Automatic Language Identification of Indigenous Mexican Languages: A Transfer Learning Approach for Nahuatl and Totonaco

Ricardo Alvarez-Sanchez[1], Geraldine Lomeli[1], David Pinto[1]*, Fernando Perez-Tellez[2], Enrique Varela[1], and Sofía Paniagua Rivera[1]

[1] Benemérita Universidad Autónoma de Puebla, Puebla, Mexico
as224470485@alm.buap.mx, lp224470490@alm.buap.mx, david.pinto@correo.buap.mx, enrique.varela@correo.buap.mx, sofiapaniagua@gmail.com
[2] Technological University Dublin, Dublin, Ireland
Fernando.PerezTellez@tudublin.ie

Abstract. Nahuatl and Totonaco are among Mexico's 68 legally recognized indigenous languages, yet both remain severely under-resourced digitally, restricting community access to public services. This paper presents a Transfer Learning approach for Automatic Language Identification (LID) between these two Mesoamerican languages by fine-tuning Wav2Vec 2.0 XLS-R, a cross-lingual self-supervised model pre-trained on 128 languages, on the publicly available TonalliCorpus (72 field-recorded interviews, 24 hours, covering five Nahuatl and four Totonaco regional variants). A data-explosion segmentation strategy converts long-form recordings into over 14,000 Transformer-ready 5-second chunks, while a speaker-independent split ensures generalization to unseen speakers. The convolutional encoder is frozen during fine-tuning to preserve pre-trained acoustic representations; only the Transformer layers and a lightweight classification head are updated. The model achieves 99.22% accuracy on the held-out test set and a 5-fold cross-validation mean of 99.19% $\pm$ 0.13%, confirming both effectiveness and stability. An analysis of the results reveals a minor Totonaco recall asymmetry ($\Delta = 0.02$) attributable to corpus class imbalance. These results establish a reproducible LID benchmark and a foundation for downstream indigenous speech technologies.

Keywords: Deep Learning · Language Identification · Transfer Learning · Fine-Tuning· Wav2Vec 2.0 · XLS-R · Nahuatl · Totonaco · Low-Resource Languages · Responsible AI

1 Introduction

1.1 Background, Motivation, and Research Question

Nahuatl (Uto-Aztecan stock, ≈2M speakers [7]) and **Totonaco** (Totonacan-Tepehua family [13]) are among the 68 indigenous languages officially recognised in Mexico [8]. Despite their profound cultural importance as carriers of

oral histories, cosmovision, and community identity, both languages are severely under-resourced in the digital domain [17]. The General Law on Linguistic Rights of Indigenous Peoples [4] mandates that these languages be valid in any public or legal procedure, yet the technological infrastructure required to support such a mandate, such as automatic language routing, Automated Speech Recognition (ASR), and machine translation, is largely absent. This institutional gap is not merely an inconvenience; it actively restricts indigenous communities' access to healthcare, education, legal representation, and digital participation [17]. Automatic Language Identification (LID) is the foundational prerequisite for all such downstream applications, making it the natural first target for computational efforts in this space.

The linguistic challenges are substantial and multi-layered. Both languages are morphologically rich (Nahuatl in particular is agglutinative and polysynthetic [5,11], encoding complex grammatical relationships within single long words), which means that a single spoken utterance may compress information that would require an entire sentence in an analytic language such as English. At the acoustic level, this morphological complexity increases the effective phonetic inventory that any model must implicitly learn. Additionally, conventional text-processing tools such as Byte Pair Encoding (BPE) [18] and overlap-based evaluation metrics such as BLEU [16] are ill-suited to these languages, a limitation that motivates the use of raw-audio self-supervised models that bypass text representations entirely. The central research question of this study is therefore:

> *How effectively can a Transfer Learning framework based on cross-lingual self-supervised representations (Wav2Vec 2.0 XLS-R) be fine-tuned on a limited, regionally-diverse corpus to achieve high-accuracy LID between Nahuatl and Totonaco in non-studio acoustic environments?*

1.2 Related Work and Technical Foundation

Low-Resource NLP for Indigenous Languages of the Americas. The Americas host approximately 86 language families, the vast majority of which remain computationally unexplored [6]. The scarcity of parallel and monolingual data is the central bottleneck: without sufficient text or audio corpora, even the most powerful architectures cannot acquire language-specific representations. Shared tasks such as AmericasNLP [12] have galvanised the community around low-resource machine translation, with multilingual pre-trained models such as mBART [9] demonstrating that pre-training on high-resource languages followed by fine-tuning on target LRLs is a viable and sample-efficient strategy. These text-based findings directly motivate the analogous speech approach taken in this work. While these text-based findings demonstrate that transfer learning is viable for indigenous languages, there remains a critical gap in applying these methodologies directly to raw acoustic data for Mesoamerican languages, which this study addresses.

Self-Supervised Learning for Speech. The paradigm shift for low-resource speech processing was catalysed by **Wav2Vec 2.0** [2]. The model passes raw waveforms through a multi-layer convolutional feature encoder that produces a sequence of latent representations at a resolution of ≈25ms per frame. These representations are then quantised into a finite codebook of learned speech units and fed into a deep Transformer network. Training proceeds by masking spans of the latent representation and requiring the Transformer to identify the correct quantised unit from a set of distractors, a contrastive self-supervised objective that requires no transcriptions whatsoever. The result is a model that, even when fine-tuned on as little as ten minutes of labelled audio, achieves competitive Word Error Rates on standard benchmarks [2]. This makes it uniquely suited to the LRL setting.

Its cross-lingual successor, **XLS-R** [1], scales this approach to 128 languages by training on hundreds of thousands of hours of publicly available multilingual speech (VoxPopuli, MMS, CommonVoice, and others). The resulting model learns a shared phonetic representation space that generalises across typologically diverse languages, including those absent from the pre-training data. XLS-R has established state-of-the-art results on multiple LID and ASR benchmarks for under-resourced languages, and fine-tuning it with a frozen convolutional encoder is the current best-practice strategy for LRL speech tasks [19]. This study applies and validates that strategy for two Mesoamerican languages that are not represented in the XLS-R pre-training corpus.

This article presents a high-accuracy Automatic Language Identification (LID) system for Nahuatl and Totonaco using the Wav2Vec 2.0 XLS-R model. Following an Introduction on the digital divide of indigenous languages, the Methodology details a transfer learning strategy and audio segmentation approach tailored for low-resource data. The Experiments and Results demonstrate a robust 99.22% accuracy, validated through speaker-independent cross-validation. Finally, the Discussion and Conclusion address service equity considerations regarding class imbalance and propose extending the framework to regional dialect identification and downstream speech technologies.

2 Methodology

2.1 Dataset and Preprocessing

All audio data used in this study come from the publicly available **TonalliCorpus** [15], a multilingual parallel corpus of Nahuatl and Totonaco collected through field recordings in classroom environments across the Sierra Nororiental of the State of Puebla, Mexico. The corpus provides 72 classroom-recorded interviews (44 Nahuatl, 28 Totonaco), spanning five Nahuatl and four Totonaco regional variants. Each interview is approximately 20 minutes long. No private or proprietary recordings were used in this study.

Because Transformer architectures are constrained by sequence length and GPU memory, raw 20-minute waveforms are unsuitable for direct ingestion. To

overcome this and to maximise data utility, a critical concern in any LRL setting, a *data-explosion* segmentation strategy was applied. All MP3 files were first decoded and resampled to 16 kHz, the sampling rate expected by the Wav2Vec 2.0 feature extractor [2]. Each interview was then systematically sliced into non-overlapping **5.0-second** chunks. Audio was sliced sequentially without explicit Voice Activity Detection (VAD) or silence filtering, meaning segments may contain natural pauses, classroom background noise, or overlapping speech. This segment length was chosen deliberately: segments shorter than ≈3 seconds risk discarding the prosodic and suprasegmental cues (stress patterns, tone, intonation) that are highly discriminative between Nahuatl and Totonaco, while segments longer than 10 seconds exceed GPU memory budgets at reasonable batch sizes. At 5 seconds, each chunk preserves sufficient phonetic and prosodic context while remaining computationally tractable. Each segment was subsequently normalised to zero mean and unit variance to mitigate the volume variability and low-level noise differences inherent to classroom field recordings.

The resulting corpus comprises **11,360 training chunks** and **3,074 test chunks**. To prevent data leakage and ensure a realistic estimate of generalisation, the train/test split was applied at the interview level (80%/20%). This speaker-independent partitioning guarantees that no audio segment from any individual appears in both partitions, directly measuring the model's ability to handle previously unseen speakers, an essential property for any real-world deployment scenario.

2.2 Model Architecture and Fine-Tuning Strategy

The Wav2Vec 2.0 XLS-R architecture [1] consists of three components: (i) a **convolutional feature encoder** (seven convolutional blocks that project raw waveforms into 512-dimensional latent frame representations at 20ms intervals); (ii) a **Transformer context network** (a deep stack of multi-head self-attention layers, with the 300M-parameter variant used here having 24 layers, that contextualise these frame representations across the full segment); and (iii) a **linear classification head**, added specifically for this task, that projects the mean-pooled final hidden state onto two output classes (Nahuatl / Totonaco) via Softmax. Training minimises the cross-entropy loss:

$$\mathcal{L} = -\sum_{c \in \{N,T\}} y_c \log \hat{p}_c \tag{1}$$

where y_c is the one-hot ground-truth label and $\hat{p}_c$ is the predicted probability for class c. Following established best practice [19], the convolutional feature encoder was **frozen** throughout training, preserving the low-level acoustic feature extraction capabilities acquired during large-scale multilingual pre-training and preventing catastrophic forgetting on the small LRL corpus. Only the Transformer layers and the classification head were updated. This selective fine-tuning is critical: the Transformer layers adapt their high-level contextual representations toward the phonological contrasts distinguishing Nahuatl and Totonaco,

while the frozen encoder continues to provide stable, high-quality acoustic features as input.

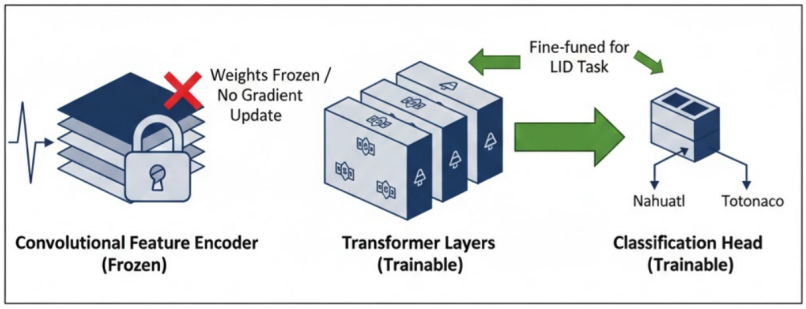

Fig. 1. Wav2Vec 2.0 XLS-R architecture for LID. The convolutional feature encoder (grey) is frozen throughout fine-tuning. The Transformer layers and the appended classification head (blue) are updated to distinguish Nahuatl and Totonaco.

2.3 Hyperparameter Configuration and Hardware

Training was executed on an NVIDIA A100 GPU (40 GB VRAM) [14]. Two hardware-specific optimisations were applied: Brain Float 16 (BF16) mixed precision, native to the A100's Ampere architecture, which accelerates matrix operations while preserving the dynamic range needed for stable deep Transformer fine-tuning; and TensorFloat-32 (TF32) for matrix multiplication, providing a further throughput increase with negligible numerical impact. Together, these reduced the total training time to ≈8 minutes for 3 epochs, compared to ≈1–2 hours on a conventional T4 GPU. An effective batch size of 64 was achieved through a physical batch size of 32 with 2 gradient accumulation steps, optimising VRAM utilisation while ensuring stable gradient estimates. The base learning rate of 3×10^{-4} was scheduled with cosine decay and a linear warm-up phase [10]: the warm-up prevents excessively large initial updates from destabilising the pre-trained weights, and the cosine schedule subsequently reduces the step size smoothly, which is particularly effective for fine-tuning scenarios where the model is close to a good solution from the start. Training was limited to **3 epochs**; as Table 1 shows, the validation loss had already converged to ≈0.04 by Epoch 3, confirming that additional epochs would risk overfitting rather than improving generalisation.

2.4 Ethical Considerations

All research adhered to the anonymisation and informed-consent protocols established by the TonalliCorpus [15]. No personally identifiable information (PII) was stored or processed during model training.

3 Experiments and Results

3.1 Training Convergence

Table 1 reports per-epoch training and validation metrics. Convergence was remarkably rapid: validation loss fell from ≈0.57 at Epoch 1 to ≈0.04 at Epoch 3 (a reduction of over 93%) while accuracy climbed from 75.3% to 99.2%. The sharp improvement between Epochs 1 and 2 (validation accuracy: 75.3% → 97.7%) indicates that the XLS-R representations already encoded highly transferable cross-lingual phonetic structure [1] prior to fine-tuning; the task-specific classification head required only minimal exposure to the target languages to align. The continued improvement from Epoch 2 to Epoch 3 hypothetically reflects the Transformer layers adapting to the acoustic and phonological boundaries between the classes, though further explainability studies are required to confirm the precise features driving this. The close tracking between training and validation loss across all epochs confirms that the speaker-independent split effectively prevented overfitting to speaker-specific artefacts.

Table 1. Training and validation metrics per epoch.

Epoch	Train Loss	Val. Loss	Accuracy
1	0.6598	0.5739	75.31%
2	0.1531	0.1022	97.69%
3	0.0237	0.0396	**99.22%**

3.2 Test-Set Performance and Confusion Matrix

The final model achieved **99.22**% mean accuracy on the 3,074 held-out test chunks (Table 2). Nahuatl recall reached almost 1.00 (virtually all Nahuatl segments were correctly identified with an almost zero false-negative rate), while Totonaco recall was 0.98, with 1.8% of Totonaco segments misclassified as Nahuatl. This minor asymmetry is consistent with the corpus class imbalance (44 Nahuatl vs. 28 Totonaco source interviews [15]) and is analysed further in Section 4. Importantly, the false-positive rate in the opposite direction is negligible (0.11% of Nahuatl segments misclassified as Totonaco), confirming that the model's slight bias operates in one direction only.

Table 2. Normalised confusion matrix on the held-out test set.

True \ Predicted	Nahuatl	Totonaco
Nahuatl	0.999	0.001
Totonaco	0.018	0.982

3.3 5-Fold Cross-Validation

To confirm that the reported test-set result is not an artefact of a single favourable train/test partition, a rigorous 5-fold cross-validation was conducted on the complete segmented corpus. Folds were constructed at the interview level, preserving strict speaker independence in every partition, and the full fine-tuning procedure, including hyperparameter settings and epoch count, was repeated identically for each fold.

Table 3. 5-fold cross-validation accuracy results.

Fold	**1**	**2**	**3**	**4**	**5**	**Mean ± Std**
Acc.	99.38%	99.12%	99.27%	99.01%	99.19%	$\mathbf{99.19\% \pm 0.13\%}$

The narrow accuracy range of 99.01%–99.38% and a standard deviation of 0.13% demonstrate that the high performance is a stable property of the methodology. Notably, even the worst-performing fold (Fold 4: 99.01%) substantially exceeds what a CNN-MFCC baseline would be expected to achieve on this noisy, limited corpus [6]. The cross-validation mean (99.19%) aligns closely with the independently obtained test-set accuracy (99.22%), providing dual confirmation of the result's reliability and the absence of distributional artefacts.

4 Discussion

4.1 Data Limitations and Generalisation Risks

Despite the high accuracy, several limitations of the corpus must be acknowledged. The class imbalance in the TonalliCorpus [15] (44 Nahuatl versus 28 Totonaco interviews) is the most likely explanation for the observed asymmetry between Nahuatl recall (1.00) and Totonaco recall (0.98). Although the segmentation strategy expands the effective sample count substantially, the underlying source-file imbalance persists and may cause the model to weight Nahuatl-specific acoustic patterns more heavily during gradient updates. Future corpus collection should target greater parity between classes.

Second, because all recordings originate from classroom settings, the model's performance under real-world acoustic conditions (e.g., outdoor markets, village squares, or telephone channels) remains an open question. Domain mismatch between controlled and unconstrained environments is a well-documented failure mode in speech models, and deployment robustness cannot be assumed from indoor test results alone. Furthermore, because silences and environmental noises were not filtered out during segmentation, there is a risk that the model might leverage acoustic artifacts specific to the recording environment (e.g., room acoustics, background classroom noise) rather than purely linguistic features. While the cross-validation confirms generalization to unseen speakers

within this corpus, testing on external datasets representing distinct acoustic domains (such as radio broadcasts or outdoor field recordings) is essential to verify true domain robustness.

Third, the binary LID framing does not probe the model's ability to discriminate among the five Nahuatl or four Totonaco regional variants [15]; the high overall accuracy may mask differential performance across dialects that would only become apparent in a finer-grained classification task.

4.2 Methodological Critique

Three aspects of the experimental design warrant critical reflection. First, the **dependency on pre-trained weights**: the model's performance is fundamentally contingent on the quality and coverage of XLS-R's pre-training corpus [1]. Since neither Nahuatl nor Totonaco appeared in that corpus, the model relies entirely on cross-lingual transfer from phonetically similar languages. Any systematic phonetic gap between those proxy languages and the target languages constitutes an unquantified source of risk. Second, the **absence of a CNN-MFCC baseline**: while comparable low-resource benchmarks suggest that such a baseline would achieve approximately 70–80% accuracy on noisy field data [6], the lack of a directly trained baseline prevents a precise quantification of the benefit attributable to the Transformer architecture and pre-trained representations specifically. The 5-fold cross-validation (Table 3) confirms stability but does not substitute for ablation. Including such a comparison is recommended for future work. Third, the **black-box nature of the Transformer**: the model provides no interpretable account of which acoustic features (place of articulation, tone, vowel harmony, prosodic rhythm) drive its decisions [3]. This limits the model's value as a tool for linguistic research and makes it difficult to anticipate failure modes in deployment.

4.3 Service Equity Considerations

The majority-class bias discussed above (Δrecall = 0.02) represents a quantifiable disparity that could translate into inequitable service quality if the model is deployed in a language-routing application: Totonaco speakers would experience a 2% higher misidentification rate than Nahuatl speakers [3]. While small in absolute terms, such disparities tend to compound across downstream tasks.

4.4 Contributions

This work makes four concrete contributions. (i) A validated, reproducible *data-explosion* preprocessing pipeline that transforms 72 long-form, non-studio field recordings from the TonalliCorpus [15] into over 14,000 Transformer-ready segments, directly addressing the data scarcity problem in LRL speech research. (ii) A high-accuracy LID benchmark (99.22% on the held-out test set; $99.19\% \pm 0.13\%$ via 5-fold cross-validation) for the regional variants of Nahuatl and Totonaco present in the Puebla State corpus, providing a reproducible reference

point for future work [5]. (iii) A hardware-aware training protocol leveraging BF16 and TF32 on the NVIDIA A100 [14] that reduces fine-tuning time from $\sim$2 hours to $\sim$8 minutes, making the approach accessible beyond large-scale compute environments. (iv) A rigorous evaluation framework based on strict speaker-independent splitting that can serve as a reusable template for reliable AI development in indigenous language contexts. The validated model and TonalliCorpus [15] together provide a ready foundation for downstream tasks including ASR, machine translation [9,12], and dialect-level identification.

5 Conclusion and Future Work

This research demonstrated that fine-tuning Wav2Vec 2.0 XLS-R on a small, publicly available corpus (TonalliCorpus [15]) is sufficient to achieve near-perfect Automatic Language Identification between Nahuatl and Totonaco. The model attained $\mathbf{99.22}\%$ test accuracy and a cross-validation mean of $\mathbf{99.19}\% \pm \mathbf{0.13}\%$, confirming both the effectiveness and stability of cross-lingual self-supervised transfer learning [1,19] for low-resource indigenous speech. The result directly supports the technological demands imposed by Mexico's General Law on Linguistic Rights [4] by providing a reliable, scalable LID component for routing and processing indigenous-language speech in public-sector applications.

Four immediate research directions follow from this work. First, **dialect identification**: extending the classification head to discriminate among the five Nahuatl and four Totonaco regional variants documented in the TonalliCorpus [15], which would require fine-grained dialect-level annotation as a prerequisite. Second, **robustness and fairness**: applying data augmentation techniques (noise injection, reverberation simulation, room impulse response convolution) to improve model performance under real-world acoustic conditions, and collecting additional recordings to balance the class representation and mitigate the allocative harm identified in Section 4. Third, **downstream integration**: repurposing the fine-tuned encoder's latent representations for ASR and machine translation systems [9,12], using this LID model as a first-stage language router, and integrating Explainable AI (XAI) techniques such as attention visualisation to expose the specific phonetic features driving classification decisions [3], bridging the gap between computational performance and linguistic insight. Fourth, **linguistic and baseline validation**: future work must include empirical comparisons against traditional acoustic baselines (such as MFCCs or Mel-spectrograms) to precisely quantify the benefit of the Transformer architecture. Additionally, leveraging forced-alignment techniques or incorporating the available Spanish translations as a comparison class could facilitate a deeper, purely linguistic analysis of the specific phonemes and pronunciation patterns driving the model's classifications, thereby avoiding experimental ambiguities.

Acknowledgments. This work was carried out with support from lacunafund.org and google.org.

Disclaimers. The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, or CENIA.

References

1. Babu, A., et al.: Xls-r: Self-supervised cross-lingual speech representation learning at scale. In: Interspeech (2022)
2. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
3. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in nlp. In: Proceedings of the 58th Annual Meeting of the ACL (2020)
4. Cámara de Diputados del H. Congreso de la Unión: Ley general de derechos lingüísticos de los pueblos indígenas. Diario Oficial de la Federación, Mexico (2003)
5. Gutierrez-Vasques, X., Mijangos, V.: Comparing morphological complexity of spanish, otomi and nahuatl. In: Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing (COLING 2018). Association for Computational Linguistics (2018)
6. Hedderich, M.A., Lange, L., Adel, H., Strötgen, J., Klakow, D.: A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (2021)
7. Instituto Nacional de Estadística y Geografía (INEGI): Censo de población y vivienda 2020: Lengua indígena. INEGI, Mexico (2021), `https://www.inegi.org.mx/programas/ccpv/2020/`
8. Instituto Nacional de Lenguas Indígenas (INALI): Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de méxico. Diario Oficial de la Federación, Mexico (2008)
9. Liu, Y., et al.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
10. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017)
11. Mager, M., Gutierrez-Vasques, X., Sierra, G., Meza-Ruiz, I.: Challenges of language technologies for the indigenous languages of the americas. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING) (2018)
12. Mager, M., Oncevay, A., Ebrahimi, A., et al.: Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. pp. 202–217 (2021)
13. McFarland, T.A.: The Phonology and Morphology of Filomeno Mata Totonac. Ph.D. thesis, University of California, Berkeley (2009)
14. NVIDIA Corporation: Nvidia a100 tensor core gpu architecture. Whitepaper, NVIDIA (2020), `https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf`
15. Pinto, D., Deance, I., Perez-Tellez, F., Paniagua, S., Durry, B., Giovanny, V., Techalotzi, A.: Tonallicorpus: Towards the construction of a multilingual parallel corpus of nahuatl and totonac variations for further ai applications. International Journal of Combinatorial Optimization Problems and Informatics (2026)

16. Reiter, E.: A structured review of the validity of bleu. Computational Linguistics **44**(3), 393–401 (2018)
17. Rosa, F.R.: From community networks to shared networks: The paths of latin-centric indigenous networks to a pluriversal internet. Information, Communication & Society **26**(11), 2326–2344 (2022)
18. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the ACL (2016)
19. Wang, Q., et al.: Transfer learning for speech recognition on low-resource languages. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022)