. IAMAI (I am Artificial Intelligence) platform for conversion of text to speech and audio to video.

B. Video synthesis (the second step)

For lip-syncing the Wav2Lip GAN [11] trained on a custom data set is used. The key feature of Wav2Lip GAN lies in being trained with an expert discriminator.

The generator in Fig. 3 contains three blocks: (i) Face encoder, (ii) Audio encoder, and (iii) Face decoder. The Face encoder is a stack of residual convolutional layers that encode a random reference frame, concatenated with a pose-prior (target-face with lower-half masked) along the channel axis. The Audio encoder is also a stack of 2D convolutions to encode the input speech segment (MFCC features or Mel spectrograms), which is then concatenated with the face representation. The Face decoder is also a stack of convolutional layers, along with transpose convolutions for upsampling. The discriminator checks whether the generated frame and the input audio are in sync [11].

The custom data set contained audio/video samples of different speakers, and to avoid bias, speakers of different ages, gender and geographical origins are used in this research and development. The custom data set is a proprietary of SentienceLab company and acquired in a longer period of

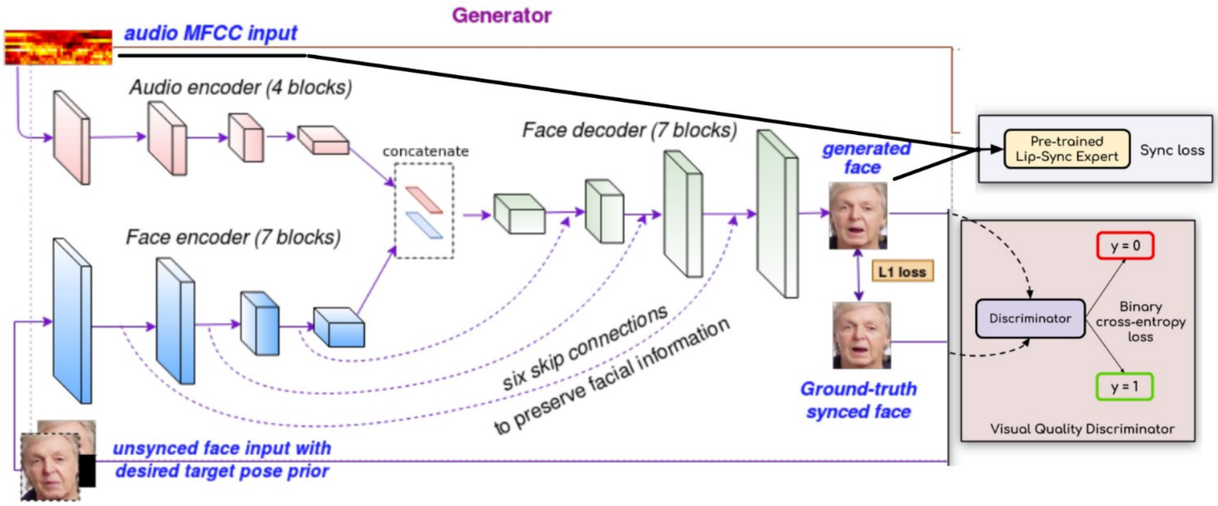


Fig. 3. Wav2Lip GAN model architecture [11-12].

time from social media sources having in mind all privacy and data protection concerns. Some part of data set is collected via crowd sourcing approach by selecting wide variety of subjects. The GAN generates lip movements on a target image of a speaker by using audio Mel spectrograms as an input, independently from the language.

Target speaker can be represented by either a static image or a video. In case of video, each frame (24 per second) is treated as a static photo and lip-syncing is applied to it based on the speech spectrogram. A video template instead of static image enables typical head movements and facial expressions, although they are independent from text and repeat themselves in the same manner.

C. Types of virtual speakers

A virtual speaker can be created from the image or video of any person on which face detection can be applied. Video templates can be used for generating different content. As a video template, a recording of a non-speaking person can be used.

Facial accessories such as glasses or anything that partially covers a person's face landmarks will cause invalid results, and the same holds for faces with beards. In such cases talking lips can sometimes be found on arbitrary positions within the generated video.

D. Legal and ethical considerations

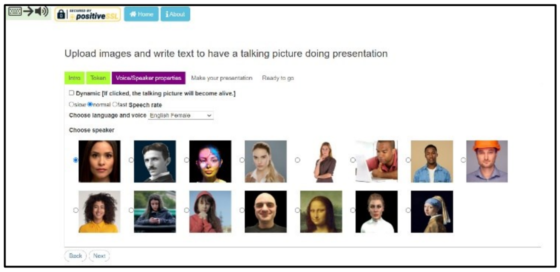
Video material created in the described manner fits into the category of synthetically created material. As such, it is the subject of a number of ethical, and even legal implications. This method could be used for creating fake material pulled out of context. To mitigate those issues, several steps are currently taken. Watermark is added that explicitly indicates AI created content. The access to target speakers and video templates is restricted and speaker templates can be updated only via a strict procedure. In the future, solutions based on block chain such as non-fungible token (NFT) may be used as an inspiration for solving the authenticity problems.

IV. Results

The video generation platform with speech synthesis in the Serbian language has been tested during a period of 8 months by approximately 60 elementary and high school teachers. In the first 3 weeks of testing 101 videos were created. The early users have been "recruited" during a series of 3 webinars organized by Sentience Lab in cooperation with the School Rajak from Novi Sad, which were attended by around 600 participants from over 50 cities in Serbia, Croatia, Bosnia and Herzegovina, Montenegro and North Macedonia. Early users have received a short training related to the IAMAI platform and content creation is described in Figs. 4-6.

IAMAI platform was created by Sentience Lab, company established in the Netherlands but relocated to Spain. Sentience Lab has created IAMAI platform for the needs of its own online school for teaching microskills, I AM AI academy.

IAMAI platform is hosted on internal Sentience Lab (www.sentiencelab.com) infrastructure where inference is done on NVIDIA consumers GPUs like GTX 1070, RTX 2070, RTX 3060 TI and RTX 3090. Also, due to internal inference optimizations and caching, CPU inference is possible with a slight degradation of performance comparing to GPU (two times longer inference duration on CPU). Some part of infrastructure in testing period was deployed to Amazon Web Services (AWS) and particularly Elastic Inference on AWS is used for scaling the resources and balancing the system load.



Upload images and write text to have a talking picture doing presentation

Intro | Token | Voice/Speaker properties | Make your presentation | Ready to go

☐ Dynamic [If clicked, the talking picture will become alive.]

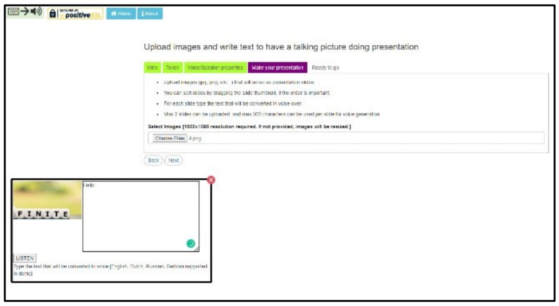
○slow ●normal ○fast Speech rate

Choose language and voice English Female

Choose speaker

Back Next

Fig. 4. Selecting speech language, speaker, speech rate and dynamic or static speaker. In case of dynamic, the head and facial movements will be applied to resulting video, otherwise a static photo will have only animated lip movements.



Upload images and write text to have a talking picture doing presentation

Intro | Token | Voice/Speaker properties | Make your presentation | Ready to go

- Upload images (jpg, png, etc.) that will serve as presentation slides
- You can sort slides by dragging the slide thumbnail, if the order is important
- For each slide type the text that will be converted in voice over
- Max 2 slides can be uploaded and max 300 characters can be used per slide for voice generation

Select Images [1920x1080 resolution required. If not provided, images will be resized.]

Choose Files 4.png

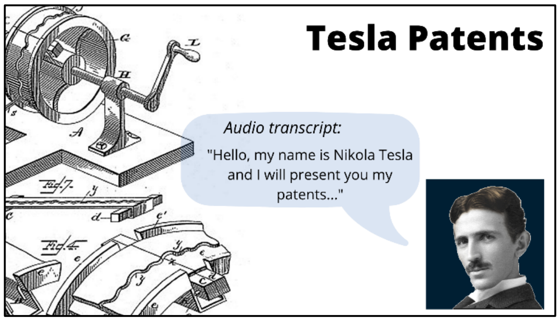
Back Next

Hello

LISTEN

Type the text that will be converted to voice [English, Dutch, Russian, Serbian supported in demo]

Fig. 5. Showing background slide and text input. The written text will be converted to speech and generated speech can be previewed before video creation.



Tesla Patents

Audio transcript:

"Hello, my name is Nikola Tesla and I will present you my patents..."

Fig. 6. Virtual speaker can be either a real-life person or 3D generated avatar or a historical person like a well know scientist, explorer, poet, etc. In principle, the best experience is achieved by selecting a virtual speaker corresponding to a given lecture topic.

Seventeen short video lessons were made as a part of the „Knowledge bites" series, which can be found at https://www.youtube.com/watch?v=glSrrUb_PwI&list=PLFcvPzGOoOAC2YA00yp2GJ6Qg6B1lhg9b. Many of the videos in the series have been seen on YouTube cca 1000 times on average. In order to leave a better impression on their students, teachers have used different historical figures, like scientists, writers and explorers which were „brought to life" and animated, by GAN models.

One video lesson among official public teaching lessons broadcast on national TV uses the video generated by the platform. This can be seen at 3:29 at: https://mojaskola.rtsplaneta.rs/show/2472569/811/os4-

srpski-jezik-51-cas-pepeljuga-narodna-bajka-obrada. Furthermore, two commercial online courses, a total of 4 hours of video material, have been made entirely with IAMAI platform.

A. Feedback from teaching community

Teachers have expressed great satisfaction with achieved results in both speech synthesis and lip-syncing, and given a strong recommendation for the use of the proposed platform. Some of the comments received by professors and teachers follow:

- "I used the material that I prepared for students using IAMAI platform. The reactions were very positive. The children reacted nicely and point out that it is easier for them to remember when they listen and watch than when they read in a classic presentation."
- "One of my wishes was to hold a class without saying a word. Of course, with students motivated to follow it. This makes it possible!"
- "So it means I can also make my photo alive."

B. Feedback from students

School children are quite excited while watching virtual speakers, especially the younger ones. Positive feedback is received from high school students as well. According to elementary school teachers:

- "The children at school were taken aback by the way Vuk Karadžić addressed them, they did not know how that was possible." (https://en.wikipedia.org/wiki/Vuk_Karadžić)
- "My children (4, 6, 7, 8 years old) asked us a hundred times to play the video of Vuk because they really liked it."

Although, people generally prefer human-generated lecture, the proposed system provides advantages in motivation of students when images of well-known persons become "live" lecturers. The auto-generated lectures could be improved with the progress of adopted technical solutions, both the TTS component and the Wav2Lip GAN module. Their main advantage in comparison to human-generated lecture is easy update.

V. Conclusion

A solution for automatization of video lectures creation is presented in the paper. It is based on a two-step application of AI: TTS and virtual speaker. The automatization solves more challenges that lecturers meet in the creation of video lectures. The challenges are elaborated and the proposed solution is explained in detail.

Evaluation of the system through a pilot project has shown the enthusiasm of teachers and motivation of students who see well-known scientists and historical persons as virtual speakers, i.e. "live" lecturers. Stimulation of both auditory and visual senses contribute to better attention and focus on the topic. Automation enables easy update of audio and video content, while animation of selected speakers can motivate students. These conclusions from the pilot project will be quantified in the following research within the project "Application of artificial intelligence to the preparation of video-lectures" dedicated to the development of tools for automated creation of video lectures using AI, both for TTS synthesis and for animating the speaker.

The proposed solution is based on text as input to the system. Language independent conversion of speech to lip movements enables each lecturer to give his/her lecture in any language using suitable TTS. It is enough that someone translate the prepared text in the lecture.

New achievements in TTS development enable speech morphing to any voice. Lecturers should just write the text that they want to be converted to speech which can be very similar to their voice in any language. Such possibilities open some ethical and even legal implications – also briefly elaborated in this paper.

Acknowledgment

This research was supported by AlfaNum Ltd., Novi Sad, Serbia, by providing the TTS engine for the described pilot project. The authors are grateful to AlfaNum engineering team for their knowledgeable and quick response in critical moments. The authors are also grateful to the teachers that tested the developed system and all students that gave their valuable feedback.

References

[1] Johny Wood, "These 3 charts show the global growth in online learning" – https://www.weforum.org/agenda/2022/01/online-learning-courses-reskill-skills-gap/ - World Economic Forum

[2] B. Walker, "Reading vs. Listening – Which is More Effective for Learning and Remembering" – Reading versus Listening - which is better for learning? (transcriptionoutsourcing.net) – Transcription Outsourcing, LLC, Denver

[3] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in Proc. EUROSPEECH 1999, Budapest, Hungary, 1999, pp. 835–838

[4] M. Moberg, K. Pärssinen, and J. Iso Sipilä, "Cross-lingual phoneme mapping for multilingual synthesis systems," in Proc. ICSLP 2004, Jeju Island, Korea, 2004, pp. 1029–1032

[5] Z. Liu and B. Mak, "Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers," arXiv preprint arXiv:1911.11601, 2019. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1911.11601

[6] T.V. Nosek, S.B. Suzić, D.J. Pekar, R.J. Obradović, M.S. Sečujski, and V.D. Delić, "Cross-lingual neural network speech synthesis based on multiple embeddings," The Int. Journal of Interactive Multimedia and Artificial Intelligence - IJIMAI, pp. 110-120, December 2021, DOI: 10.9781/ijimai.2021.11.005

[7] T.V. Nosek, S.B. Suzić, M. Vujović, D.J. Pekar, M.S. Sečujski, and V.D. Delić, "Explicit control of the level of expressiveness in DNN-based speech synthesis by embedding interpolation", Proc. 23rd Int. Conf. Speech and Computer – SPECOM, St. Petersburg, Russia, LNAI 12997, pp. 472-482, Sept. 2021, DOI: 10.1007/978-3-030-87802-3_43

[8] M. Sečujski, D. Pekar, S. Suzić, A. Smirnov, and T. Nosek, "Speaker/style-dependent neural network speech synthesis based on speaker/style embedding", Journal of Universal Computer Science, vol. 26, no. 4, pp. 434–453, 2020

[9] S.B. Suzić, T.V. Delić, D.J. Pekar, V.D. Delić, M.S. Sečujski, "Style transplantation in neural network based speech synthesis," Acta Polytechnica Hungarica, Journal of Applied Sciences, 16(6):171-189, September 2019, DOI: 10.12700/APH.16.6.2019.6.11

[10] D.J. Pekar, A novel method for speaker adaptation in parametric speech synthesis, PhD thesis, University of Novi Sad, Serbia, September 2021

[11] K.R. Prajwal, M. Rudrabha, V.P. Namboodiri, and C.V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild", Proc. of the 28th ACM Int. Conf. on Multimedia, pp. 484-492, Oct. 2020, Association for Computing Machinery, New York, USA

[12] K.R. Prajwal, M. Rudrabha, P. Jerin, J. Abhishek, N. Vinay, and C.V. Jawahar, "Towards Automatic Face-to-Face Translation", Proc. of the 27th ACM Int. Conf. on Multimedia, Oct. 2019, Nice, France, pp. 1428-1436