Linguistically Interpretable Hierarchical CTC for Universal Phone Recognition

Kalvin Chang* [1] kalvin1204@gmail.com   Chin-Jou Li* [1]   Shih-Heng Wang* [1]   Eunjung Yeo [1]   Kwanghee Choi [1]
Aaricia Herygers [2]   Farhan Samir [3]   Jian Zhu [3]   Shinji Watanabe [1]   David R. Mortensen [1]

[1]Carnegie Mellon University   [2]alphaspeech   [3]University of British Columbia

Motivation

- Applications of phone recognition
  - atypical speech assessment
  - sociolinguistic coding
  - endangered language documentation
  - pronunciation training
- Low accuracies in multilingual phoneme recognizers
- Current outputs: broad phonemic trancription
  - ex: /b æ t/ "bat"
  - phonemes are language-specific
- Goal: allophone-level transcription
  - ex: [p æ t] "bat"
  - language universal

Architecture

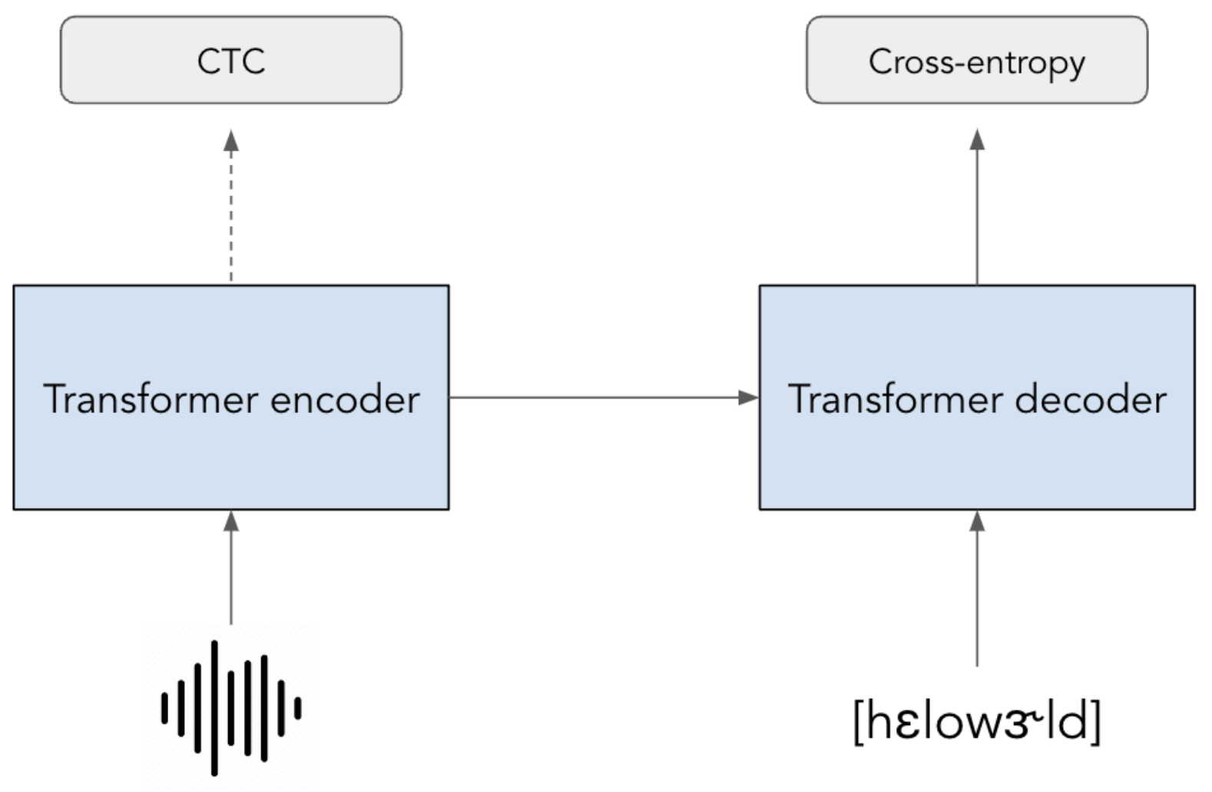

Figure 1. Encoder-decoder for phone recognition

- weakly supervised phoneme transcriptions (like Whisper and OWSM)
- auxiliary CTC loss to ensure monotonic alignments (Kim et al 2017)
- Hierarchical CTC approach (Higuchi et al 2023)

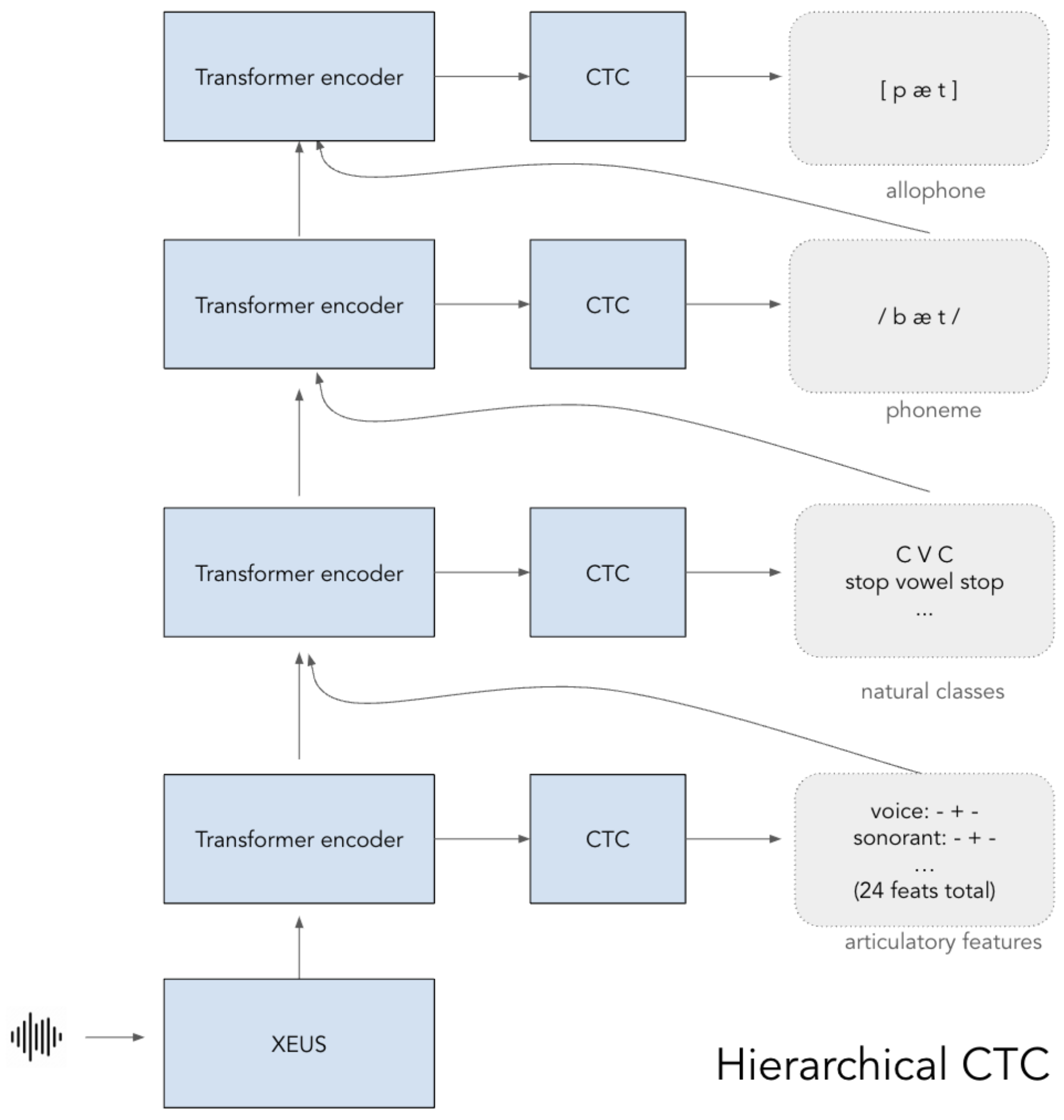

Figure 2. Proposed hierachical CTC approach

Datasets

- IPAPack (Zhu et al 2024)
  - 1000 hours (eventually 20000 hours)
  - 115 languages
  - phonemic transcriptions obtained via G2P

Baselines

- XEUS features: S3M pretrained on 4,057 languages, 1.1m hours (Chen et al 2024)
- auxiliary CTC losses to predict the articulatory features (Glocker et al 2023)
- Frozen XEUS
  - Transformer encoder + phoneme CTC (29.71% CER)
  - Transformer encoder + phoneme CTC + 24 articulatory feature CTC
- Finetuned XEUS
  - phoneme CTC
  - phoneme CTC + 24 articulatory feature CTC

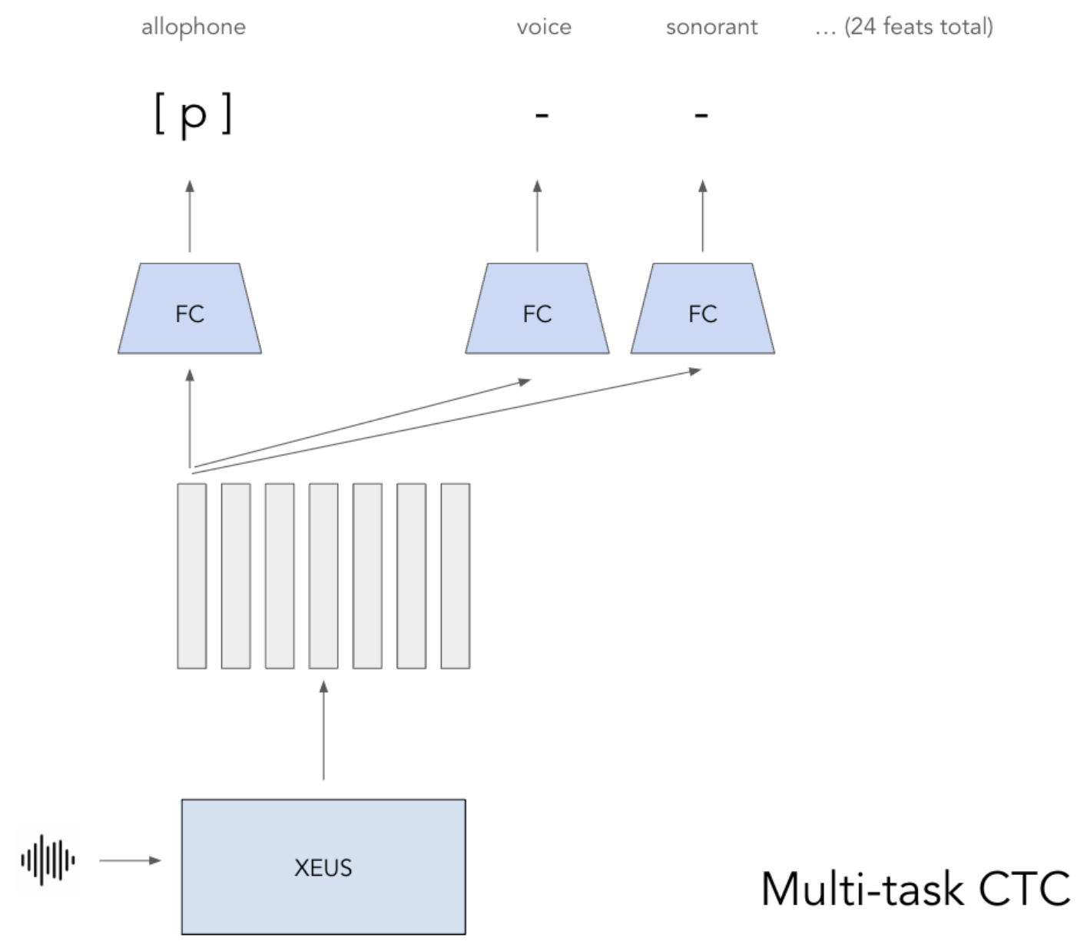

Figure 3. Multi-task CTC baseline with 24 articulatory feature losses

Future work

- Foundation model
  - G2P on OWSM's pre-training data
  - Foundation model enables in-context learning
- Joint phone recognition & forced alignment
  - forced alignment important for downstream phonetic analysis
  - learn forced alignment in unsupervised fashion
  - extract phone-level alignment from a modified CTC (Huang et al 2024) to address peakiness of CTC
  - or forward-sum loss (Shih et al 2019, Badlani et al 2021, Zhu et al 2024, Koriyama 2024)

Disclaimer

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

* denotes equal contribution