Toward a Coarse-Labeled Spoken Language Identification Dataset for Central Alaskan Yup'ik and Samoan from US Broadcast Archives

Yangyang Chen **Kyeongmin Rim** **James Pustejovsky**
Department of Computer Science
Brandeis University
{yangyangchen, krim, jamesp}@brandeis.edu

Abstract

Publicly available spoken language identification (LID) systems give sparse and inconsistent coverage of the languages spoken in US communities beyond the contiguous mainland — Alaska Native languages and the languages of the US Pacific Island territories. No system on HuggingFace covers Central Alaskan Yup'ik except the largest variant of Meta's MMS-LID family, and only three MMS-LID variants cover Samoan, while Whisper and VoxLingua107-based models lack both despite including other Polynesian languages. We describe an ongoing effort to build a coarse-labeled LID dataset for Yup'ik and Samoan from US public broadcast archives, benchmark publicly available LID systems on it, and train a simple MLP classifier on top of two frozen pretrained speech encoders as prototypes. We report preliminary corpus statistics, off-the-shelf model performance, and prototype results. Guided by the distinctive phonological typology of the target languages, we outline a phonologically-informed fine-tuning direction as future work.

1 Introduction

Spoken language identification is a prerequisite for nearly every downstream speech-processing task in multilingual settings, including automatic speech recognition, transcription, search, and indexing of audio archives, yet off-the-shelf LID coverage of the languages spoken in US communities beyond the contiguous mainland — Alaska Native languages and those of the US Pacific Island territories — is sparse and inconsistent. Table 1 surveys the publicly available spoken LID systems on HuggingFace against our two target languages — Samoan (ISO 639-3: `smo`) and Central Alaskan Yup'ik (hereafter Yup'ik;[1] ISO 639-3: `esu`) — together with related languages from the same families (Hawaiian and Maori for Polynesian; Central Siberian Yupik for Eskimo-Aleut) that illustrate the inconsistency of existing coverage. Whisper (Radford et al., 2023), VoxLingua107-based models (Valk and Alumäe, 2021), and ESPnet OWSM (Peng et al., 2024) include Hawaiian and Maori but neither Samoan nor any Yupik variety. Meta's Massively Multilingual Speech (MMS) LID family (Pratap et al., 2023) covers Samoan only at its 1024-label size and above, and covers Yup'ik only at the largest 4017-label variant. At 4017 labels, the per-class prior is approximately 2.5×10^{-4}, and zero-shot performance on under-represented languages is accordingly poor. To our knowledge, `mms-lid-4017` is the only publicly available spoken LID model that covers both target languages.

		Target		Related		
System	#langs	`smo`	`esu`	`haw`	`mi`	`ess`
Whisper large-v3	99	—	—	✓	✓	—
VoxLingua107	107	—	—	✓	✓	—
ESPnet OWSM v4	∼150	—	—	✓	✓	—
mms-lid-126/256/512	126–512	—	—	—	—	—
mms-lid-1024	1024	✓	—	—	—	—
mms-lid-2048	2048	✓	—	—	—	—
mms-lid-4017	4017	✓	✓	✓	✓	✓

Table 1: Coverage of the two target languages (Samoan `smo`, Yup'ik `esu`) and three related languages from the same families (Hawaiian `haw`, Maori `mi` for Polynesian; Central Siberian Yupik `ess` for Eskimo-Aleut) across publicly available spoken LID systems. Only `mms-lid-4017` covers both target languages. That it also covers `ess` (the primary source of Yup'ik confusion in Section 5) means the model has both labels in its vocabulary yet cannot reliably distinguish them.

Against this backdrop, we describe an ongoing effort to build LID capability for Yup'ik and Samoan using archival broadcast recordings from US public media (Section 4). This setting differs from standard LID corpora in three important

[1] We follow the standard orthographic convention in which the apostrophe in Yup'ik marks the geminate /pː/ specific to the Central Alaskan variety; *Yupik* without the apostrophe refers to the broader language family.

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 203–211

ways: (i) speech is interspersed with English in code-switching and bilingual programming contexts, (ii) audio quality varies across decades from analog tape to digital capture, and (iii) a large fraction of the raw material is non-speech (music, ambient sound, silence) that must be filtered before any language labeling.

The pairing of these two languages is driven by the archive and the communities it serves, not chosen arbitrarily. The American Archive of Public Broadcasting (AAPB) — a national collection of digitized US public radio and television — holds decades of non-English programming in many under-resourced languages; Yup'ik and Samoan are two for which the holdings are both substantial and tied to a concrete access need, comprising long-running Yup'ik public-television programming from Alaska and Samoan-language broadcasting from American Samoa. For the speaker communities that produced this material, language is the primary axis of access: without language-level metadata the recordings remain effectively unfindable to the people they most concern. Supplying that metadata reliably is the role of LID here, and doing so responsibly requires sustained outreach with those communities, who are to lead the larger-scale annotation effort that this pilot is designed to bootstrap (Section 7). This effort is one part of a broader program to make US public-broadcast collections more accessible to the communities they document; a parallel line of work prompts vision–language models to extract structured entities, including Hawaiian personal names, from on-screen television chyrons (Lynch et al., 2026). Neither language is served by existing tooling: each is absent from or marginal in every off-the-shelf LID system we surveyed (Table 1). The pairing is also methodologically useful, because the two languages sit at near-opposite poles of phonological typology (Section 2): a single system that must separate both from English and from their respective family neighbors is a demanding testbed for the phonologically-informed direction we pursue.

This paper makes four contributions. First, we describe a pipeline for coarse segment-level LID annotation over long-form broadcast material, designed to be executable by non-native researchers on a short timeline without sacrificing label reliability. Second, we report preliminary corpus statistics for a small in-house annotated set covering both target languages. Third, we benchmark publicly available LID systems on this set, documenting how off-the-shelf models perform when the target language is either absent from or severely under-represented in the training label space. Fourth, we train a simple MLP head on top of two frozen pretrained speech encoders (XLS-R-300M and Whisper-medium) as prototype classifiers, showing that even minimal in-domain supervision substantially improves Yup'ik recall over zero-shot MMS-LID, and that the choice of pretrained encoder is itself a major lever at this scale. The results motivate a phonologically-informed fine-tuning direction, which we sketch in Section 7.

2 Target Languages

Central Alaskan Yup'ik and Samoan occupy maximally distant points in several dimensions of phonological typology. Table 2 summarizes the contrasts we return to in Section 7.

Yup'ik, an Eskimo-Aleut language spoken in western Alaska, exhibits a relatively rich consonantal inventory including velar and uvular consonants, voiceless sonorants, consonant gemination, and permissive consonant clustering (Jacobson, 1995). Its phonology additionally includes a four-vowel system with contrastive vowel length and an iambic prosodic system associated with systematic consonant lengthening (Woodbury, 1987). These properties produce dense consonantal transitions and distinctive durational patterns in running speech.

In contrast, Samoan, a Polynesian language of the Austronesian family, has a much smaller consonant inventory and highly restrictive syllable structure. Samoan syllables are strictly of the form (C)V, disallowing consonant clusters and codas entirely. The language additionally exhibits phonemic vowel length and a phonemic glottal stop, yielding speech with high vocalic density and relatively regular rhythmic structure (Mosel and Hovdhaugen, 1992).

These typological contrasts are directly relevant to LID. Yup'ik contains several acoustically salient but cross-linguistically rare features, including velar and uvular consonants as well as voiceless nasals, while Samoan is characterized by highly regular vowel-rich phonotactics and minimal consonant complexity. The two languages also differ substantially from English along multiple dimensions of syllable structure and segment inventory, suggesting that phonologically-informed representations may provide useful inductive bias for low-

Feature	Yup'ik	Samoan	English
Vowel inventory	4	5+length	11+
Consonant inventory	∼25	∼13	∼24
Velar and uvular consonants	✓		
Voiceless nasals	✓		
Consonant clusters	✓		✓
Gemination	✓		
Syllable structure	CVC	(C)V	complex
Glottal stop phonemic		✓	

Table 2: Contrastive phonological features across the three languages in our LID task.

resource spoken LID.

3 Related Work

Our paper follows a line of AmericasNLP work that documents dataset construction for an individual under-resourced language as the central contribution, with modest or no experimental results. Reyes Pérez and García Zuñiga (2024)'s description of the curated Amuzgo dataset is a direct precedent in this tradition, pairing a detailed linguistic description with a multi-phase data collection and annotation pipeline. A long line of work on St. Lawrence Island Yupik (Schwartz et al., 2021; Chen et al., 2020) has produced textual corpora and morphological analyzers but no speech resources for LID evaluation.

On the methods side, phonologically-informed LID is an active area. Liu et al. (2022) unify acoustic-phonetic and phonotactic information into a single encoder (PHO-LID) and report over 40% relative improvement on AP17-OLR. Shahin et al. (2023) show that a two-stage pipeline (pretrain on manner and place of articulation detection, then fine-tune on LID) improves code-switching LID by 5–6% relative. Universal phone recognizers such as Allosaurus (Li et al., 2020) and Allophant (Glocker et al., 2023) provide a pathway to obtain phonetic pseudo-labels in any language, which can be mapped to articulatory feature vectors via PanPhon (Mortensen et al., 2016) and PHOIBLE (Moran and McCloy, 2019). We leverage these directions in Section 7.

4 Data and Annotation

4.1 Source material

We take data from the AAPB Online Reading Room,[2] a publicly available collection of digitized public radio and television content from stations across the United States. The Yup'ik material originates with KYUK, a public broadcasting station in western Alaska that has produced Yup'ik television news and programming since the 1970s; the recordings used here are video, and contain a mixture of fully Yup'ik segments, English segments, and code-switching within single broadcasts. The Samoan material comes from KVZK, the public television broadcaster in American Samoa, and covers Samoan-language news and cultural programming. The archive spans several decades and includes both edited broadcast output and raw field recordings, with substantial variation in audio quality. All materials used in this pilot study are accessed under the AAPB's existing access and use policies; we do not redistribute the source audio.

4.2 Annotation protocol

Our annotation protocol is designed for a short timeline with in-house researcher annotators rather than expert native speakers. Two observations justify this choice. First, programming from Alaska in particular has a high base rate of Yup'ik in its non-English content: the overwhelming majority of non-English speech in this broadcast context is Yup'ik rather than another language. Similarly, the American Samoa-source material is predominantly Samoan and English. In both cases, program metadata (titles, descriptions) and audiovisual contexts provide additional disambiguation. Second, our target label set is deliberately coarse. We annotate at the segment level with five labels: `english`, `esu`, `smo`, `mixed`, and `other`. The `other` category includes non-speech material and unintelligible audio. This coarse label set is sufficient to train and evaluate a LID system while being reliably producible by a non-native researcher annotator working from a video source with context.

The annotation pipeline is as follows:

1. Voice activity detection and segmentation to strip non-speech regions and produce speech chunks suitable for annotation.

2. Manual segment-level labeling using an in-house annotation interface, with video context available to the annotator throughout.

3. *Dual annotation*: each segment in the current evaluation set is labeled independently by two authors, and inter-annotator agreement is reported in Section 4.4.

[2] https://americanarchive.org/

4.2.1 Labeling guidelines

Broadcast material mixes languages in structured ways that our coarse label set must handle consistently across annotators. We adopt the following rules, finalized after an initial annotation pilot:

- **Dominant-language rule.** If a segment contains speech in more than one language but one language clearly dominates, the segment receives the dominant language's label rather than `mixed`. A segment is labeled `mixed` only when the two languages are roughly balanced or alternate in genuine code-switching.
- **News broadcast convention.** News segments in which a target-language anchor delivers content first and an English broadcaster restates the same content afterwards are labeled with the target language (`esu` or `smo`), not `mixed`, since the English portion is a repetition rather than parallel code-switching.
- **Proper nouns.** Proper nouns (personal names, place names, organizations) embedded in otherwise-monolingual speech do not trigger the `mixed` label; the segment takes the label of the surrounding discourse.
- **Numbers and borrowed English vocabulary.** Yup'ik broadcasters frequently read numbers, weather figures, and commercial catch report data in English inside otherwise-Yup'ik speech. We treat such stretches as `esu` under the dominant-language rule, because the English tokens are routine lexical borrowings rather than genuine code-switching.

These rules are encoded in the annotator guidelines we plan to hand off to native-speaker annotators in subsequent work.

4.3 Corpus statistics

Table 3 reports corpus statistics for the current annotated set, using the primary annotator's labels as gold.

4.4 Inter-annotator agreement

A second author independently annotated 10 of the 15 videos (6 Yup'ik, 4 Samoan), producing 1,342 dual-annotated segments covering 9.3 hours of audio. Table 4 reports the confusion matrix between the two annotators. Raw agreement is 75.3% and Cohen's κ is 0.67, indicating substantial agreement.

	Yup'ik	Samoan
Videos annotated	9	6
Total duration	5.1 h	5.1 h
Total segments	589	1,189
labeled `esu`/`smo`	283	685
labeled `eng`	238	308
labeled `mixed`	53	122
labeled `other`	15	74

Table 3: Corpus statistics for the in-house annotated set (primary annotator). Segments labeled `other` (non-speech) are excluded from all evaluations.

	Annotator 2				
Annotator 1	`eng`	`esu`	`mix`	`oth`	`smo`
`eng`	249	2	39	6	11
`esu`	10	112	6	5	0
`mixed`	39	17	113	0	29
`other`	6	4	0	83	26
`smo`	35	0	49	43	458

Table 4: Inter-annotator confusion matrix over 1,342 dual-annotated segments (raw agreement 75.3%, Cohen's $\kappa = 0.67$).

The primary source of disagreement is the `mixed` label: 179 of 327 disagreements (55%) involve one annotator choosing `mixed` and the other choosing a specific language or `other`. This is consistent with the inherent subjectivity of the dominant-language rule (Section 4.2.1), which requires a judgment call on whether code-switching is "genuine" or incidental. Agreement on the two target-language labels is considerably higher: Yup'ik segments agree 84% of the time (112/133) and Samoan segments agree 78% (458/585). These per-class agreement rates support the reliability of the gold labels used in the MMS-LID and prototype evaluations (Sections 5–6).

5 Benchmarking Off-the-Shelf LID

We evaluate the three MMS-LID variants that cover at least one target language (Table 1) on our annotated evaluation set, excluding segments labeled `other` (non-speech and unintelligible audio). Whisper, VoxLingua107, and ESPnet OWSM cover neither target language and are omitted from this evaluation. Table 5 reports per-class precision, recall, and F1 for the target-language and English labels within each language group. The `mixed` class is never predicted by any MMS-LID model (which outputs a single language label) and therefore has zero recall across the board; we omit it from the table.

System	Class	Yup'ik videos P	R	F1	Samoan videos P	R	F1
mms-lid-1024	eng	.96	.96	.96	.94	.65	.77
	target	—	—	—	.92	.57	.71
mms-lid-2048	eng	.96	.93	.94	.95	.52	.67
	target	—	—	—	.93	.58	.72
mms-lid-4017	eng	.98	.92	.95	.97	.50	.66
	target	.85	.41	.55	.98	.61	.75

Table 5: Zero-shot per-class precision (P), recall (R), and F1 for the MMS-LID model family on our annotated evaluation set, broken down by language group. "target" = esu for Yup'ik, smo for Samoan. Dashes mark models that cannot emit the target label. Segments labeled other and mixed are excluded.

Gold Label	Predicted Label	Count
esu	ess	111
esu	esu	110
esu	esk	16
esu	esi	7
smo	smo	390
smo	haw	128
smo	ton	93
smo	jav	6

Table 6: Top predicted labels for gold Yup'ik (esu) and Samoan (smo) clips under MMS-LID-4017. Most errors remain within related language families.

Finding 1: Samoan recall is moderate but precision is high. All three MMS-LID variants that cover Samoan achieve high precision on smo (0.92–0.98) but only moderate recall (0.57–0.61): the models recognize Samoan when they predict it, but frequently assign related Polynesian labels, primarily Hawaiian (haw) and Tongan (ton), to segments that are in fact Samoan. This Polynesian confusion is expected given that these languages share much of their phonological profile (small inventories, strict CV syllable structure, phonemic vowel length). Performance does not improve with model size, suggesting the confusion is driven by acoustic similarity rather than label-space sparsity.

Finding 2: Yup'ik is systematically confused with other Yupik varieties. Only mms-lid-4017 can emit the esu label. It achieves high precision (0.85) but low recall (0.41): the majority of Yup'ik segments are assigned to Central Siberian Yupik (ess) or Northwest Alaska Inupiatun (esk) instead. This is not a general-purpose model failure: Central Alaskan and Central Siberian Yupik are close relatives within the Yupik subgroup of Eskimo-Aleut, sharing most of their segmental inventory. Table 6 shows that the most frequent errors concentrate on genealogically related languages. The confusion is best read as a linguistically principled limitation that the aggregate MMS-LID label space does not resolve. It also constitutes direct motivation for phonologically-informed modeling: the two varieties differ in specific features (e.g., certain uvular realizations, subsets of the voiceless sonorant series, and prosodic detail) that an articulatory-aware encoder could in principle exploit. We revisit this in Section 7.

Finding 3: English recall degrades in Samoan-source videos. English is recognized with 92–96% recall in Yup'ik-source videos but only 50–65% in Samoan-source videos. We attribute this to the Samoan material containing more code-switching and borrowed English vocabulary, which MMS-LID tends to assign a Polynesian label rather than English.

6 Prototype Classifier

To test how much of the off-the-shelf performance gap can be closed without fine-tuning the speech encoder, we train a lightweight MLP classification head on top of a *frozen* encoder and repeat the experiment with a second frozen encoder of comparable scale. We compare XLS-R-300M (Conneau et al., 2020), a self-supervised wav2vec 2.0 model, against the encoder of Whisper-medium (Radford et al., 2023), a multilingual ASR model whose encoder has about 307M parameters. Both produce 1024-dimensional hidden states. For each segment we extract mean-pooled final-layer activations (over the audio-valid frames in the Whisper case, whose input is padded to 30 seconds) and train a two-layer MLP (1024→256→4, with ReLU and dropout) on those fixed features. Holding the head, the data, and the recipe constant lets the comparison isolate two effects: the value of in-domain supervision relative to zero-shot MMS-LID, and the choice of pretrained encoder.

The classifier is trained with cross-entropy over four labels {eng, esu, smo, other}: segments labeled mixed in the annotation are collapsed into the target language of their source video (Section 4.2.1). We evaluate with 15-fold leave-one-video-out cross-validation across all Yup'ik and Samoan videos combined; each fold trains on 14

Encoder	Class	P	R	F1
XLS-R-300M	eng	.68	.75	.71
	esu	.79	.85	.82
	smo	.80	.69	.74
Whisper-medium	eng	.93	.83	.88
	esu	.96	.92	.94
	smo	.91	.91	.91

Table 7: Prototype classifier with a trainable MLP head on top of two frozen encoders, evaluated with 15-fold leave-one-video-out cross-validation. Overall accuracy / macro F1 are .74 / .76 for XLS-R-300M and .89 / .91 for Whisper-medium. Segments labeled `other` and `mixed` are excluded from the reported metrics; `mixed` is collapsed into the video's target language during training.

videos and evaluates on the held-out video. Table 7 reports per-class precision, recall, and F1 for both encoders, excluding the `other` class to match the MMS-LID evaluation in Table 5.

Comparison with MMS-LID, and across encoders. Despite using only frozen encoders, both prototypes substantially outperform zero-shot MMS-LID-4017 on Yup'ik: XLS-R-300M lifts `esu` recall from 0.41 to 0.85 (F1 0.55→0.82), and Whisper-medium lifts it further to 0.92 (F1 0.94). MMS-LID's principal Yup'ik failure mode — the `esu`/`ess` confusion in Table 6 — is absent here by construction, since our head emits only the four annotation labels; the relevant question is how cleanly each encoder's representations separate those four classes, and the gap between XLS-R and Whisper answers it. For Samoan, where MMS-LID was already moderately strong (F1 0.75), XLS-R offers parity (F1 0.74); Whisper-medium reaches F1 0.91, with the off-diagonal `smo`→`eng` count dropping from 162 under XLS-R to 19. English precision follows the same pattern: 0.68 with XLS-R, recovering to 0.93 with Whisper, comparable to MMS-LID's 0.97–0.98. At this parameter scale, the choice of pretrained encoder dominates the supervision recipe: a model whose pretraining objective rewards phonetic discriminability across many languages (Whisper, multilingual ASR) yields markedly more LID-discriminable representations than one trained by self-supervised reconstruction (XLS-R).

Encoder scale. We also ran the same head on top of Whisper-large-v3, roughly twice the encoder parameter count of Whisper-medium: overall macro F1 is essentially identical (0.91 vs. 0.91), and per-class scores move within noise. We speculate that at this scale of supervision and label granularity — about 1.6k labeled clips and four coarse classes — the bottleneck has shifted away from encoder capacity, and that the additional parameters in large-v3 chiefly refine fine-grained acoustic detail that a coarse four-class LID head cannot exploit.

Per-video variance. Two Yup'ik folds stand out as hard cases under XLS-R-300M (Folds 4 and 8 in the LOO-CV, accuracies 0.52 and 0.60 against the Yup'ik-group average of 0.87). Switching to Whisper-medium resolves one of them: Fold 8, a raw field recording with markedly noisier audio than the surrounding broadcast segments, climbs to 0.94, while Fold 4, an edited broadcast clip set, remains the hardest at 0.59 even with the stronger encoder. The split points to two distinct sources of difficulty — acoustic distance from the encoder's pretraining distribution, which a stronger encoder can absorb, and fold-specific factors that persist across encoders and warrant further investigation — and further motivates encoder adaptation (Section 7).

Pilot-scale caveat. The headline macro F1 of 0.91 under Whisper-medium should be read against the scale of the supporting evidence: 15 broadcast videos, about 1.6k labeled segments, and one held-out video per fold. The picture is likely to shift once the larger native-speaker-led annotation effort (Section 7) yields a more domain-diverse training and evaluation set, and the numbers reported here should be treated as pilot-scale evidence rather than settled levels.

7 Discussion and Future Work

7.1 Discussion

The benchmark results in Section 5 reveal a shared failure mode across both target languages: MMS-LID achieves high precision but low recall, because it systematically confuses target-language segments with closely related languages within the same family: Hawaiian and Tongan for Samoan, Central Siberian Yupik and Northwest Alaska Inupiatun for Yup'ik. These are not random misclassifications: they reflect genuine acoustic similarity within language subgroups that a model trained on broad typological coverage cannot fully resolve.

To probe whether these confusions align with broad phonological tendencies, we ran the universal phone recognizer Allosaurus over the annotated clips and computed two simple phone-level proxies

Gold label	Vowel ratio	Uvular-like ratio
eng	0.409	0.018
esu	0.441	0.051
smo	0.516	0.023

Table 8: Exploratory phone-level statistics computed from Allosaurus pseudo-phone sequences. Samoan exhibits the highest vowel-token ratio, consistent with its strict (C)V syllable structure, while Yup'ik shows the highest rate of uvular-like phones, consistent with its documented velar and uvular inventory.

from the resulting pseudo-phone sequences. These outputs are not treated as gold phonetic transcriptions, but as approximate acoustic-phonetic representations suitable for exploratory analysis. Table 8 summarizes two aggregate measures by gold label: vowel-token ratio and uvular-like phone ratio.

The resulting distributions align with the typological contrasts outlined in Section 2. Samoan clips exhibit the highest vowel-token ratio, consistent with the language's strict (C)V syllable structure and high vocalic density, while Yup'ik clips show the highest rate of uvular-like phones, consistent with its documented velar and uvular inventory. These exploratory results support the interpretation that the MMS-LID confusions are phonologically structured rather than random.

This pattern directly motivates a phonologically-informed approach to LID of closely related varieties. For example, Central Alaskan Yup'ik shares substantial similarities with other Yupik varieties, but comparative descriptions between Central Alaskan and Central Siberian Yupik also note several systematic differences, such as retroflex segments reported for Central Siberian Yupik that are not typically reported in Yup'ik, and differences in gemination and prosodic timing (Section 2) (Jacobson, 1990). Features like these may help reduce the frequent confusions observed in our experiments. Similarly, although Samoan, Hawaiian, and Tongan share many phonological properties, finer-grained differences in segment inventories and specific prosodic patterns such as tonal marking of absolutive case (Yu, 2021) may support improved discrimination within the language family. These are precisely the dimensions that a loss designed around articulatory targets can steer an encoder toward. The prototype results in Section 6 already provide indirect empirical support: under the same MLP head, Whisper-medium (multilingual ASR pretraining) substantially outperforms XLS-R-300M (self-supervised reconstruction), suggesting that pretraining objectives rewarding phonetic discriminability across languages already carry much of the inductive bias LID benefits from.

We sketch a two-stage phonologically-informed fine-tuning pipeline as the direction of our continuing work. Stage 1 follows Shahin et al. (2023): the wav2vec 2.0 encoder is first fine-tuned on articulatory feature classification (manner and place of articulation) with pseudo-labels obtained from universal phone recognizers such as Allosaurus or Allophant, mapped to feature vectors via PanPhon. This requires no target-language transcription. Stage 2 fine-tunes the resulting articulatory-aware encoder for LID on the annotated broadcast data, optionally augmenting the classification loss with a regularization term that encourages representations whose inter-language distances correlate with PHOIBLE inventory distances. Recent work (Choi et al., 2026) has shown that self-supervised speech models already encode distinctive features as linear subspaces in their hidden states; the proposed fine-tuning should sharpen exactly the dimensions that are discriminative within the Yupik subgroup, and should help most in the short-segment regime where aggregate statistical LID degrades.

7.2 Ongoing Annotation and Future Directions

Beyond the modeling direction, immediate next steps include transitioning primary annotation to trained native-speaker annotators, who will expand the dataset over the coming months; expanding the label set to include code-switching boundaries and speaker turns; testing the MMS-LID Samoan result under noisier, non-studio recording conditions; and adding speaker and recording-era metadata to support fairness evaluation across the multi-decade archive.

Since the initial submission of this paper, a larger-scale annotation effort has begun with community language fellows working directly on Yup'ik and Samoan broadcast materials. These annotators are contributing segment-level review, correction, and expansion of the coarse-label dataset using the same annotation interface and guidelines described in Section 4. Their ongoing work will substantially expand both the scale and linguistic reliability of the corpus, and will enable future evaluation under native-speaker supervision.

8 Conclusion

We present a pilot spoken language identification benchmark for Central Alaskan Yup'ik and Samoan using archival broadcast recordings. Experiments with off-the-shelf multilingual LID systems show that exact-label recognition remains challenging, with many errors concentrated within the related language families. Lightweight supervised adaptation substantially improves performance, and the choice of frozen pretrained encoder is itself a major lever: under an otherwise identical head, an ASR-pretrained encoder yields substantially fewer errors than a self-supervised one of comparable scale. Future work will expand annotation, incorporate native-speaker review and explore phonologically informed modeling approaches.

Limitations

The annotated set reported in this paper is small (9 Yup'ik and 6 Samoan videos) and was produced by in-house researcher annotators rather than native speakers. A deliberately coarse five-label scheme and high base rate of Yup'ik and Samoan in non-English content mitigate annotator noise, and dual annotation yields substantial agreement ($\kappa = 0.67$), but native-speaker validation is required before strong claims can be made about absolute label quality. The dominant source of annotator disagreement is the `mixed` label, which is inherently subjective; the target-language labels themselves agree at 78–84%. The prototype classifiers are trained and evaluated on a dataset far smaller than typical low-resource LID settings; the reported numbers, including the Whisper-medium macro F1 of 0.91, should be read as pilot-scale evidence and may not generalize to a larger or more domain-diverse set. The multi-decade time span of the broadcast source introduces acoustic variability (analog vs. digital capture, channel and codec differences) that we do not yet explicitly model. Finally, the phonologically-informed approach in Section 7 is proposed, not yet implemented or evaluated.

Acknowledgments

We thank KYUK in western Alaska and KVZK in American Samoa for providing access to the broadcast recordings that made this study possible. These recordings were provided through the American Archive of Public Broadcasting [3], a collaboration between GBH and the U.S. Library of Congress. This research was supported in part by Andrew W. Mellon Foundation.

[3] https://americanarchive.org/

References

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved finite-state morphological analysis for St. Lawrence Island Yupik using paradigm function morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2676–2684.

Kwanghee Choi and 1 others. 2026. Self-supervised speech models discover phonological vector arithmetic. *Computing Research Repository*, arXiv:2602.18899.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. volume arXiv:2006.13979.

Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. Allophant: Cross-lingual phoneme recognition with articulatory attributes. In *Proceedings of Interspeech 2023*.

Steven A. Jacobson. 1990. Comparison of central alaskan yup'ik eskimo and central siberian yupik eskimo. *International Journal of American Linguistics*, 56(2):264–286.

Steven A. Jacobson. 1995. *A Practical Grammar of the Central Alaskan Yup'ik Eskimo Language*. Alaska Native Language Center, University of Alaska Fairbanks.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of ICASSP 2020*.

Hexin Liu, Leibny Paola Garcia Perera, Andy W. H. Khong, Suzy J. Styles, and Sanjeev Khudanpur. 2022. PHO-LID: A unified model incorporating acoustic-phonetic and phonotactic information for language identification. In *Proceedings of Interspeech 2022*.

Kelley Lynch, Owen King, Kyeongmin Rim, Gabrielle Keen, Yangyang Chen, and James Pustejovsky. 2026. Structured entity extraction from Hawaiian television chyrons using vision-language models. In *Proceedings of the SIGUL 2026 Workshop: Towards Inclusivity and Equality — Language Resources and Technologies for Under-Resourced and Endangered Languages*, Palma, Mallorca, Spain. Joint workshop with ELE, EURALI, and DCLRL, co-located with LREC 2026.

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016*.

Ulrike Mosel and Even Hovdhaugen. 1992. *Samoan Reference Grammar*. Scandinavian University Press, Oslo.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-weon Jung, Soumi Maiti, and Shinji Watanabe. 2024. OWSM v3.1: Better and faster open Whisper-style speech models based on E-Branchformer. In *Proceedings of Interspeech 2024*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. volume arXiv:2305.13516.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Antonio Reyes Pérez and Hamlet Antonio García Zuñiga. 2024. From field linguistics to NLP: Creating a curated dataset in Amuzgo language. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 127–131.

Lane Schwartz, Emily Chen, Hyunji Hayley Park, Edward Jahn, and Sylvia L.R. Schreiner. 2021. A digital corpus of St. Lawrence Island Yupik. In *Proceedings of the Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Mostafa Shahin, Zheng Nan, Vidhyasaharan Sethu, and Beena Ahmed. 2023. Improving wav2vec2-based spoken language identification by learning phonological features. In *Proceedings of Interspeech 2023*.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: A dataset for spoken language recognition. In *Proceedings of IEEE SLT 2021*.

Anthony C. Woodbury. 1987. Meaningful phonological processes: A consideration of Central Alaskan Yupik Eskimo prosody. *Language*, 63(4):685–740.

Kristine M. Yu. 2021. Tonal marking of absolutive case in Samoan. *Natural Language & Linguistic Theory*, 39:291–365.