Geschlechtergerechte Schreibung als Herausforderung für gelungene Textrealisation. *Der Sprachdienst* 64(1–2), 31–45.
Nübling, D. (2018). Und ob das Genus mit dem Sexus. Genus verweist nicht nur auf das Geschlecht, sondern auch auf Geschlechterordnung. *Sprachreport* 34(3), 44–50.
Standpunkt (2020). Standpunkt der Gesellschaft für deutsche Sprache zu einer geschlechtergerechten Sprache. *Der Sprachdienst* 64(1–2), 51–62.
Ulrich, W. (2024). Nicht länger *Säugling*, nur noch *Stilling*? Leiden wir unter einer sprachlichen Diskriminierungsphobie? *Der Sprachdienst* 68(3), 81–100.

David Drevs, M.A.
Internationale Hochschule SDI München

"Linguistic Equality" or "Voluntaristic Acts"?
The Academic Debate Surrounding Gender-Fair Language:
Brief Insights from Five Volumes of *Der Sprachdienst* Journal (2020–2024)

This paper provides some insight into scientific articles published in the German periodical Der Sprachdienst *on the topic of gender-sensitive language policies in Germany from 2020 to 2024. The author's aim is to obtain an (albeit fragmentary) overview of the ongoing discussion and its dynamic over time.*

Key words: *German linguistics, gender-sensitive language, gender-neutral designations, gender star, gender gap, diversity, equality, anti-discrimination, gender-equitable official language.*

Hernandez, Dylan
International University SDI Munich
Supervisor – Drevs, David, M. A.
https://doi.org/10.33989/pnpu.791.c3286

Artificial Intelligence and Dubbing: A Literature Review of the Current and Possible Future Capabilities of AI in Film and Television Dubbing

There are few translation tasks as complex and multi-modal as dubbing. A successful dub requires the coordination of a team of translators, directors, voice actors, and sound engineers exercising attention on an array of interdependent elements of an audiovisual medium in order to transfer these into a different language and cultural context. These elements must be carefully balanced, as they restrict one another while simultaneously contributing to the audience's experience of the finished dub in unique ways. The intricacy of this process renders dubbing a task far beyond the capabilities of simple machine translation (MT).

Nevertheless, as artificial intelligence (AI) continues to garner a significant amount of interest from scientific circles and commercial entities, efforts are being made to adapt the technology for application in dubbing in order to expedite and reduce the cost of what is typically a fairly lengthy and expensive process. Many professionals in language services are well aware of the AI wave that has crashed upon the industry in recent years and may be concerned for the security of their careers, with ever more language-centric, AI-powered solutions threatening to replace human labor. Furthermore, for language service professionals without backgrounds in disciplines such as computer science or machine learning, or who simply have not interfaced with much of the relevant scientific literature, the exact capabilities and

limitations of AI in these contexts are somewhat unclear. Professionals involved in complex translation tasks such as dubbing therefore find themselves threatened by an entity which is not easily understood, increasing anxieties about how AI will impact the near future of dubbing.

This work seeks to address those concerns by summarizing conclusions drawn from an intensive review of contemporary scientific literature on the subject of AI-driven automatic dubbing conducted in 2024. This literature review entailed critical analysis of research methodology and synthesized findings and trends across various works in order to provide a clearer illustration of the current capabilities and limitations of AI in dubbing, and to offer insight into what consequences these may have for human involvement in dubbing.

A total of 22 scientific works which presented technology either possessing high potential to be incorporated into automatic dubbing or technology which was expressly designed for that task were evaluated in the literature review. These technologies included dubbing corpora intended for training end-to-end automatic dubbing models, isochrony-aware MT models, expressive text-to-speech models, visual dubbing models (i.e., AI models designed to edit mouth movements in target video content to better match target audio), and more. Literature for the review was selected primarily on the basis of three criteria: Firstly, only contemporary literature was to be included in the review. Most works included were published no earlier than 2022. Second, the reviewed literature was required to be scientific, meaning that no commercial models were evaluated, and that the works presented data and findings from scientific research rather than simply promoting a proposed automatic dubbing solution. Finally, literature was selected which presented research and proposed AI solutions specifically in the field of automatic dubbing. Literature exploring the capabilities of generalized MT, for example, was excluded from the review.

It was found in the review of the selected works that state-of-the-art automatic dubbing solutions generate dubs which adhere well to constraints of isochrony, and possess some unique capabilities such as cloning source speaker vocal qualities and modifying target video lip movements to match spoken dialogue, but struggle with translation quality under dubbing constraints, as well as with synthesizing vocal performances which are found equal or better in quality compared to human vocal performances.

Test sets and evaluation parameters in the reviewed literature were found to share certain commonalities that indicate a problem of assessing the potential effectiveness of their proposed automatic dubbing solutions in the case of dubbing primarily narrative-driven media such as film, television, animation, and video games. Almost all of the audiovisual content used in test sets tends toward a similar profile. These test sets typically use source video featuring a single active speaker in frame at a time, a mostly, if not completely, static camera and background, and audiovisual content which remains limited to educational, informative, or otherwise non-narrative material, such as interviews, speeches, and lectures (see, for example, Sahipjohn et al., 2024 and Virkar et al., 2022).

While evaluating automatic dubbing solutions on audiovisual content of this profile may provide some insight into their basic functionality, it does not test their

robustness and the upper limits of their capabilities when presented with audiovisual content that does not adhere to a clean and fairly uniform template. Film, television, video games, and animation frequently feature multiple characters on-screen simultaneously, and instances of simultaneous or overlapping speech are not uncommon. It is unclear based on the results of evaluations using such test sets whether any automatic speech recognition systems utilized would still interpret dialogue accurately and be able to assign the correct dialogue to the correct character, for example. Likewise, a completely static camera and background is far from standard in the media mentioned above. Action sequences in films that feature highly dynamic shots and movement from the onscreen actors may prove particularly problematic for visual dubbing solutions which have only been tested on audiovisual content with mostly consistent and fixed camerawork and backgrounds, for instance. Synthesized speech may be not only acceptable to audiences, but completely indiscernible from natural human speech in an automatically dubbed lecture or instructional video. However, whether a high caliber acting performance in a television drama will be as well-received in its automatically dubbed version as it would be in a professionally dubbed version featuring a human voice actor remains widely unexplored.

Subjective evaluations in the reviewed literature also tended to include fairly few participants. No study was found to survey more than 50 participants, with the majority of studies hovering closer to approximately 20 participants. It is questionable whether such sample sizes are sufficient to reflect how the material presented would generally be received by a much larger audience.

Perhaps the most pervasive trend observed in the reviewed literature was the lack of evaluation against a human-produced standard. Of all reviewed works, only 5 (Parada-Cabaleiro et al., 2023, Kim et al., 2019, Swiatkowski et al., 2023a & 2023b, Saboo & Baumann, 2019) were found to include a human-produced dub (or other form of human-produced content) in their evaluation for comparison, and none of them yielded data which reflected superior performance of AI-driven automatic dubbing over human dubbing. Considering the observed lack of automatic dubbing literature that evaluates proposed solutions against a human standard, it seems apparent that automatic dubbing research has yet to produce any verifiable evidence to support the claim that automatic dubbing is a viable replacement for professional human dubbing.

Perhaps the most apparent hindrance to automatic dubbing within the observed literature is the massive amount of training data needed, not only to improve overall quality and performance, but simply to equip AI models with the basic information necessary to carry out the tasks they are designed for. As of the completion of the literature review, the only notable corpora compiled specifically for automatic dubbing were the Heroes corpus (Öktem et al., 2018) and Anim-400K (Cai et al., 2024). The amount of data in the Heroes corpus was deemed by multiple researchers to be “far too little to train an NMT model on,” (Saboo & Baumann, 2019, p. 4) and Anim-400K, being a fairly recent corpus, remains underutilized in automatic dubbing research, though the homogeneity of the corpus with regard to medium, style, and language pair could potentially prove problematic for application in AI-driven

dubbing systems (Anim-400K consists exclusively of Japanese Anime dubbed in Japanese and English). Because of the lack of dedicated corpora and the difficulty in acquiring appropriate training data due to not only the scale of what is needed but also legal protections of intellectual property, researchers in automatic dubbing face a considerable hindrance without an immediately apparent circumvention.

While a full replacement of human dubbing with automatic dubbing therefore seems unlikely in the near future, there may be potential for an integration of AI-driven automatic dubbing technology into human dubbing in order to accelerate and optimize the process. Aside from more obvious use cases, such as MT for generating rough translations for post-editing, certain automatic dubbing technology possesses unique capabilities which do not directly correlate with any of the tasks of the typical human dubbing process. Perhaps most exemplary of this is visual dubbing. Visual dubbing models provide a unique means of improving phonetic synchrony by adjusting the target video to better complement the target audio; in the typical human dubbing process, the target video is not altered whatsoever. The implementation of visual dubbing into the human dubbing process could provide translators more leniency, particularly in instances where phonetic synchrony is of greater importance, such as closeup shots, and improve viewer experience by presenting a more convincing illusion of organic speech.

Whether or not an AI-augmented dubbing process becomes more commonplace in entertainment industries, the results observed in the evaluations of automatic dubbing technology in the examined literature, as well as the general trends of the literature itself, indicate that human involvement in the dubbing process is still, and will likely continue to be, for the foreseeable future, essential to creating audiovisual translations of acceptable quality to most audiences.

This work therefore asserts that AI-driven automatic dubbing is currently not a viable replacement for human dubbing, and likely will not be for the foreseeable future. While augmentation of the dubbing process by AI is a possibility, high-quality dubbing will require a primarily human-driven process.

References

Agarwal, M., et al. (2023). *Findings of the IWSLT 2023 evaluation campaign*. IWSLT. ACL Anthology. https://aclanthology.org/2023.iwslt-1.1v2.pdf

Anastasopoulos, A., et al. (2022). *Findings of the IWSLT 2022 evaluation campaign*. IWSLT. ACL Anthology. https://aclanthology.org/2022.iwslt-1.10v2.pdf

Brannon, W., et al. (2023). Dubbing in practice: A large-scale study of human localization with insights for automatic dubbing. *Transactions of the Association for Computational Linguistics, 11*, 419–435. https://doi.org/10.1162/tacl_a_00551

Cai, K., et al. (2024). *ANIM-400K: A large-scale dataset for automated end-to-end dubbing of video*. arXiv. http://arxiv.org/abs/2401.05314

Chronopoulou, A., et al. (2023). *Jointly optimizing translations and speech timing to improve isochrony in automatic dubbing*. arXiv. http://arxiv.org/abs/2302.12979

Cong, G., et al. (2023). Learning to dub movies via hierarchical prosody models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14687–14697. https://doi.org/10.1109/CVPR52729.2023.01411

Cong, G., et al. (2024). *StyleDubber: Towards multi-scale style learning for movie dubbing*. arXiv. http://arxiv.org/abs/2402.12636

Guan, J., et al. (2023). *StyleSync: High-fidelity generalized and personalized lip sync in style-based generator*. arXiv. http://arxiv.org/abs/2305.05445
Kim, H., et al. (2019). Neural style-preserving visual dubbing. *ACM Transactions on Graphics, 38*(6), 1–13. https://doi.org/10.1145/3355089.3356500
Lakew, S. M., et al. (2022). *Isometric MT: Neural machine translation for automatic dubbing*. arXiv. http://arxiv.org/abs/2112.08682
Li, J., et al. (2024). *Joint multi-scale cross-lingual speaking style transfer with bidirectional attention mechanism for automatic dubbing*. arXiv. http://arxiv.org/abs/2305.05203
Li, Y., et al. (2023). *Zero-shot emotion transfer for cross-lingual speech synthesis*. arXiv. http://arxiv.org/abs/2310.03963
Öktem, A., et al. (2018). Bilingual prosodic dataset compilation for spoken language translation. *IberSPEECH 2018*, 20–24. https://doi.org/10.21437/IberSPEECH.2018-5
Pal, P., et al. (2023). *Improving isochronous machine translation with target factors and auxiliary counters*. arXiv. http://arxiv.org/abs/2305.13204
Parada-Cabaleiro, E., et al. (2023). Perception and classification of emotions in nonsense speech: Humans versus machines. *PLOS ONE, 18*(1), e0281079. https://doi.org/10.1371/journal.pone.0281079
Rao, Z., et al. (2023). Length-aware NMT and adaptive duration for automatic dubbing. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, 138–143. https://doi.org/10.18653/v1/2023.iwslt-1.9
Saboo, A., & Baumann, T. (2019). Integration of dubbing constraints into machine translation. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 94–101. https://doi.org/10.18653/v1/W19-5210
Sahipjohn, N., et al. (2024). *DubWise: Video-guided speech duration control in multimodal LLM-based text-to-speech for dubbing*. arXiv. http://arxiv.org/abs/2406.08802
Saunders, J., & Namboodiri, V. (2024). *Dubbing for everyone: Data-efficient visual dubbing using neural rendering priors*. arXiv. http://arxiv.org/abs/2401.06126
Swiatkowski, J., et al. (2023a). Cross-lingual prosody transfer for expressive machine dubbing. *INTERSPEECH 2023*, 4838–4842. https://doi.org/10.21437/Interspeech.2023-437
Swiatkowski, J., et al. (2023b). Expressive machine dubbing through phrase-level cross-lingual prosody transfer. *INTERSPEECH 2023*, 5546–5550. https://doi.org/10.21437/Interspeech.2023-441
Tam, D., et al. (2022). *Isochrony-aware neural machine translation for automatic dubbing*. arXiv. http://arxiv.org/abs/2112.08548
Virkar, Y., et al. (2021). Improvements to prosodic alignment for automatic dubbing. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7543–7574. https://doi.org/10.1109/ICASSP39728.2021.9414966
Virkar, Y., et al. (2022). *Prosodic alignment for off-screen automatic dubbing*. arXiv. http://arxiv.org/abs/2204.02530
Wu, Y., et al. (2023). VideoDubber: Machine translation with speech-aware length control for video dubbing. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(11), 13772–13779. https://doi.org/10.1609/aaai.v37i11.26613

Dylan Hernandez
Sprachen und Dolmetscher Institut München

Artificial Intelligence and Dubbing: A Literature Review of the Current and Possible Future Capabilities of AI in Film and Television Dubbing

This article summarizes the findings of a literature review conducted in 2024 around the subject of AI in the dubbing of audiovisual media. It was concluded that AI-driven automatic dubbing is not currently a viable replacement for a human-driven dubbing process.

Key words: *artificial intelligence, dubbing, literature review, audiovisual translation, media translation, machine translation.*