On the Nature of Discrete Speech Representations in Multilingual Self-supervised Models

Badr M. Abdullah & **Mohammed Maqsood Shaik** & **Dietrich Klakow**

Language Science and Technology [**LST**]
Saarland University, Germany

SIGTYP Workshop
EACL 2023 — Dubrovnik, Croatia

UNIVERSITÄT DES SAARLANDES

SIGTYP

EACL 2023

On the Nature of Discrete Speech Representations in Multilingual Self-supervised Models

Badr M. Abdullah & Mohammed Maqsood Shaik & Dietrich Klakow

Language Science and Technology [**LST**]

Saarland University, Germany

SIGTYP Workshop

EACL 2023 — Dubrovnik, Croatia

Multilingual Self-supervised Speech Models

- **Self-supervision** is an effective paradigm for learning representations of spoken language from raw, **untranscribed audio**
- Self-supervised speech models can be **pre-trained** on a large sample of languages
  - ➡ **multilingual** models with **transferable** representations across languages
  - ➡ facilitate transfer learning for **low-resource languages**
- A **shared quantization module** within the model's architecture
  - ➡ transforms the continuous acoustic input into a **sequence of discrete units**

Multilingual XLSR-53 Model

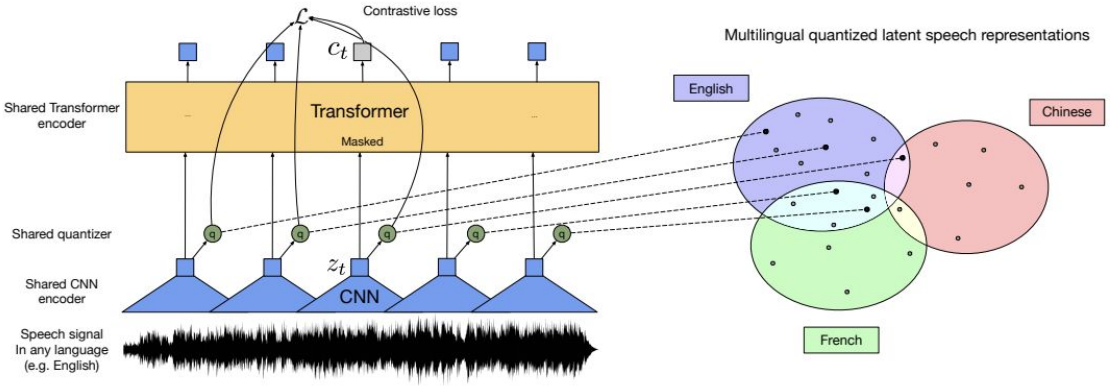
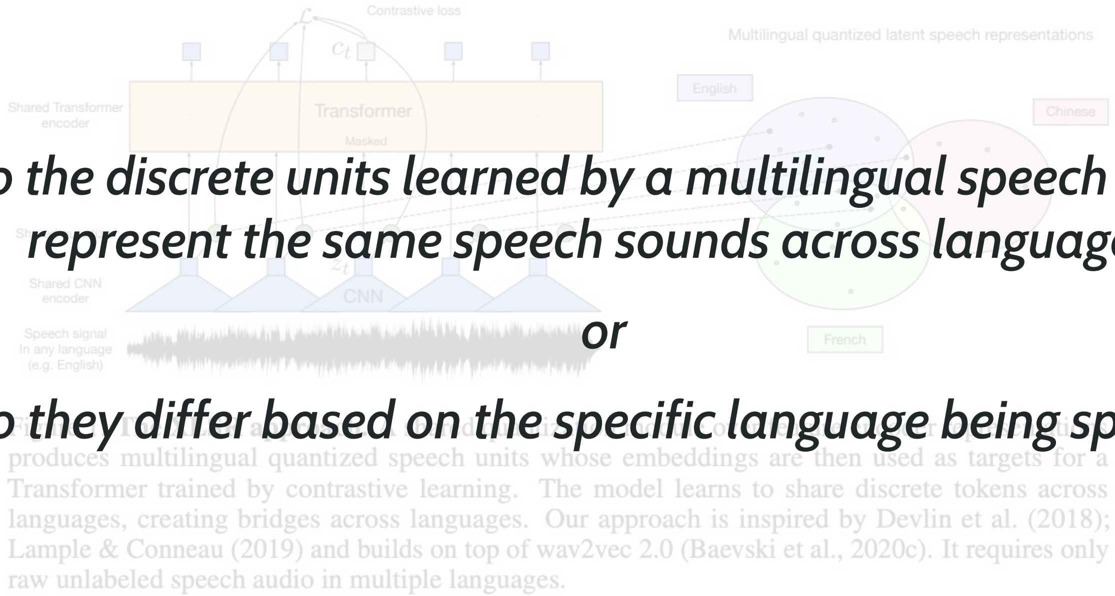


Figure 1: **The XLSR approach.** A shared quantization module over feature encoder representations produces multilingual quantized speech units whose embeddings are then used as targets for a Transformer trained by contrastive learning. The model learns to share discrete tokens across languages, creating bridges across languages. Our approach is inspired by Devlin et al. (2018); Lample & Conneau (2019) and builds on top of wav2vec 2.0 (Baevski et al., 2020c). It requires only raw unlabeled speech audio in multiple languages.

Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning for speech recognition."

Research Question

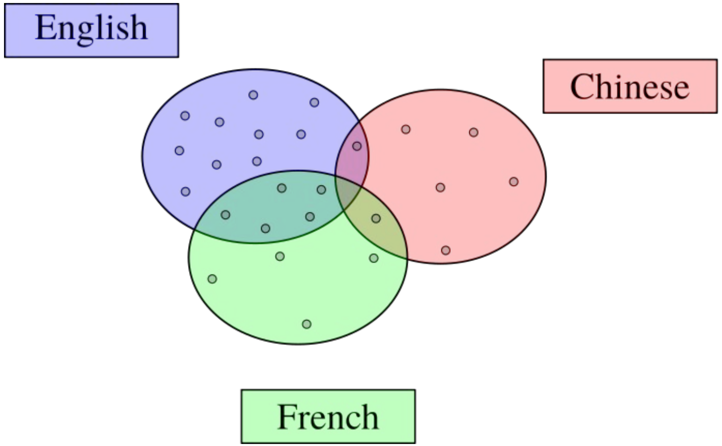


Do the discrete units learned by a multilingual speech model represent the same speech sounds across languages?

or

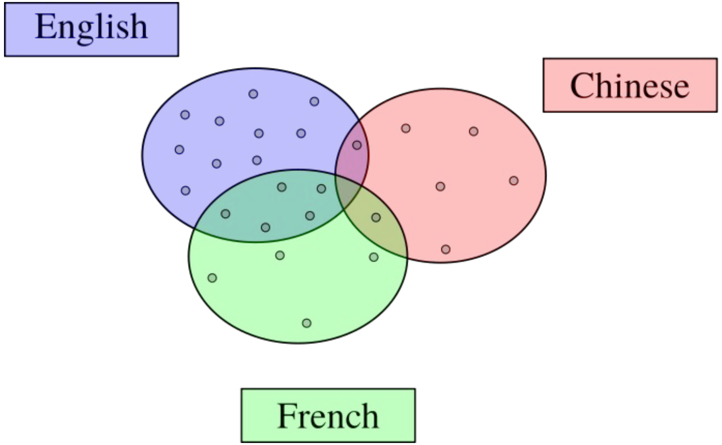
do they differ based on the specific language being spoken?

Discrete Speech Representations

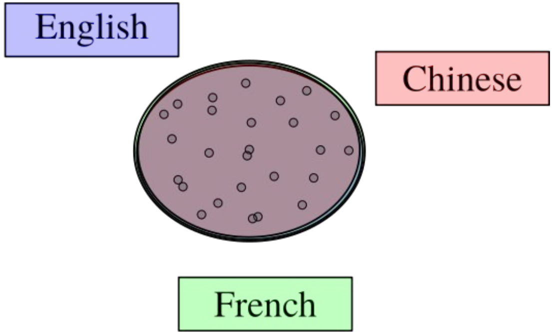


Expectations

Discrete Speech Representations



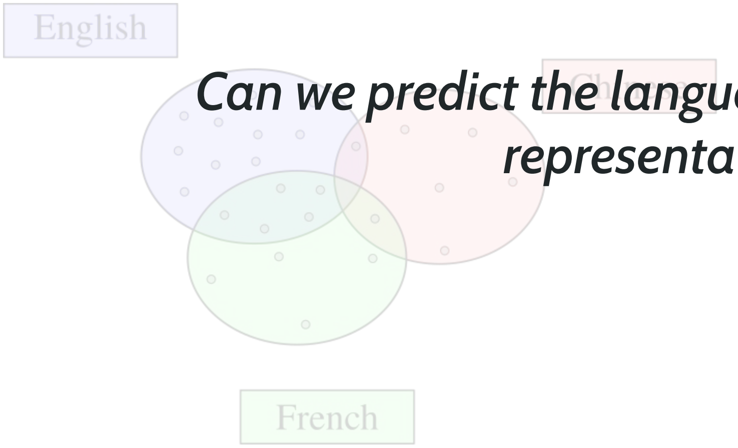
Expectations



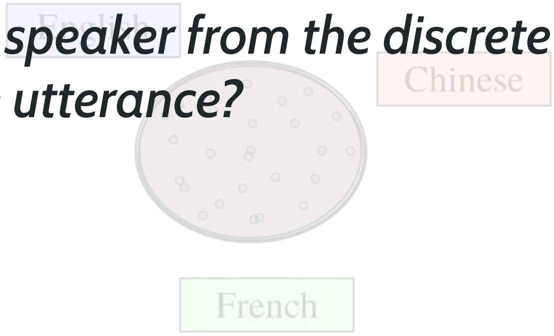
Reality

(Revised) Research Question

Can we predict the language of the speaker from the discrete representation of the utterance?

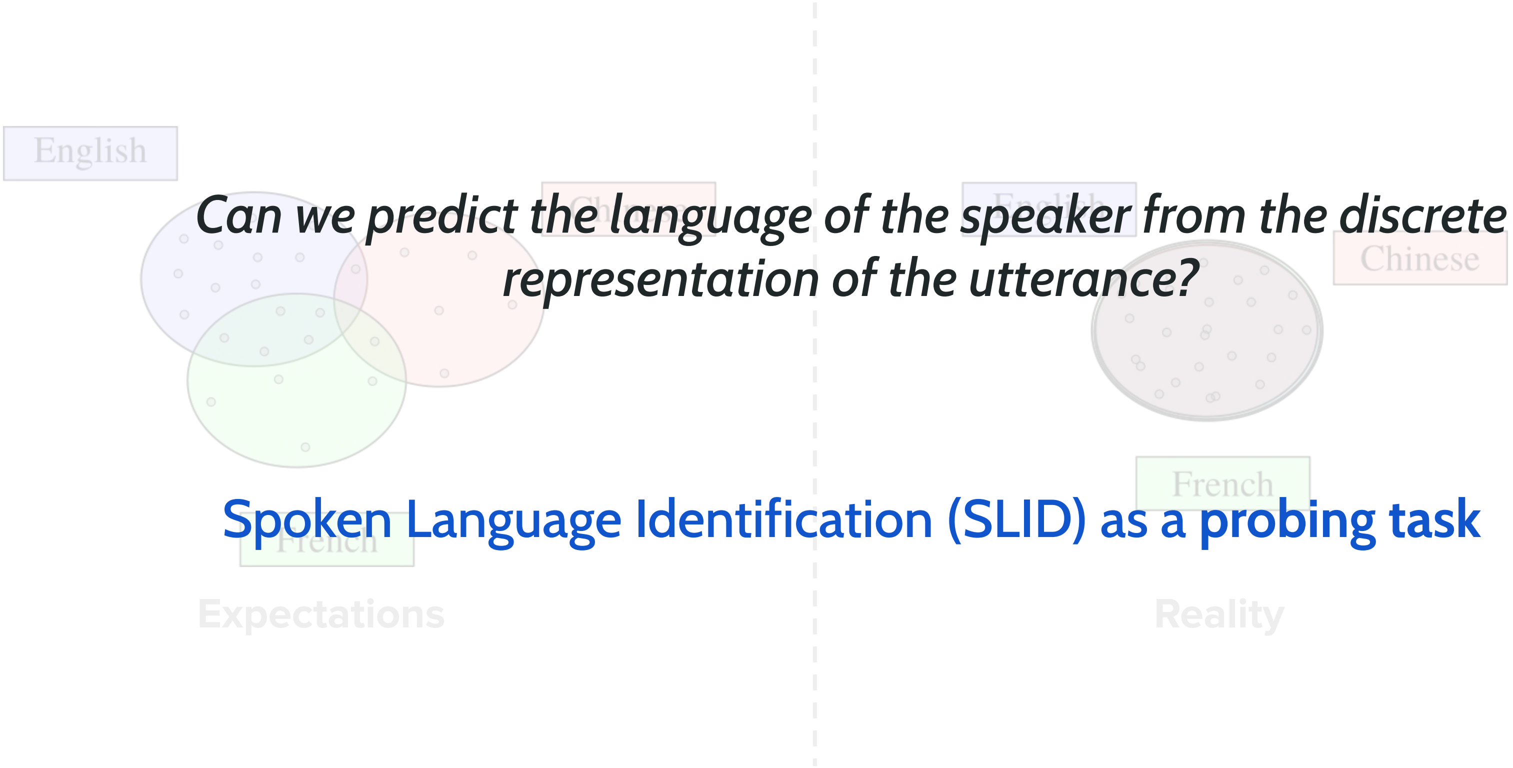


Expectations



Reality

(Revised) Research Question



Can we predict the language of the speaker from the discrete representation of the utterance?

Spoken Language Identification (SLID) as a **probing task**

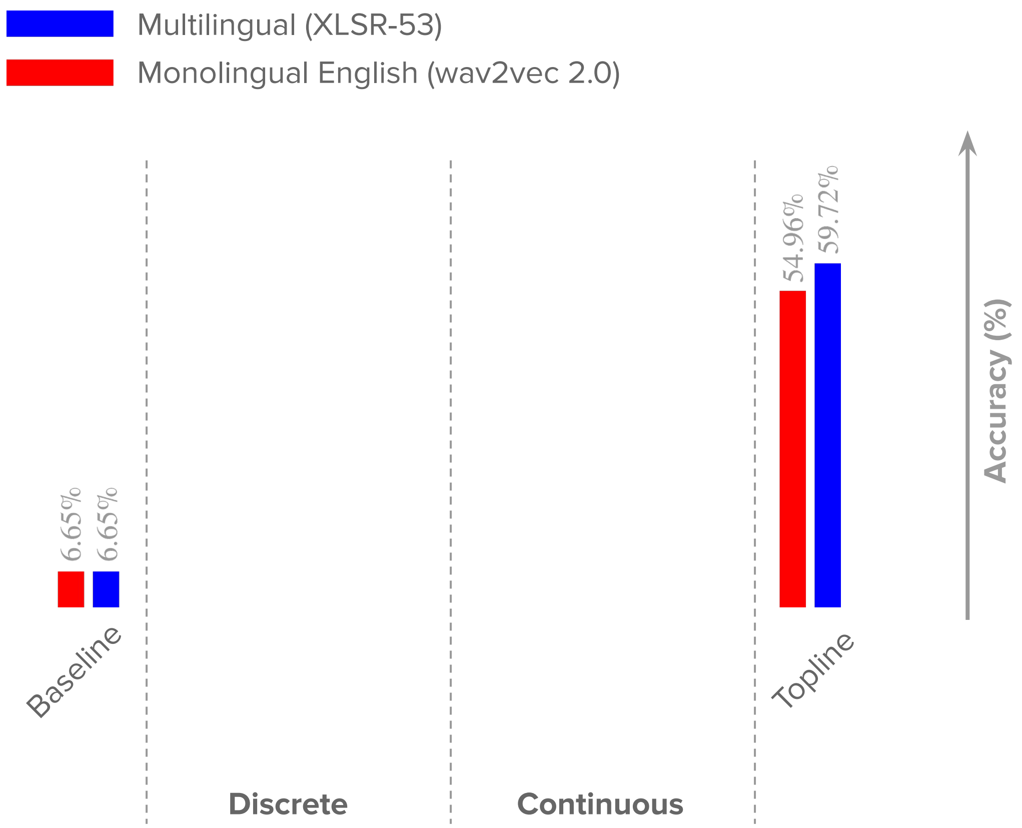
Language Sample

Common Voice speech corpus

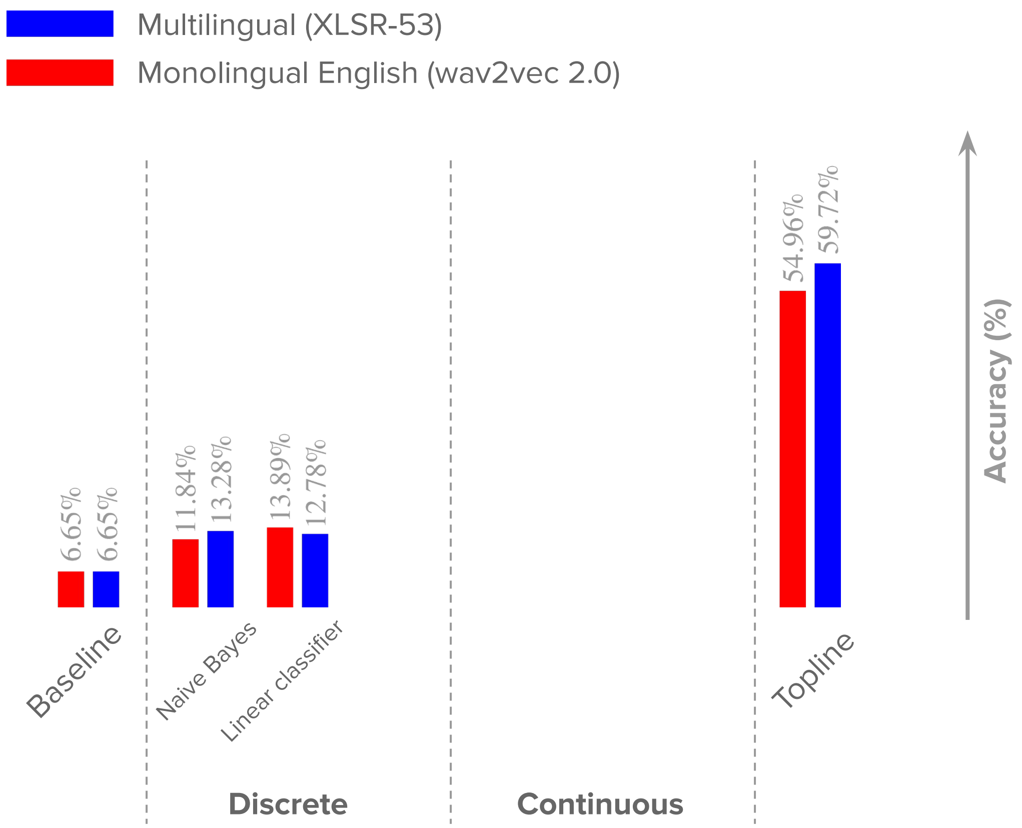
16 Indo-European languages

Romance	Germanic	Slavic	Celtic	Hellenic	Indo-Iranian
Catalan	German	Ukrainian	Welsh	Greek	Persian
Portuguese	Dutch	Russian	Breton		
French	Swedish	Polish			
Spanish	Frisian				
Italian					

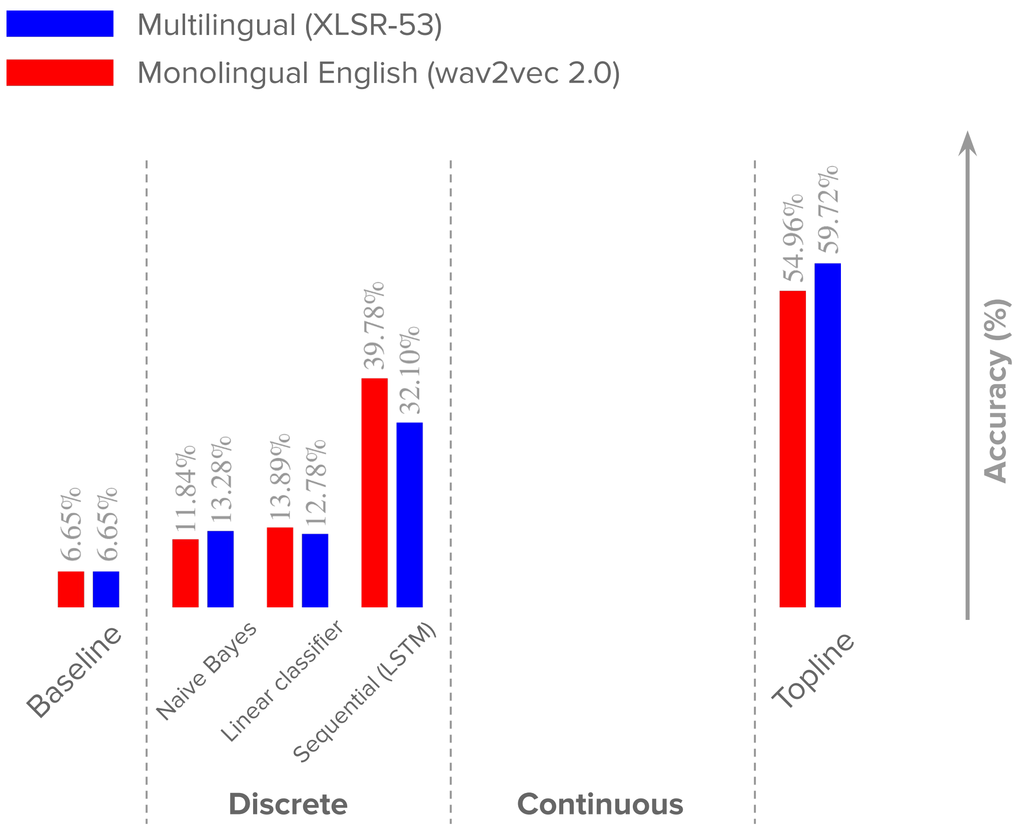
Experimental Results



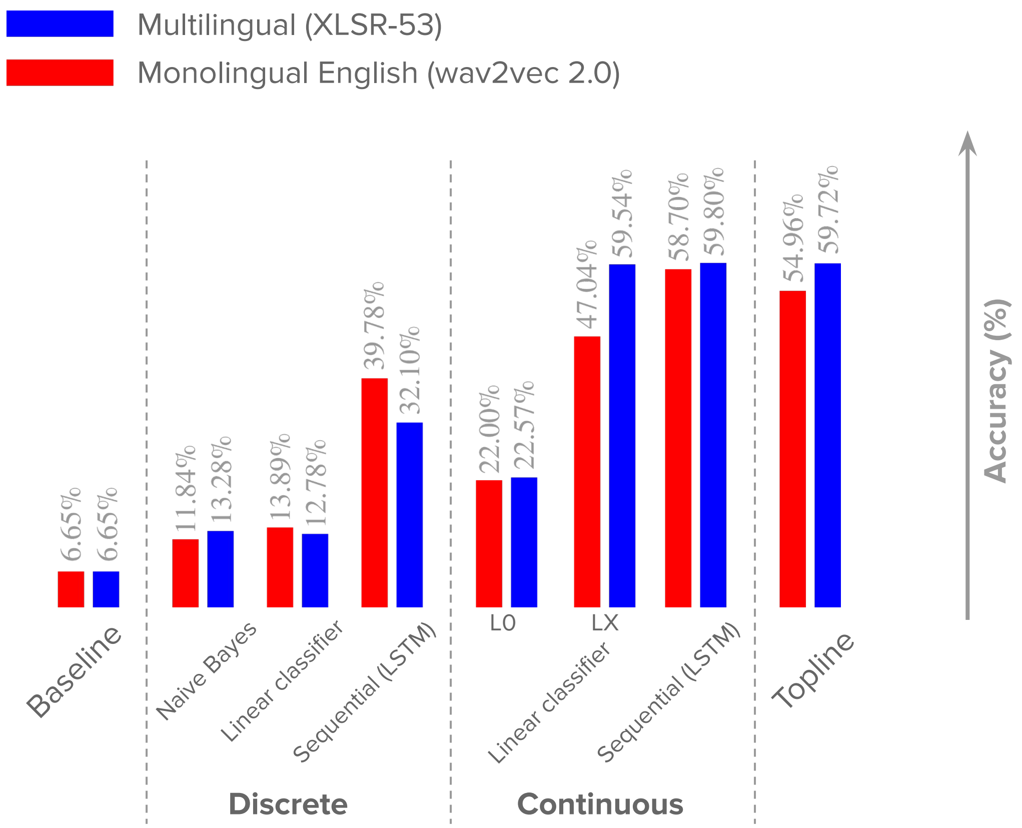
Experimental Results



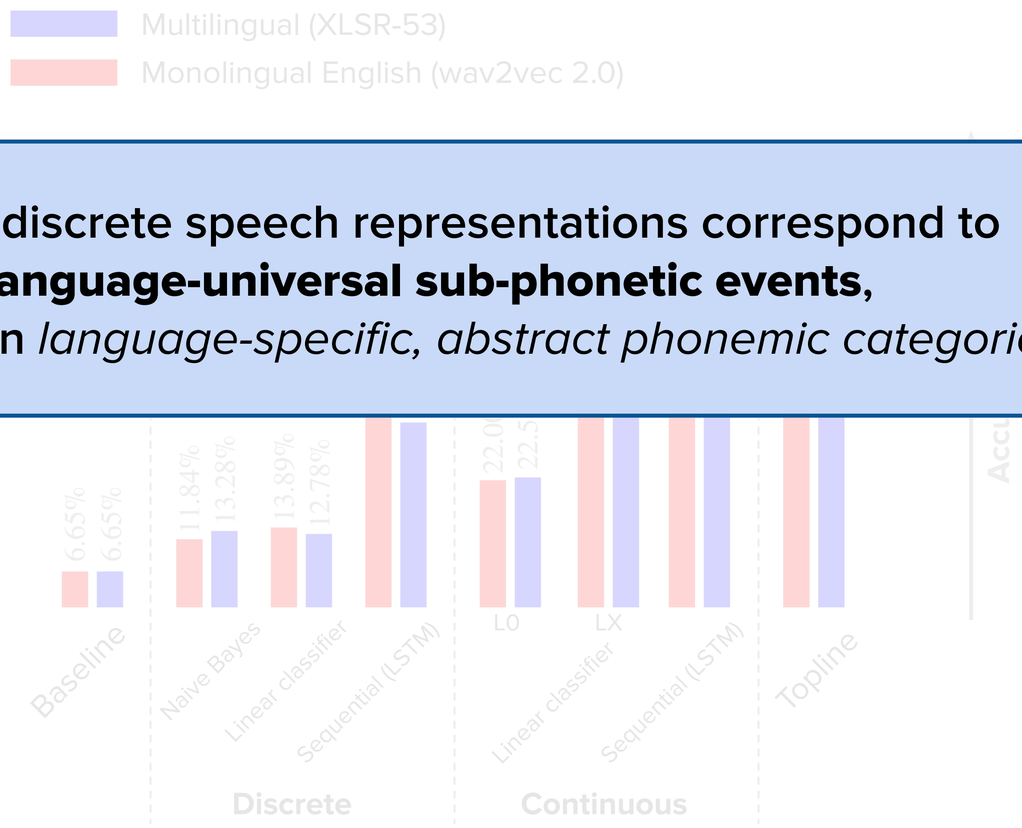
Experimental Results



Experimental Results

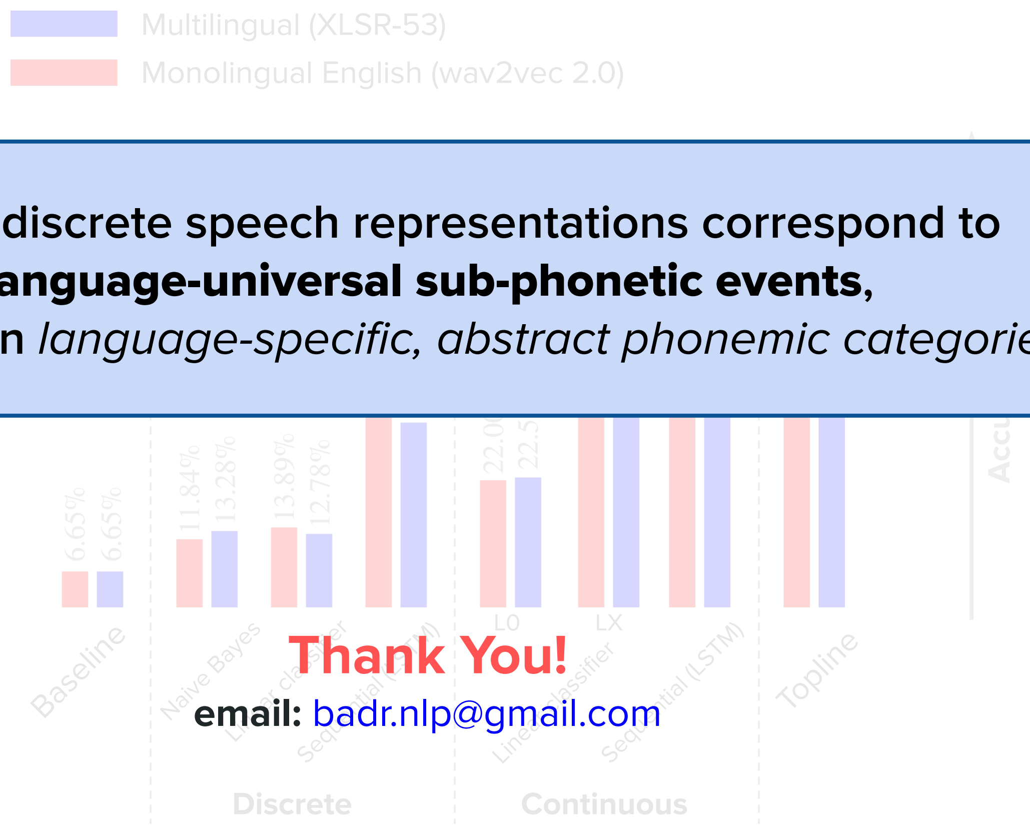


Experimental Results



Latent discrete speech representations correspond to
language-universal sub-phonetic events,
rather than *language-specific, abstract phonemic categories*

Experimental Results



Latent discrete speech representations correspond to
language-universal sub-phonetic events,
rather than *language-specific, abstract phonemic categories*

Thank You!

email: badr.nlp@gmail.com