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

William Brannon*
MIT
wbrannon@mit.edu

Yogesh Virkar
AWS AI Labs
yvvirkar@amazon.com

Brian Thompson
AWS AI Labs
brianjt@amazon.com

*Work performed during an internship at Amazon

Transactions of the Association for Computational Linguistics (2023) 11: 419–435

Motivation

How do humans dub video content between languages?

ML meets humanities:

- Qualitative work on human dubbing
- ML work on automatic dubbing

Dubbers face many constraints, but can’t satisfy all of them. How do they trade off?

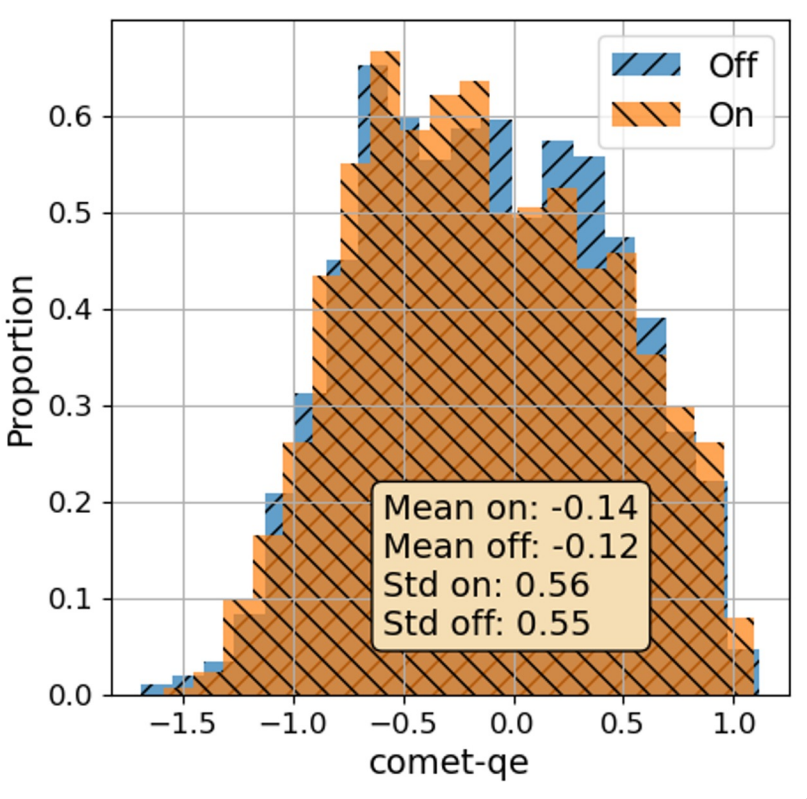
Data Sources

- **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.

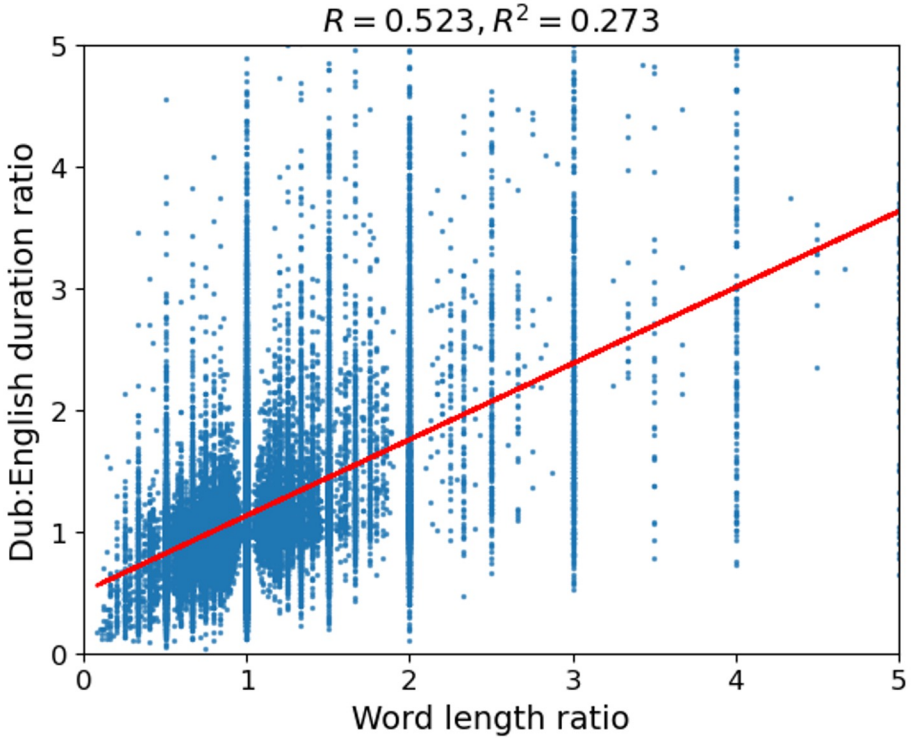
Answer: No measurable worsening of translation quality!



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?

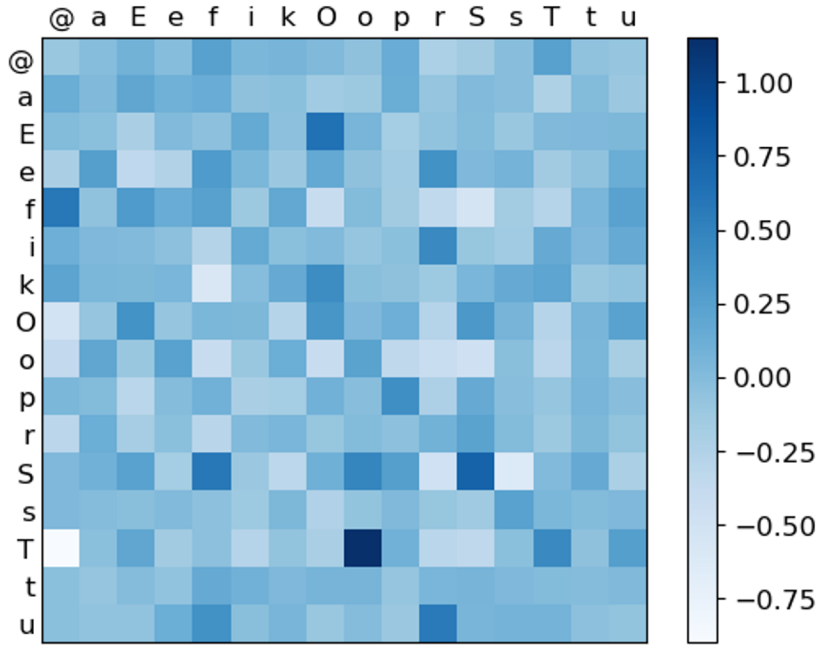
Answer: Longer speech! Dubbers would rather break timing constraints than vary speaking rates.



Lip Sync

Question: Does dub speech better align with mouth movements when onscreen (actors' mouths visible)?

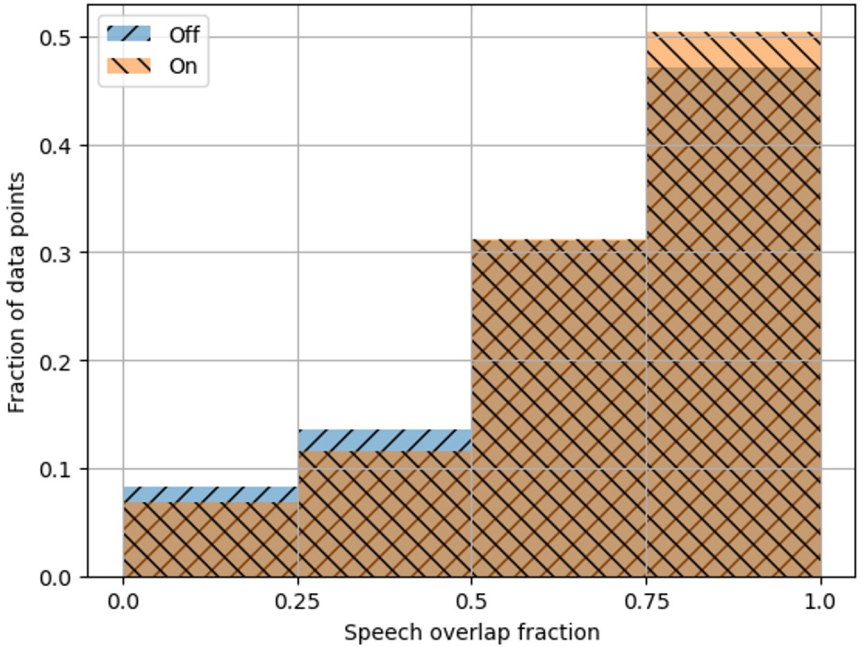
Answer: Yes, but not by much. There's little pattern visible in English / dub viseme (mouth movement class) cooccurrence plot.



Isochrony

Question: Are original timing constraints respected? Specifically, does source/dub speech timing match up more onscreen than off? Onscreen is more constraining.

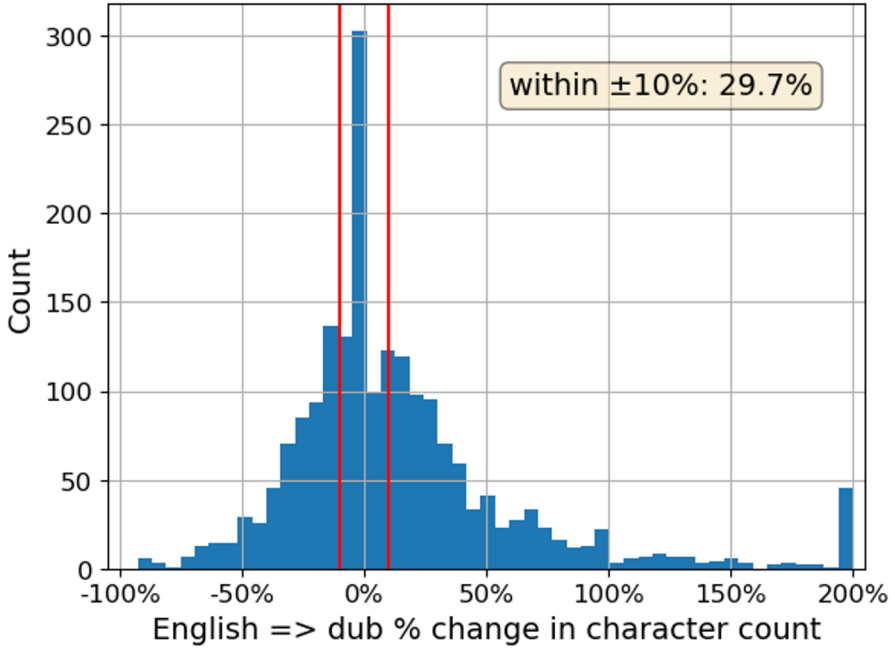
Answer: Less than expected. Isochrony is strong; response to onscreen constraint is not.



Isometry

Question: Are original and dub texts about equally long? Do human dubs follow prior ML work's ±10% length threshold?

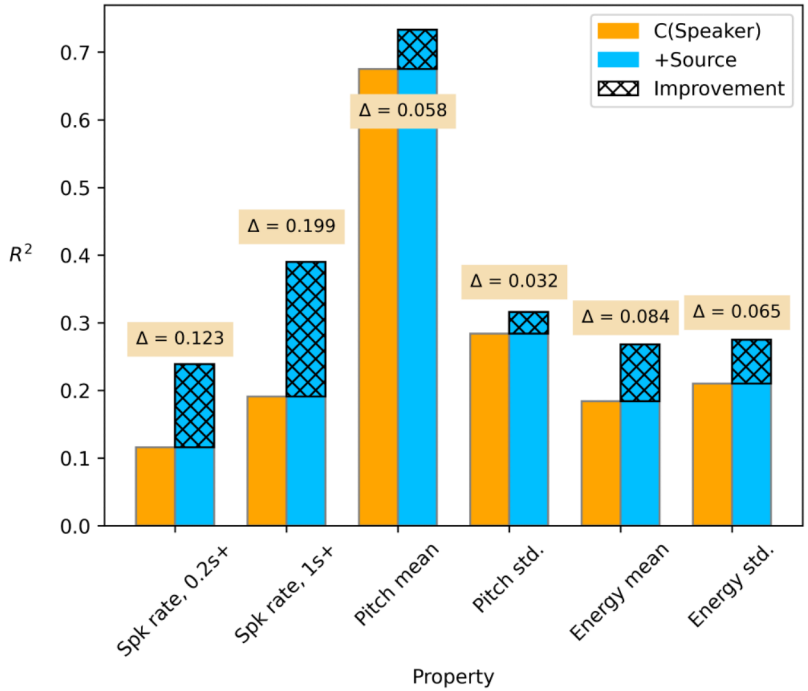
Answer: No! Most human dubs are not isometric.



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).



Conclusions

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

➡️ Isochrony and lip sync matter, but not as much

✅ 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 likely **not useful** for automatic dubbing

TACL Paper:

doi.org/10/gr9cbz

Questions? Want to collaborate? Interested in working/interning at Amazon?

wbrannon@mit.edu

brianjt@amazon.com