Comparing Phonological Feature Sets for Low-Resource ASR

Introduction. Automatic speech recognition (ASR) tools have proven to be helpful to linguists in a variety of language documentation tasks, including reducing the amount of time spent transcribing speech audio completely from scratch (Ćavar, 2016; Prud'hommeaux, 2021). The task of an ASR model that can phonetically transcribe speech is to discover the mapping from utterance waveform to IPA symbol string. While this can be formulated directly as a training objective, training ASR models to encode more phonetically informed structure, such as articulatory features, have yielded positive results for transcription in a variety of deep learning architectures (Koreman et al., 1998; Xu et al., 2021; Taguchi et al., 2023; Lee et al., 2025). Phonet (Vásquez-Correa, 2019) is a bank of parallel bidirectional RNNs that extract phonological features from speech audio by estimating their posterior probabilities. Despite generally being outperformed by modern ASR systems, Phonet models remain valuable for a relatively straightforward comparison between featurally informed ASR and phoneme level ASR. In this preliminary work, we retrained Phonet models on two different feature sets to see the extent to which specific theories of phonological features facilitate better phoneme recognition, using a low-resourced language (Yan-nhangu, Pama-Nyungan) as a testing ground for performance. Using a naïve decoding strategy, we found that IPA features led to the best transcription accuracy, followed closely by a featureless baseline.

Feature sets. We use models trained in three feature settings: 1) naïve place/manner of articulation features of the IPA, 2) PHOIBLE features (Moran & McCloy, 2019; loosely based on the feature system in Hayes, 2009, with additions from Moisik & Eisling, 2011), and 3) a no-feature baseline in which symbols of the International Phonetic Alphabet (IPA) are predicted directly from speech audio. The features included in the IPA and PHOIBLE feature sets are included in Table 1.

Language(s). The corpus for this test includes selected field recordings of the Yan-nhangu language, a member of the Yolŋu subgroup of Pama-Nyungan (Glottolog code YANN1237; ISO-639 JAY). Yan-nhangu's phoneme inventory includes both place and manner contrasts, with stops at 6 points of articulation. There are two stop series distinguished by both length and VOT, and 6 vowels distinguished by height, length, and frontness.

Methodology. We retrained three Phonet models: two models predicting features from a predefined feature set (IPA or PHOIBLE features) and one baseline model predicting a phone sequence directly from the audio (no features). Featural models output *posteriorgrams*, matrices storing an activation from 0 to 1 for each feature for every 10 ms audio frame. To decode, for each audio frame, we computed the Euclidean or cosine distance between the posteriorgram's feature activations to the feature vector of each phone in the target language's inventory. A phone was predicted greedily by selecting the phone that minimizes distance in the resulting distance vector over the target phones. Since multiple audio frames can span a single phone, our naïve decoding strategy requires that for a window length of n phones in either direction, a phone is removed if it occurs less than m times in that window (n=3, m=4 worked best). Collapsing consecutive identical phones after this removal yields a predicted phone sequence, which was used to calculate average phone error rate (PER) against a human annotated label.

Results. Table 2 presents the PER for the three models across the two distance metrics. We found the model using IPA features yields the lowest PER (highest accuracy: 58%, compared to chance performance at 1/(# target phones)). Models trained on PHOIBLE features or decoding with cosine distance consistently worsened performance.

Conclusion. We presented preliminary results as a first step towards a more comprehensive exploration of the impact of different assumptions about the feature space in phonetically informed ASR models, compared to phoneme level ASR. We found that of our four configurations of feature set vs. distance metric, models making use of the place/manner of articulation features used for the IPA and a decoding strategy that uses Euclidean distance performed best. However, usage of more sophisticated decoding strategies, e.g., incorporating beam search, may reveal a different story. Future work aims to include such strategies, as well as expanding to cross-linguistic comparison, and qualitative analysis of errors in different feature systems.

IPA Features	**PHOIBLE Features**
voiced, bilabial, velar, palatal, dental, alveolar, retroflex, glottal, plosive, nasal, trill, lateral, approximant, vocalic, high, low, front, center, back, round, short, long, silence	syllabic, long, consonantal, sonorant, approximant, trill, nasal, lateral, labial, round, coronal, anterior, distributed, dorsal, high, low, front, back, tense, periodicGlottalSource, constrictedGlottis, silence

Table 1. Features included in the two feature sets.

	Euclidean distance	**Cosine distance**
IPA features	0.42145	0.43556
PHOIBLE features	0.52793	0.57085
Phone-level (baseline)	0.45352	

Table 2: PER (Phone Error Rate) across models and feature-to-phone similarity metrics

References. Ćavar, M., Ćavar, D., & Cruz, H. (2016). Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4004–4011. Hayes, B. (2009). *Introductory phonology*. Blackwell. Koreman, J. C., Andreeva, B., & Barry, W. J. (1998). Do phonetic features help to improve consonant identification in ASR?. In *ICSLP*. Lee, J., Mimura, M., & Kawahara, T. (2025). Leveraging IPA and Articulatory Features as Effective Inductive Biases for Multilingual ASR Training. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. Moisik, S. R., & Esling J. H. (2011). The "whole larynx" approach to laryngeal features. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 1406–1409). Moran, S. & McCloy, D. (eds.) (2019). PHOIBLE 2.0. *Jena: Max Planck Institute for the Science of Human History.* (http://phoible.org) Prud'hommeaux, E., Jimerson, R., Hatcher, R., & Michelson, K. (2021). Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation* 15: 491-513.Taguchi, C., Sakai, Y., Haghani, P., & Chiang, D. (2023). Universal automatic phonetic transcription into the international phonetic alphabet. *arXiv preprint arXiv:2308.03917*. Xu, Q., Baevski, A., & Auli, M. (2021). Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*.