Weakly supervised spoken term discovery using cross-lingual side information

Sameer Bansal[1], *Herman Kamper*[1,2], *Sharon Goldwater*[1], *Adam Lopez*[1]

[1]Institute for Language, Cognition, and Computation
[2]Centre for Speech Technology Research
School of Informatics, University of Edinburgh, UK
sameer.bansal@ed.ac.uk, kamperh@gmail.com, {sgwater, alopez}@inf.ed.ac.uk

Big picture

- High-quality ASR systems are built using hundreds of hours of transcribed speech data and pronunciation dictionaries.
- Available for a tiny fraction of the world's spoken languages as most are zero or low resource.
- *Zero-resource* speech technology aims to develop useful systems in such scenarios.
- Learning from audio alone is very challenging.
- We ask whether using side information could improve performance.

Unsupervised Term Discovery

UTD systems search for pairs of audio segments that are similar using dynamic time warping (DTW) distance.

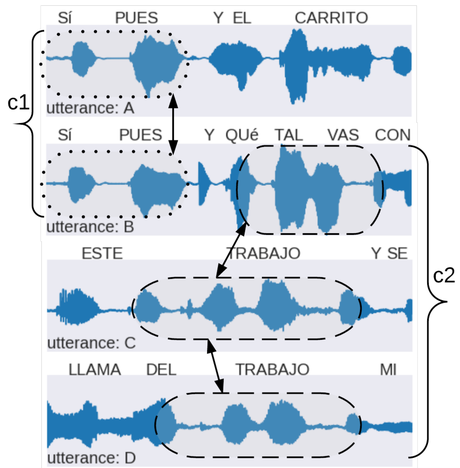


Figure 1: Acoustic pattern detection

Susceptible to errors due to:

- Acoustic variability between speakers
- Background noise.
- How to handle phonetically similar, but semantically different utterances?

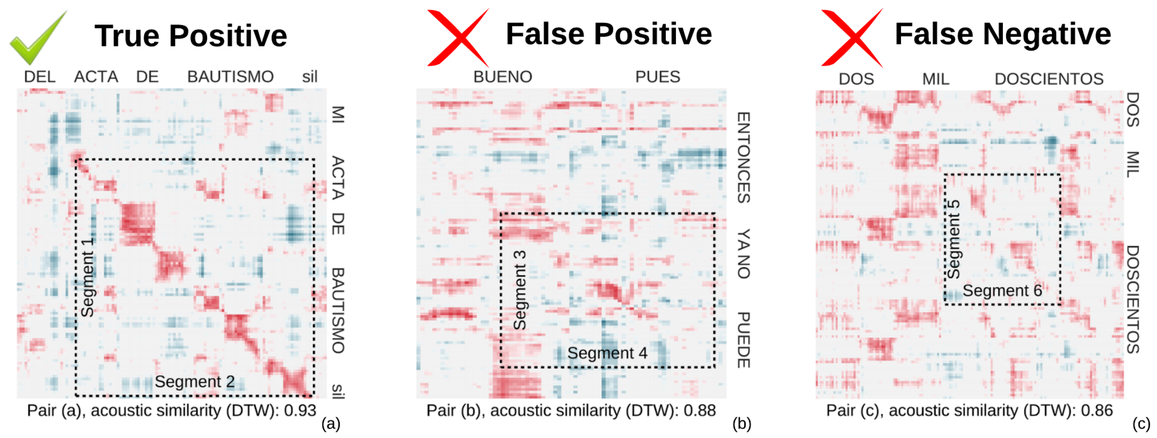


Figure 2: Acoustic similarity (dtw score) for utterance pairs.

English translations	Acoustic similarity	Translation similarity
to tell them to send me my *baptism act*	0.93	0.125
going to need the sacrament of *baptism paper*		
not now now then he cant anymore	0.88	0
yes well its good well yeah		
okay this the address *two thousand two hundred*	0.86	0.600
two thousand two hundred		

Side information

- Translations easier to obtain than transcriptions
- In disaster relief scenarios such as the 2010 Haiti earthquake, translations rapidly crowd-sourced.
- An option for languages without a written form
- Experiments on a noisy multi-speaker corpus of telephone calls in Spanish, and their crowdsourced English translations.

Method

- Use translation similarity (Jaccard score) as a noisy signal to improve UTD.
- Audio and translations aligned at utterance level (not word level).

Similarity between a pair of translations (E_1, E_2):

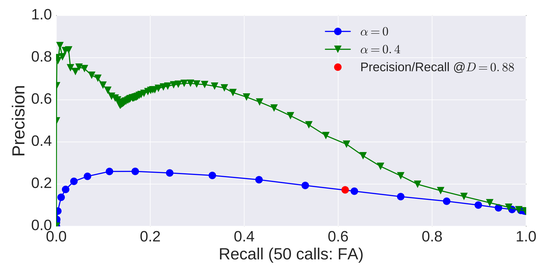
$$J = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$$

Rescore pairs returned by UTD using their translation similarity:

$$score_i = (1 - \alpha) \times dtw_i + \alpha \times J_i$$

Experiments

- Metric: *average precision*.
- Using a weighted score of translation similarity 40% and acoustic similarity 60% gives best results.



audio	audio only	with Eng. text
3 hrs.	0.34	0.58
7 hrs.	0.18	0.45

Table 1: Average precision results for baseline zero-resource and our system with translations.

Important Result

- Low-resource settings are more representative of real world scenarios.
- Simple method shows large gains in precision of acoustic pattern detection task by using side information.

Future work

- "Towards speech-to-text translation without speech recognition", S. Bansal, A. Lopez, H. Kamper, and S. Goldwater. In Proc. EACL, 2017.

Acknowledgments

Thanks to Ida for help with this poster. Federico, Marco, Joana, Sorcha and Clara for review comments.