

Introducing ReVoice: A tool for investigating the learning effects of own-voice narration

Josef Buchner ¹ and Simon Egli¹

Abstract: Research has shown that voice plays an important role in video-based learning, as it can shape learners' cognitive engagement and their experience of instruction. However, previous work has mainly focused on unfamiliar human or synthetic voices, whereas the potential of a learner's own voice has remained largely unexplored. Recent advances in generative artificial intelligence enable a new approach: replacing the original narration of instructional videos with a synthetic version of the learner's own voice. This demo contribution presents the ReVoice tool, a local application for educational voice cloning that supports a controlled study workflow, including voice recording, video transcription, speech synthesis, and video rendering. Initial self-testing results and future experimental and field-based research perspectives are outlined.

Keywords: Voice cloning, instructional video, tool development synthetic media, own-voice narration, instructional technology

Introduction

Instructional videos are among the most widely used media in digital education. They provide flexible, scalable, and reusable access to explanations, demonstrations, and worked examples. Learners can watch them regardless of time and place, pause them, replay segments, and work through content at their own pace. For these reasons, instructional videos have become a central element of many instructional approaches in schooling, higher education, and professional learning [FyHP22; Win24].

Research has shown that learning with videos can be as effective as face-to-face teaching [NGDS21]. However, at the same time, it has become evident that the effectiveness of videos depends strongly on their design. In particular, principles of multimedia learning have proven beneficial for the design of instructional videos. Examples include the combination of visualizations and spoken explanations, the avoidance of unnecessary representations, and the use of social cues such as directly addressing the audience [Buch18; MaFS20; NNSE25].

Voice has also emerged as an important factor in learning with videos. Mayer [Maye20] explains this through the voice principle: when an appealing voice is used in videos, a form of social connection may emerge between the presenter and the learner. This can lead to greater cognitive engagement and, in turn, to improved learning. Previous research

¹ St. Gallen University of Teacher Education, Institute of Digital and Computer Science Education, St. Jakobstrasse 56a, CH-9000, St. Gallen, josef.buchner@phsg.ch,  <https://orcid.org/0000-0001-7637-885X>; simon.egli@phsg.ch

on the role of voice in video-based learning has consistently examined unfamiliar human or synthetic voices. Recent developments in generative artificial intelligence now enable an entirely new approach: learning with videos narrated in one's own voice.

From a theoretical perspective, potential effects of learning with one's own voice can be explained by assumptions from multimedia learning research and self-referential processing. In the Cognitive Theory of Multimedia Learning [Maye20] and in its extension, the Cognitive-Affective-Social Theory of Learning in Digital Environments [SBNS21], social cues are regarded as factors that influence cognitive and motivational processes. If specific design features strengthen the social connection between learning materials and learners, learning may improve because more cognitive resources are activated and engaged. One may therefore assume that one's own voice constitutes such a social cue, more specifically a self-related social cue. Research on self-referential processing has shown that information linked to the self is processed more deeply and remembered better [LWLH24; SyJo97].

Research in this field is still at an early stage, and only a small number of studies have examined learning with one's own voice in comparison with learning with an unfamiliar voice. However, first findings from research in medical education suggest that this approach holds considerable potential for the design of educational media [SXJH26].

In order to investigate the influence of learning through one's own voice, the ReVoice tool was developed, which is presented in the following.

ReVoice for educational voice cloning

The ReVoice tool is a local Electron-based application designed to support a controlled, prototypical study workflow on a Mac. Although it appears as a desktop application, it is internally built on web technologies such as React, HTML, and CSS. The actual data processing is handled by a local Python service.

The tool consists of two separate interfaces:

- Study app: guides participants through login, consent, recording, pre-questionnaire, learning video, post-questionnaire, and completion.
- Admin board: displays study status, allows test resets, manages the completion URL, starts preview runs, and maintains the mixed-transcript mode.

A central design decision is the clean-folder principle. Complete result data are stored only after regular completion of the study procedure. After completion, temporary voice and media data such as raw recordings, normalized recordings, voice presets, synthesized audio segments, snippets, and dubbing artefacts are deleted. In the case of dropout, no complete result file is generated. Only a minimal dropout log is retained, containing technical metadata such as participant ID, time, and reason for termination. Partial content

responses, voice files, and media artefacts are not stored in the results folder if the study is discontinued. Participants interact exclusively with the study app. The study administrator interacts exclusively with the admin board. Both interfaces communicate with the same local Python service, which validates, processes, and stores the data.

For automatic video transcription, the audio track of the selected study video is first extracted. The speaker's voice is then separated from accompanying sound or background elements. Technically, this is achieved using a local source-separation component based on `demucs_mlx`. The result is a speech track that is as isolated as possible. This isolated speech track is then transcribed using `faster-whisper`. `Faster-whisper` is an efficient implementation of the `Whisper` speech-recognition model. In the current backend, the speech-to-text model runs on the CPU and generates segments containing start time, end time, and transcript text.

For the study videos, `vad_filter=True` is used unless an admin transcript is activated. VAD refers to voice activity detection, that is, the detection of those parts of the audio signal that actually contain speech. If the mixed-transcript mode is active, this live detection is skipped during the study run. In that case, the app uses the segments for female and male voices that have previously been prepared and published in the admin board. This reduces runtime and ensures that all subsequent study runs use the same corrected text.

For speech synthesis, `Qwen3-TTS` is used via a local `MLX` runtime. `MLX` is Apple's machine-learning framework for Apple Silicon. For voice cloning, the local setup uses the `Qwen3-TTS-Base` model in its `MLX` variant (`mlx-community/Qwen3-TTS-12Hz-1.7B-Base-8bit`). The recognized or published video transcript is not synthesized as one long block but divided into smaller segments. For each segment, the backend passes the following information to `Qwen3-TTS`:

- the text to be spoken
- the detected or predefined language
- the participant's reference recording
- the reference transcript for that recording

`Qwen3-TTS` then generates `WAV` audio files intended to reproduce the content of the original video in the participant's recorded voice. These generated audio snippets are subsequently normalized, temporally aligned with the segment windows of the video, mixed into a new dubbed audio track, and then muxed with the video.

Here, muxing refers to recombining the video track and the new audio track into a single `MP4` file. `MLX`-based synthesis runs locally on the study computer. The voice is not sent to an external cloud service. After completion or dropout, the reference recording, voice artefacts, and residual media files are deleted in accordance with the cleanup logic.

Initial testing and outlook

To date, the tool is already operational and has undergone several rounds of self-testing by the authors. Swiss German videos, videos featuring Austrian dialects, and Standard German videos were uploaded to the platform and transformed. The initial results for Swiss German and Austrian dialect videos were not fully satisfactory. However, by incorporating transcripts into the workflow, these videos could also be rendered satisfactorily with the authors' voices. For the Standard German videos, the platform produced highly satisfactory results from the outset. The criteria for quality comprised two aspects: the correct pronunciation of words with respect to their intended meaning, and the pronunciation of words without altering their phonetic form.

As a next step, the tool will be further refined and prepared for use in online and laboratory-based studies. Subsequently, randomized experiments will be conducted, either in controlled laboratory settings or via Prolific, to examine whether learning with one's own voice is more effective than learning with an unfamiliar voice. In addition to learning outcomes, assessed through knowledge and transfer measures, cognitive variables such as cognitive load as well as emotional and motivational variables (e.g., enjoyment and interest) will be collected in order to better explain potential effects.

From the outset, potential moderating variables will also be considered. In particular, gender will be considered, as previous studies have yielded mixed findings regarding the relevance of learning with a voice matching one's own gender [LiGS10, ScSZ21]. Additionally, perceived speaker similarity will be considered, as own-voice narration generated by ReVoice represents a synthetic approximation of the learner's actual voice.

Further steps will include examining potential effects in authentic field settings and in combination with established principles from multimedia learning research.

References

- [Buch18] Buchner, J.: How to create Educational Videos: From watching passively to learning actively. In: *Open Online Journal for Research and Education* Bd. 12 (2018), S. 1–10
- [FyHP22] Fyfield, M.; Henderson, M.; Phillips, M.: Improving instructional video design: A systematic review. In: *Australasian Journal of Educational Technology* Bd. 38 (2022), Nr. 3, S. 155–183
- [LiGS10] Linek, S. B.; Gerjets, P.; Scheiter, K.: The speaker/gender effect: does the speaker's gender matter when presenting auditory text in multimedia messages? In: *Instructional Science* Bd. 38 (2010), Nr. 5, S. 503–521
- [LWLH24] Liu, Z. et al.: The effectiveness of self: A meta-analysis of using self-referential encoding techniques in education. In: *British Journal of Educational Psychology* Bd. 94 (2024), Nr. 1, S. 112–137

- [MaFS20] Mayer, R. E.; Fiorella, L.; Stull, A.: Five ways to increase the effectiveness of instructional video. In: Educational Technology Research and Development (2020)
- [Maye20] Mayer, R. E.: Multimedia Learning. Third Edition. Cambridge, UK : Cambridge University Press, 2020 — ISBN 978-1-107-18750-4
- [NGDS21] Noetel, M. et al.: Video Improves Learning in Higher Education: A Systematic Review. In: Review of Educational Research Bd. 91 (2021), Nr. 2, S. 204–236
- [NNSE25] Navarrete, E. et al.: A Closer Look into Recent Video-based Learning Research: A Comprehensive Review of Video Characteristics, Tools, Technologies, and Learning Effectiveness. In: International Journal of Artificial Intelligence in Education Bd. 35 (2025), Nr. 4, S. 1631–1694
- [SBNS21] Schneider, S. et al.: The Cognitive-Affective-Social Theory of Learning in digital Environments (CASTLE). In: Educational Psychology Review (2021)
- [ScSZ21] Schrader, C.; Seufert, T.; Zander, S.: Learning From Instructional Videos: Learner Gender Does Matter; Speaker Gender Does Not. In: Frontiers in Psychology Bd. 12, Frontiers (2021)
- [SXJH26] Sun, Y. et al.: Effectiveness of AI-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial. In: Journal of Medical Internet Research Bd. 28 (2026), S. e81387–e81387
- [SyJo97] Symons, C. S.; Johnson, B. T.: The Self-Reference Effect in Memory: A Meta-Analysis. In: Psychological Bulletin Bd. 121 (1997), Nr. 3, S. 371–394
- [Win24] Winder, G. et al.: Flexible und individualisierte Entwicklung digitaler Kompetenzen von Lehrpersonen am Beispiel der Weiterbildungsplattform «aprendo – digitale Kompetenz». In: Beiträge zur Lehrerinnen- und Lehrerbildung Bd. 42 (2024), Nr. 2, S. 165–183