Historical Linguistics-Informed Speech In-Context Learning for Low-Resource Language Varieties

Kalvin Chang* [1] kalvin1204@gmail.com Ming-Hao Hsu* [2] Soh-Eun Shim* [3] Shih-Heng Wang* [1] Alex Cheng* [1]
Hung-yi Lee [2] Barbara Plank [3] Shinji Watanabe [1] David R. Mortensen [1]

[1]Carnegie Mellon University [2]National Taiwan University [3]Ludwig Maximilian University of Munich

Motivation

- Pronunciation variation leads to biases in performance against non-standard varieties (Koenecke et al 2020)
- S3Ms (self-supervised speech models) cannot generalize to unseen pronunciation variants (Chang et al 2024)
- In-context learning does not require finetuning, unlike PEFT
- Setting: Non-standard varieties of high-resource languages
- **Neogrammarian hypothesis**: sound change is systematic (regular), affecting all instances of a sound in specific contexts

Pronunciation (romanization + IPA)			text
standard	variety 2	variety 3	
gí-giân [gi gien]	**gú-giân** [gu gien]	**gír-giân** [gi gien]	語言
hī [hi]	**hû** [hu]	**hīr** [hi]	魚
lí-hó [li hɤ]	**lú-hó** [lu ho]	**lír-hó** [li ho]	你好

Figure 1. Regularity of sound change for 3 varieties of Taiwanese Hokkien

Speech in-context learning

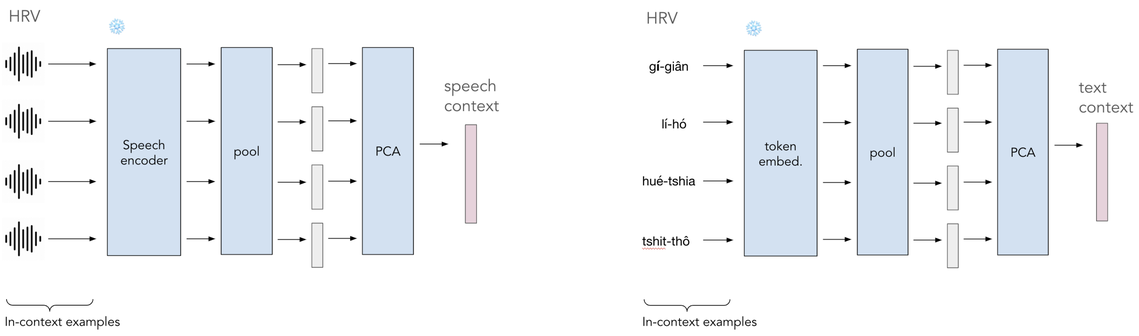

Figure 2. Extending in-context vectors (Liu et al 2023) to speech

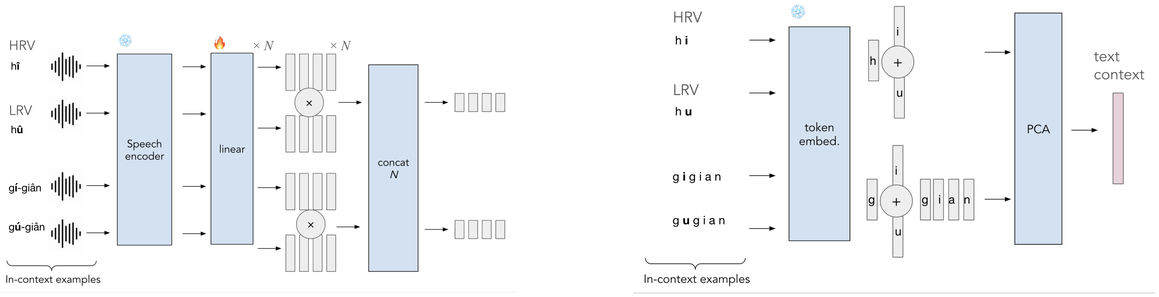

Figure 3. Extracting "accent shift" (Shao et al 2022) from sound correspondences across high-resource and low-resource varieties

Disclaimer

This paper is in the initial brainstorming stage. We're here to discuss ideas and move this further!

Datasets

Macro-language	HRV	Std Orth?	Resources
Chinese	Mandarin	Y	Center for the Protection of Languages, TAT_MOE
Swiss German	Standard German	Y	SwissDial, STT4SG-350, Swiss Parliaments
Dutch	Hollandic Dutch	N	Goeman-Taeldeman Van Reenen-Project
English	Mainstream American English	Y	SPADE, MD3
Arabic	Modern Standard Arabic	N	MGB-5, Casablanca
Italian	Standard Italian	N	Vivaldi

Table 1. Datasets with low-resource varieties of high-resource languages

Baseline

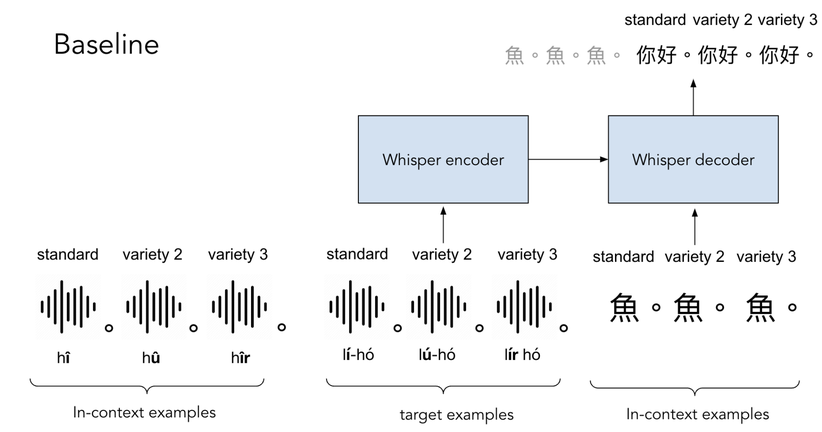

Figure 4. Wang et al (2024)'s speech in-context learning approach

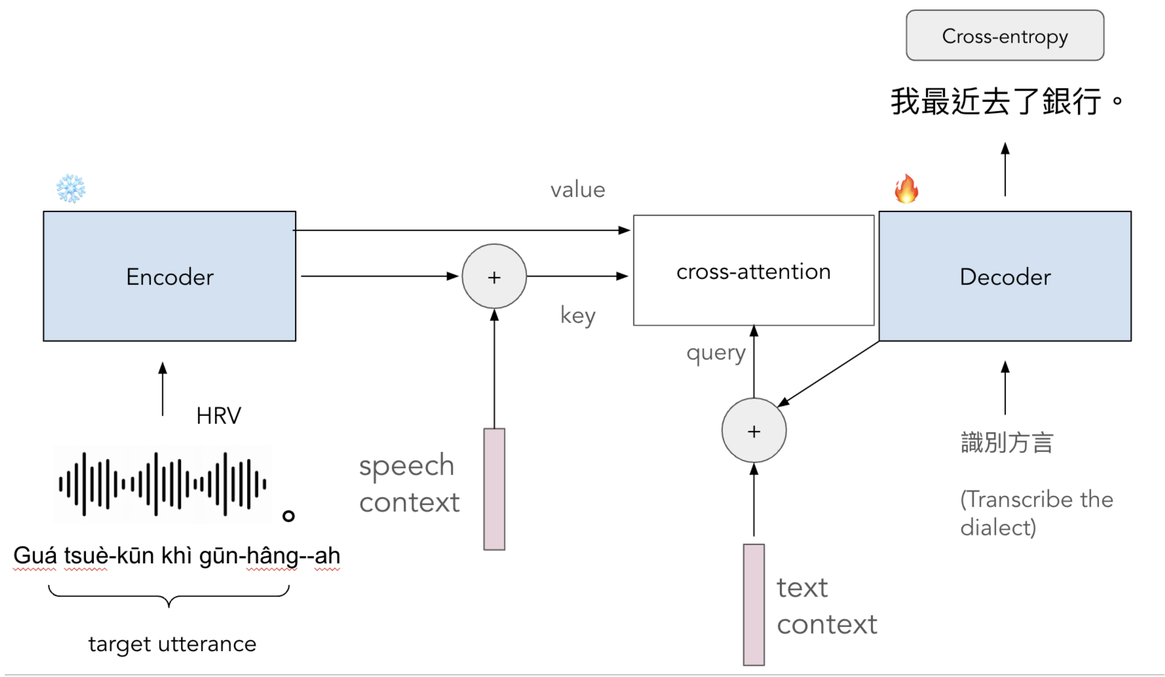

Figure 5. Modified cross-attention for in-context learning

Future work

- Data augmentation with sound correspondences
  - FST: HRV → protolanguage → LRV → phones
- Learning sound correspondences from speech
  - cognate set induction → extract semantic tokens → NMT → vocoder

* denotes equal contribution