Prosodic Phrase Alignment for Machine Dubbing

Alp Öktem
Col·lectivaT

Mireia Farrús
Universitat Pompeu Fabra

Antonio Bonafonte
Universitat Politècnica de Catalunya

ABSTRACT

Dubbing is a type of audiovisual translation where dialogues are translated and enacted so that they give the impression that the media is in the target language. It requires a careful alignment of dubbed recordings with the lip movements of performers in order to achieve visual coherence. In this paper, we deal with the specific problem of prosodic phrase synchronization within the framework of machine dubbing. Our methodology exploits the attention mechanism output in neural machine translation to find plausible phrasing for the translated dialogue lines and then uses them to condition their synthesis. Our initial work in this field records comparable speech rate ratio to professional dubbing translation, and improvement in terms of lip-syncing of long dialogue lines.

AUDIOVISUAL TRANSLATION

- ***Dubbing*** is voice acting on top of the dialogues in e.g. movies, series, commercials
- Makes the media accessible to viewers of another language
- Dubbing is preferred to subtitles in many countries and also by children and the visually-impaired
- It is carried out in professional studios involving actors, engineers and directors

lip·syncing

Reproduction of timing, phrasing and phonetic content of the original speech segments in the target language to match the lip movements of the original performers

- Comparable number of syllables
- *Similar phrasing and pausing structure*
- Matching mouth articulation movements (like opening, closing etc.)

Code and samples

github.com/alpoktem/MachineDub

Pre-print

arxiv.org/abs/1908.07226

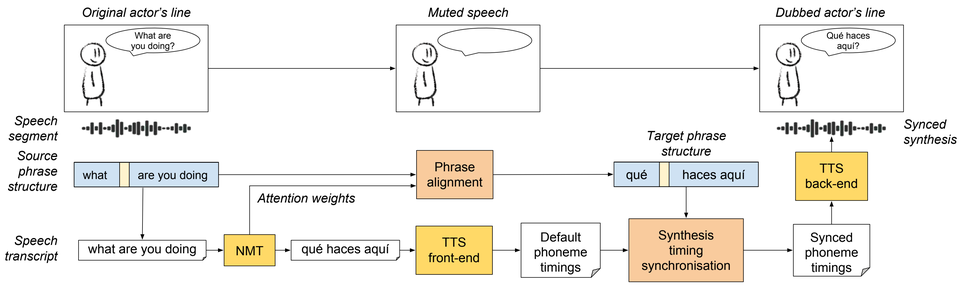
MOTIVATION

Compared to subtitling, there are relatively few automated solutions for dubbing.

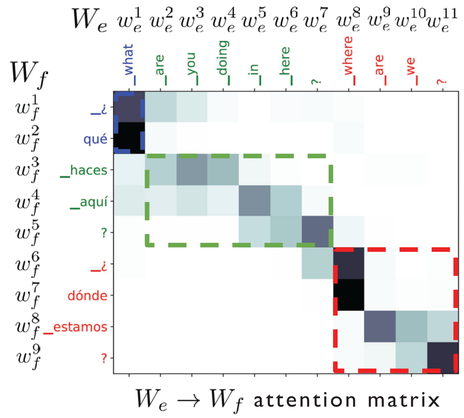
Machine translation with lip-syncing together has never been addressed before.

Realistic automatic dubbing could be useful for media content creators in reaching foreign audiences.

PIPELINE



CROSS-LINGUAL PHRASE ALIGNMENT



Translated phrase must align voiced segments to achieve lip-sync

Attention weights can be used to align prosodic phrases (PP)

How to split the MT output into phrases that reflect the input phrasing structure?

1. Populate possible target PP label sequences that contain same number of unique PP labels in the same order

$$S : \{L_f^1, \ldots, L_f^X\}$$

2. Rank sequences in S using a scoring function based on attention weights

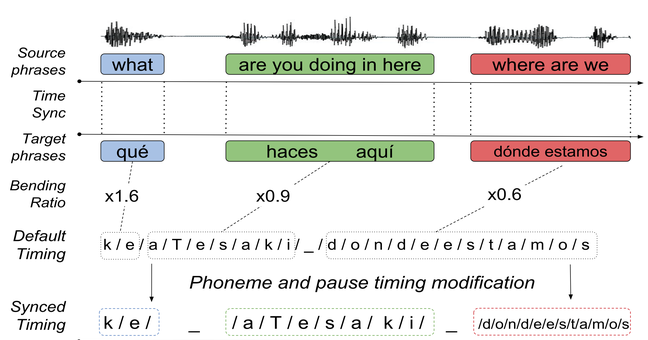
$$score(s) = \prod_n^N \sum_m^M W_{masked}^l(n;m)$$

$$W_{masked}^l(i;j) = \begin{cases} W(i;j) \text{if}, & l = l_e^j \\ 0, & \text{otherwise} \end{cases}$$

3. Select best scoring sequence as target PP sequence

$$L_f = \arg\max_{s \in S} score(s)$$

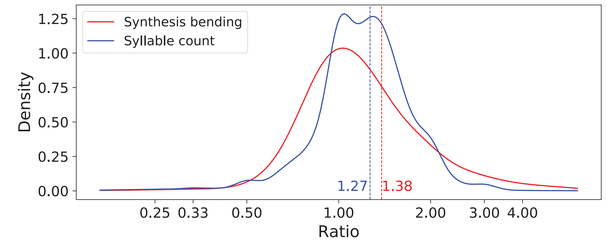
How to synchronise synthesis to the original utterance phrase timing?



DATA

- Heroes corpus contains bilingual speech segments from the TV series Heroes.
- 7000 single speaker speech segments were extracted from the original and Spanish dubbed version of 21 episodes.
- Audio segments are accompanied with subtitle transcriptions, word-level timing and prosodic/paralinguistic information.

EVALUATION



How much the aligned prosodic phrases match in terms of their articulation length?

Average speech rate ratio (1.27) comparable to professional dubbing measured in Heroes corpus (1.31).

How much of a speed-up or slow-down rate has been applied for synthesis synchronization?

TTS was more likely to be sped-up than slowed-down.

Perception Test

18 Spanish speaking participants were asked to compare our approach to subtitle reading method on dubbed video samples in terms of translation quality and lip-syncing precision.

System	MOS		Preference
	Translation	Lip-sync	
subtitle	4.08	3.44	%68
synced	2.96	3.58	%32

DISCUSSION & FUTURE WORK

- Comparable speech rate ratio was obtained in average
- Acceptable or better lip-syncing quality with automatically translated and synchronized dialogue lines
- Very high or very low bending ratios lead to unnatural sounding syntheses
- Further lip-syncing guidelines are left unadressed
- Scoring mechanism for phoneme level alignment could help achieve mouth articulation synchronization
- Using alternative translations from N-best lists can be used to find optimal length translations

CONTACT

alp@collectivat.cat
alpoktem.github.io
@OktemAlp

ACKNOWLEDGEMENTS

Travel costs of the author during presentation of this poster were sponsored by Universitat Politècnica de Catalunya and Translators without Borders. The second author is funded by the Spanish Ministry through the Ramón y Cajal program.