POWSM: A Phonetic Open Whisper-Style Speech Foundation Model

Chin-Jou Li* [1] chinjoul@andrew.cmu.edu Kalvin Chang* [2] Eunjung Yeo [3] Kwanghee Choi [3] Masao Someki [1]
Farhan Samir [4] Jian Zhu [4] David R. Mortensen [1] Shinji Watanabe [1]

[1]Carnegie Mellon University [2]UC Berkeley [3]UT Austin [4]University of British Columbia

Motivation

Phone recognition has broad applications

- Documentation of endangered languages
- Assessment of atypical speech
- Sociolinguistic analysis and coding
- Pronunciation training and feedback

Goal: Accurate phonetic transcription

- More directly aligned with the acoustic signal
  - e.g., [p æ t] for “bat”
- Challenges for multilingual phoneme recognition
  - Phoneme inventories are language-specific
  - IPA is not entirely language-independent

Method

Dataset

- **Training:** IPAPack++ [1]
  - 17k hours of multilingual speech from ASR datasets
  - Paired with phonemic IPA transcriptions
- **Evaluation:** generalization to new settings
  - Unseen languages: DoReCo, VoxAnglex
  - Sociophonetic variation: Buckeye, L2-ARCTIC

Architecture

- Encoder-decoder model based on OWSM [2]
- Multitask learning setup:
  - Phone recognition (PR): speech → IPA
  - ASR: speech → orthography
  - Audio-guided P2G: PR + orthographic context
  - Audio-guided G2P: ASR + IPA context
- Encoder is CTC-aligned to phone sequences
- Decoder is conditioned on language and task tokens
  - P2G, G2P contexts are provided as previous texts

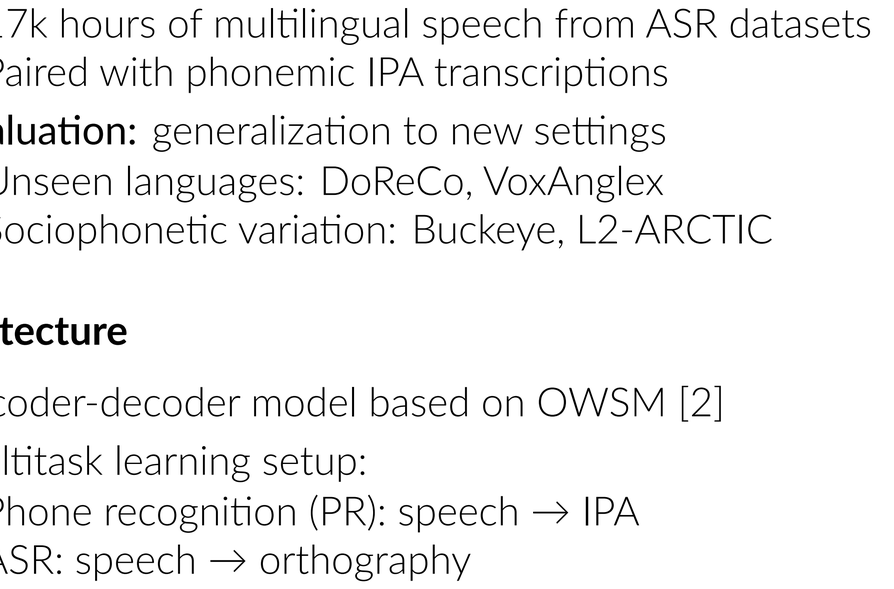

Figure 1. Overview of POWSM framework

Preliminary Results

Performance

- Training from scratch outperforms fine-tuning Whisper-small
- Multitasking improves both metrics
  - PER: phone error rate (%)
  - PFER: phonetic feature edit distance

Table 1. Comparison of PFER. (*: PR-only, 25 epochs, full IPAPack++)

Model	Param.	Struct.	Unseen DoReCo	Unseen VoxAngeles	Sociophonetic L2-ARCTIC	Sociophonetic Buckeye
Allosaurus	11M	enc	8.89	1.35	5.72	5.04
Whisper-PPT	244M	enc-dec	9.31	0.91	6.65	6.80
ZIPAs	300M	enc	5.93	0.75	3.63	3.91
POWSM*	344M	enc-dec	7.24	0.89	4.85	4.78

Table 2. PER & PFER on DoReCo. (Trained on 1k hour of IPAPack++)

Included tasks	PER	PFER
PR	59.98	9.51
PR + ASR	59.01	8.26
PR + ASR + G2P + P2G	**56.30**	**8.04**

Observations

- High proportion of vowel substitutions
  - Multilingual setting; vowels are continuous
- Lacks phonetic precision for sociophonetic variation
  - Also related to difference between transcriptions
- Frames near CTC peaks do prefer similar phones

Next Steps

Inspect current architecture

- Modify encoder targets
  - Incorporate phone similarity
  - Use label priors to reduce CTC peakiness
- Analyze multitasking benefits and potential drawbacks

Scaling up & additional tasks

- Predict articulatory features
- Explore in-context learning for unseen languages
- Predict timestamps to enable emergent phone alignment

References

[1] J. Zhu, F. Samir, E. Chodroff, and D. R. Mortensen. “ZIPA: A family of efficient models for multilingual phone recognition.” In *ACL 2025*.

[2] Y. Peng, J. Tian, W. Chen, S. Arora, B. Yan, Y. Sudo, M. Shakeel, K. Choi, J. Shi, X. Chang, J. weon Jung, and S. Watanabe. “Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer.” In *Interspeech 2024*.

* denotes equal contribution