Dubbing in Practice: A Large Scale Study of Human Localization With Insights for Automatic Dubbing

William Brannon [1] Yogesh Virkar [2] Brian Thompson [2]

[1]Massachusetts Institute of Technology (work performed during an internship at Amazon) [2]AWS AI Labs

Transactions of the ACL (2023) 11: 419-435

Motivation

How do humans dub video content between languages? Dubbers face many constraints, but they can't satisfy all of them. How do they trade them off?

Qualitative work has theorized these questions [1], [2], and ML work has built automatic dubbing systems [3]. Both make important assumptions which have not been checked by a large-scale empirical study like ours.

Answers to these questions can inform qualitative study and provide direction for ML research on automatic dubbing.

Data Sources

Very large dataset: Every Amazon Studios show (with available scripts) on Prime Video at year-end 2021. 674 episodes; 54 shows; 319.5 hours.

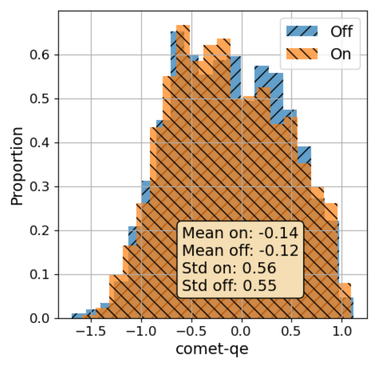
Force-aligned to transcripts and **semantically aligned** between English source and dub. Final data: same content, different languages.

Extensively filtered for quality: Drop non-English content, poor audio quality, crosstalk, incorrect alignments...

Onscreen/offscreen annotations from original scripts: When can we see actors' mouths and mouth movements?

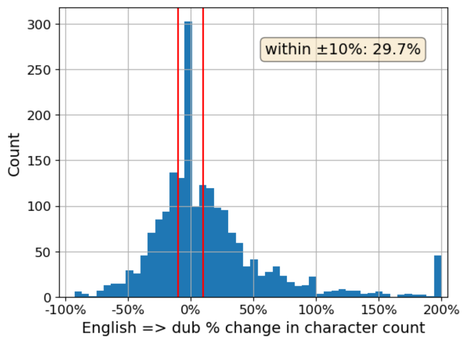
Translation Quality

- **Question**: Do adequacy / fluency suffer for other constraints? Specifically, are automatic MT metrics worse onscreen than off? Onscreen is more constraining.
- **Answer**: No measurable worsening of translation quality!



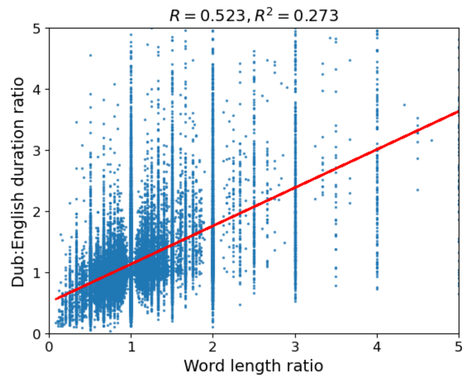
Isometry

- **Question**: Are original and dub texts about equally long? Do human dubs follow prior ML work's $\pm 10\%$ length threshold? [4]
- **Answer**: No! Most human dubs are not isometric.



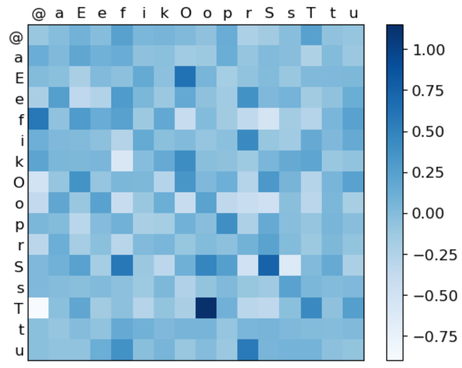
Naturalness (Speaking Rate)

- **Question**: Is speech naturalness reduced to hit other constraints? Specifically, does dub content getting longer lead to faster speaking rate or longer speech?
- **Answer**: Longer speech! Dubbers would rather break timing constraints than vary speaking rates.



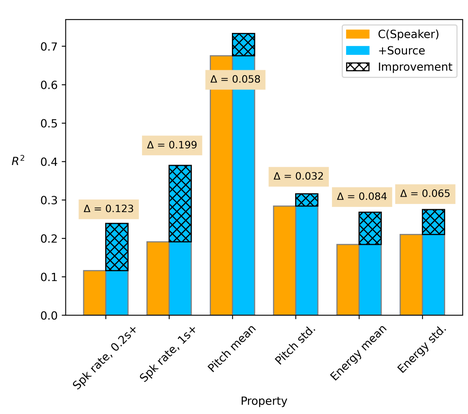
Lip Sync

- **Question**: Does dub speech better align with mouth movements when onscreen (actors' mouths visible)?
- **Answer**: Yes, but not by much. There's little pattern visible in English / dub viseme (mouth movement class) cooccurrence plot.



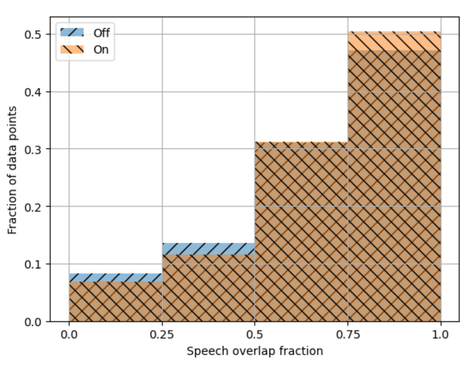
Nonverbal Influence

- **Question**: Does source speech influence the dub nonverbally (within dialogue lines)?
- **Answer**: Yes! Source audio is highly predictive of speaking rate and proxies for emotionality (even controlling for speaker identity).



Isochrony

- **Question**: Are original timing constraints respected? Specifically, does source/dub speech timing match up more onscreen than off? Onscreen is more constraining.
- **Answer**: Less than expected. Isochrony is strong; response to onscreen constraint is not.



Conclusions

Translation quality and **speech naturalness** are paramount!

Major **nonverbal influence** of source audio on dub audio.

Automatic dubbing should **focus on end-to-end systems** + incorporate audio/video, not just text, from the source content.

Isometric MT is **not a useful technique** for automatic dubbing.

Video

TACL Paper

References

[1] F. Chaume, *Audiovisual Translation: Dubbing*, 1st ed. St. Jerome, 2012, ISBN: 978-1-905763-91-7.

[2] G. S. Miggiani, *Dialogue Writing for Dubbing: An Insider's Perspective*, 1st ed. Palgrave, 2019. DOI: 10/h9s2.

[3] M. Federico, R. Enyedi, R. Barra-Chicote, *et al.*, "From Speech-to-Speech Translation to Automatic Dubbing," *ICSLT*, 2020. DOI: 10/gp7jgr.

[4] S. M. Lakew, Y. Virkar, P. Mathur, and M. Federico, "ISOMETRIC MT: Neural Machine Translation for Automatic Dubbing," *ICASSP*, 2022. DOI: 10/gqbxs2.

wbrannon@mit.edu