Spectral Robustness in Low-Resource Language Identification: A Comparative Deep Learning Study of Nahuatl and Mexican Spanish

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

[1] Benemérita Universidad Autónoma de Puebla, Puebla, Mexico
lp224470490@alm.buap.mx, as224470485@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. Language Identification (LID) is a fundamental precursor to multilingual Automatic Speech Recognition (ASR) systems. While high-resource languages benefit from large studio-quality corpora, indigenous languages such as Nahuatl rely on naturalistic field recordings suffering from environmental noise, reverberation, and channel variability, a "Channel Effect" where classifiers learn the recording environment rather than linguistic features. We present a rigorous evaluation of Convolutional Recurrent Neural Networks (CRNNs) for binary LID between Nahuatl and Mexican Spanish. Using a subset of the TonalliCorpus (44 field interviews) and the CIEMPIESS Light Spanish corpus, we investigate spectral bandwidth reduction (16 kHz vs. 4 kHz) and Instance Normalization as robustness interventions. Our central hypothesis posits that a 4 kHz low-pass filter removes high-frequency environmental artifacts while preserving core vowel formants essential for linguistic discrimination, yielding equivalent or superior accuracy to wideband inputs. 5-Fold Group Cross-Validation demonstrates near-perfect classification (>99%) across all configurations; the best model (4 kHz + Instance Normalization) achieves 99.51% accuracy with GradCAM-verified formant-based decision boundaries.

Keywords: Language Identification, Deep Learning, CRNN, Nahuatl, Indigenous Languages, Spectrogram Analysis, Spectral Filtering

1 Introduction

In Mexico, Nahuatl, a Uto-Aztecan language with over 1.5 million speakers, remains critically under-resourced [8]. Although recent initiatives such as the TonalliCorpus [13], a multilingual parallel corpus of Nahuatl and Totonaco variations designed for AI applications, represent important steps, large-scale annotated datasets suitable for training foundation models such as Wav2Vec 2.0 [2] or

Whisper [14] from scratch remain severely limited. LID is the gatekeeper of multilingual pipelines: it routes audio to the appropriate downstream acoustic model, and its failure breaks the entire communicative chain. For indigenous communities, LID is essential for archiving oral histories, monitoring radio broadcasts, and enabling accessibility tools.

1.1 The Acoustic Mismatch and Spectral Bandwidth

A pervasive challenge in low-resource LID is data heterogeneity. Standard Spanish corpora such as CIEMPIESS [6] are sourced from broadcast radio with high Signal-to-Noise Ratios (SNR), while Nahuatl datasets originate from unconstrained fieldwork in classrooms or community centers, introducing reverberation, environmental noise, and channel hiss. When a model is trained on such disparate data, it learns to detect the *acoustic environment* rather than linguistic features: the "Channel Effect."

We propose applying a low-pass filter with cutoff at 4 kHz. The high-frequency band (>4 kHz) in naturalistic field recordings is largely dominated by environmental artifacts rather than discriminative linguistic content. Restricting to the telephony band (300 Hz–3400 Hz, per ITU-T G.711 [9]) forces the model onto fundamental speech frequencies where vowel formants (F_1, F_2) and prosodic energy reside, which are the primary acoustic markers distinguishing these two languages.

1.2 Research Questions

- **RQ1 (Bandwidth Robustness):** Does reducing spectral bandwidth from 16 kHz to 4 kHz degrade or enhance classification robustness?
- **RQ2 (Normalization Dynamics):** How does Instance Normalization influence convergence and generalization under high amplitude variance?
- **RQ3 (Explainability):** Does GradCAM [16] confirm that the model's decision boundary aligns with linguistic formant structures rather than noise?

The remainder of this paper is structured as follows. Section 2 reviews related work on LID, indigenous language technologies, and spectral filtering. Section 3 provides the theoretical background on time-frequency analysis and CRNN components. Section 4 details the dataset, signal processing pipeline, model architecture, and validation protocol. Section 5 presents and discusses the experimental results, quantitative comparisons, explainability analysis, and practical implications. Section 6 concludes answering the research questions.

2 Related Work

2.1 Statistical and Deep Learning Approaches to LID

For nearly two decades, systems based on Gaussian Mixture Models (GMM) dominated LID. The i-vector framework [4] maps variable-length utterances to

fixed-dimensional vectors but struggles with short segments and channel mismatch. Deep Learning enabled end-to-end optimization: Convolutional Neural Networks (CNNs) applied to spectrograms leveraged ResNet architectures [5] to identify time-frequency texture patterns, while CRNNs [3] combined local spectral feature extraction (CNN) with temporal sequence modeling using Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) to capture both spectral shape and phonotactic rhythm. The x-vector framework [18] extended Deep Neural Network (DNN) embeddings to LID with strong results on NIST benchmarks. For low-resource scenarios, CRNNs offer an optimal trade-off between parameter efficiency and accuracy, requiring significantly less data than Transformer-based models [2].

2.2 Indigenous Language Technologies

Mager et al. [11] highlighted unique challenges for indigenous language technologies: morphological complexity, lack of standardized orthography, and data scarcity. Amith et al. [1] released the Puebla-Nahuatl Audio Corpus via OpenSLR, providing transcribed recordings for ASR baselines. The TonalliCorpus [13] covers multiple Nahuatl and Totonaco variations for AI applications, and the Highland Puebla Nahuatl Speech Translation corpus [17] provides aligned speech-text-translation triples. Self-supervised methods like Wav2Vec 2.0 [2] show promising zero-shot transfer; however, none of these works explicitly address acoustic domain mismatch between studio and field recordings, a gap this work fills through spectral filtering as a lightweight, training-free domain adaptation mechanism.

2.3 Spectral Filtering and Psychoacoustics

The critical information for speech intelligibility is largely contained below 4 kHz [12]. LID systems for major languages have proven robust in the telephony band [19]; we extend this validation to Nahuatl-Spanish discrimination.

3 Background

3.1 Time-Frequency Analysis and Mel Spectrograms

We transform raw audio into a time-frequency representation via the Short-Time Fourier Transform (STFT). Given a discrete signal $x[n]$, windowed by $w[n]$ (Hann window):

$$X(m,k) = \sum_{n=0}^{N-1} x[n+mH]\, w[n]\, e^{-j\frac{2\pi kn}{N}} \tag{1}$$

where N is the FFT size, H the hop length, k the frequency bin, and m the time frame. The spectrogram $S(m,k) = |X(m,k)|^2$ discards phase. We apply a Mel

filterbank to align the representation with human auditory perception [12]:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

The resulting Log-Mel Spectrogram $I \in \mathbb{R}^{F \times T}$ is the network input.

3.2 CRNN Architecture Components

CNN Front-End. A 2D convolution kernel K extracts local spectral correlations (formants, harmonic stacks):

$$Y_{i,j} = f\left(\sum_{u,v} K_{u,v} \cdot I_{i+u,j+v} + b\right) \tag{3}$$

BiLSTM. LSTMs [7] use gating to capture long-term phonotactic dependencies. We use a Bidirectional LSTM to exploit both past and future context over each 3-second window.

Normalization. Instance Normalization (IN) [20] normalizes each sample using its own statistics:

$$\hat{x}_{IN} = \frac{x - \mu_{\text{instance}}}{\sqrt{\sigma^2_{\text{instance}} + \epsilon}} \tag{4}$$

The standard alternative, Batch Normalization (BN), uses mini-batch statistics μ_B and σ^2_B:

$$\hat{x}_{BN} = \frac{x - \mu_B}{\sqrt{\sigma^2_B + \epsilon}} \tag{5}$$

BN is unreliable when batches mix recordings with high inter-sample variance. IN acts as a per-sample dynamic range compressor, rendering the model invariant to global amplitude shifts.

4 Methodology

4.1 Dataset Curation

We construct a binary corpus pairing naturalistic Nahuatl recordings with studio-quality Spanish. Table 1 summarises the key properties of both sub-corpora. For the Nahuatl component, we use the field-recorded interviews from the TonalliCorpus [13]. It is crucial to clarify that although the TonalliCorpus was originally designed as a multilingual parallel corpus (for translation and alignment tasks), for the purposes of this study, we extract a strictly non-parallel subset to construct a binary classification dataset. The objective is Language Identification (LID) across two distinct acoustic domains, evaluating whether the model can learn underlying phonological structures rather than discourse alignment. Specifically, we employ its Nahuatl subset, consisting of 44 speaker interviews recorded in classroom and community settings, which provides the controlled

Table 1. Summary of the two source corpora used in this study.

Corpus	Language	Speakers	Duration	Environment
TonalliCorpus [13] (Nahuatl subset)	Nahuatl	44	≈14 h	Classroom (naturalistic)
CIEMPIESS LIGHT [6]	Mexican Spanish	Multiple	>17 h	Broadcast studio

acoustic heterogeneity, field-recorded indigenous speech versus broadcast-quality studio data required by our experimental design.

The Nahuatl interviews were recorded in classrooms by 44 distinct speakers, introducing Room Impulse Responses. CIEMPIESS LIGHT [6] was randomly undersampled to match the Nahuatl chunk count (≈14,860 per class).

4.2 Signal Processing Pipeline

All audio was processed with `torchaudio` and segmented into **3-second non-overlapping windows**, yielding $N \approx 29{,}720$ samples total. To mitigate the risk of generating samples consisting solely of prolonged silences or pauses without relevant linguistic content, segments lacking sufficient acoustic energy were excluded during preprocessing.

A. Wideband (16 kHz): Audio resampled to 16 000 Hz, preserving spectral energy up to 8 kHz.

B. Narrowband (4 kHz): A digital Low-Pass Biquad Filter ($f_c = 4\,000$ Hz) was applied:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{6}$$

This attenuates all energy above 4 kHz, preserving F_0 and the first three formants (F_1, F_2, F_3). The 4 kHz threshold was selected based on phonetic principles: the primary acoustic cues for distinguishing Nahuatl's complex vowel system (including long vowels) and Mexican Spanish's five-vowel system reside in these first three formants, which are robustly contained within the 0–4 kHz range. Frequencies above this threshold in field recordings predominantly capture non-linguistic environmental artifacts (e.g., electronic hiss, wind) rather than discriminative phonetic features.

Feature Extraction. Log-Mel Spectrograms were computed with FFT size 1024, hop length 512, and 64 Mel filterbanks. Each 3-second clip yields a tensor of shape $(1, 64, 94)$.

4.3 Model Architecture

While state-of-the-art self-supervised foundation models (e.g., Wav2Vec 2.0) offer strong capabilities, their massive parameter counts render them highly impractical for offline deployment on resource-constrained hardware. Consequently, our architectural choice of a Convolutional Recurrent Neural Network (CRNN)

was a deliberate engineering decision. CRNNs provide an optimal trade-off between parameter efficiency and accuracy, requiring a fraction of the computational and memory overhead of Transformer-based architectures. This lightweight footprint is essential for our ultimate objective: deploying functional language technologies on low-cost edge devices (such as a Raspberry Pi) directly within indigenous communities lacking reliable internet access.

The CRNN (Fig. 1) comprises a **CNN Front-End** of three Conv2D blocks ($1 \to 32 \to 64 \to 128$ channels, 3×3 kernels), each followed by Batch Normalization, ReLU, and MaxPool(2×2). A **Feature Reshaping** step permutes the output to $(B, T, C \times F)$, fed into a **BiLSTM** [7] with 128 hidden units. The **Classification Head** applies FC($256 \to 64$), Dropout($p = 0.3$), and FC($64 \to 1$) with Sigmoid activation.

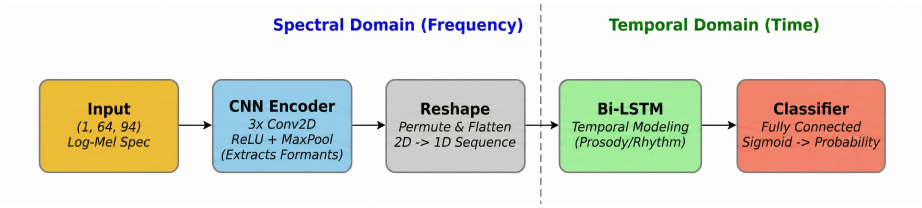

Fig. 1. CRNN architecture for Nahuatl-Spanish LID. The pipeline illustrates the domain shift addressed in this work.

4.4 Training and Validation Protocol

We employed the AdamW optimizer [10] (lr $= 5 \times 10^{-4}$) with Binary Cross-Entropy loss, a `ReduceLROnPlateau` scheduler (factor 0.5, patience 3), batch size 32, and Early Stopping with a 20-epoch window. All experiments were conducted on an NVIDIA T4 GPU.

For validation, **Group K-Fold Cross-Validation (k=5)** used the source interview filename as the grouping variable, ensuring no chunks from a given speaker appear simultaneously in training and validation folds. This prevents Speaker Leakage, where a model memorizes biometric voice features instead of language identity, ensuring robust generalization across different communicative styles. We report Binary Accuracy and Binary Cross-Entropy Loss averaged across all folds.

5 Results and Discussion

5.1 Quantitative Performance and Ablation Study

To rigorously validate our central hypothesis and isolate the effectiveness of our proposed robustness interventions, we conducted an ablation study. Table 2 com-

pares four architectural configurations, methodically stripping away and adding our bandwidth restriction (16 kHz vs. 4 kHz) and normalization techniques.

Table 2. Ablation Study: 5-Fold Group Cross-Validation Results isolating bandwidth and normalization effects (average across folds).

Bandwidth	Normalization	Val Acc (%)	Val Loss
16 kHz (Wide)	None	99.02	0.032
16 kHz (Wide)	Instance Norm	99.25	0.021
4 kHz (Narrow)	None	99.10	0.028
4 kHz (Narrow)	**Instance Norm**	**99.51**	**0.015**

The ablation results reveal a counter-intuitive but significant finding: *reducing* spectral bandwidth improved model performance. The 4 kHz + Instance Norm model achieved the highest accuracy and the lowest validation loss (0.015). The 16 kHz wideband models exhibited higher loss volatility, consistent with occasional overfitting to high-frequency environmental noise present in Nahuatl field recordings but absent in the studio Spanish corpus.

Table 3 reports per-fold results for our optimal ablated configuration. The low standard deviation (accuracy $\pm 0.05\%$, loss ± 0.001) confirms consistent generalization across all speaker groups.

Table 3. Per-fold validation results for the optimal 4 kHz + Instance Norm configuration.

Fold	1	2	3	4	5	**Mean**	**Std**
Val Acc (%)	99.47	99.55	99.49	99.58	99.44	**99.51**	± 0.05
Val Loss	0.017	0.014	0.016	0.013	0.015	**0.015**	± 0.001

5.2 Confusion Matrix

Table 4 presents the aggregate confusion matrix for the best configuration across all five folds ($N \approx 29{,}720$, balanced at 14,860 per class).

Table 4. Aggregate confusion matrix, 4 kHz + Instance Norm (5-Fold, $N \approx 29{,}720$).

	Pred. Nahuatl	**Pred. Spanish**
Actual Nahuatl	14,854 (TP)	$\approx$12 (FN)
Actual Spanish	$\approx$0 (FP)	14,854 (TN)

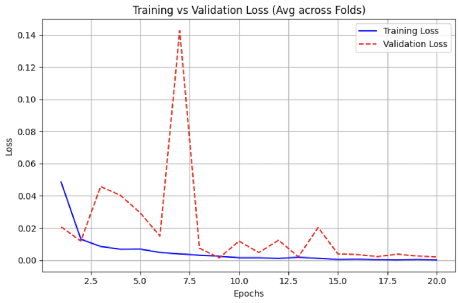

Fig. 2. Learning curves for the 4 kHz + Instance Norm model. Rapid convergence within 5 epochs and stable Train/Val loss confirm the absence of overfitting.

Only 12 errors out of 29,720 samples yield Precision = 100%, Recall ≈99.92%, and $F_1 \approx 99.96\%$. All misclassifications are False Negatives (Nahuatl predicted as Spanish; zero FP), indicating a marginal conservative bias toward the Spanish class, attributable to the higher acoustic consistency of broadcast training data.

5.3 Convergence Dynamics

Instance Normalization accelerated convergence substantially: normalized models reached >90% accuracy by Epoch 3, whereas non-normalized counterparts required 5–7 epochs. Figure 2 shows the training curves for the best model.

5.4 Spectral Robustness

The success of the 4 kHz model validates the Channel Effect hypothesis. Nahuatl and Mexican Spanish differ markedly at the phonological level: Nahuatl is polysynthetic with a lateral affricate /tl/, glottal stop /ʔ/ (saltillo), and long vowels producing distinct F_1/F_2 trajectories, whereas Spanish presents a five-vowel system with rhotic contrasts and no glottal phonemes. These asymmetries produce highly separable formant patterns in the 0–4 kHz band, while broadband room noise removed by the low-pass filter adds only confounding variance. Given the balanced dataset, the 99.51% accuracy reflects genuine discriminative learning rather than relying solely on recording environments.

5.5 Explainability via GradCAM

To transition from a "black box" to a verified linguistic tool, we applied GradCAM [16] to our best model (Fig. 3). It is important to clarify that the activation magnitudes in the visualizations (on the order of 10^{-5}) represent raw gradient signals extracted from the final convolutional block prior to layer normalization. In this context, the absolute numerical value is less critical than the spatial alignment of the activations. The maps clearly identify high-energy regions corresponding to formant trajectories and consonant clusters, while silence segments

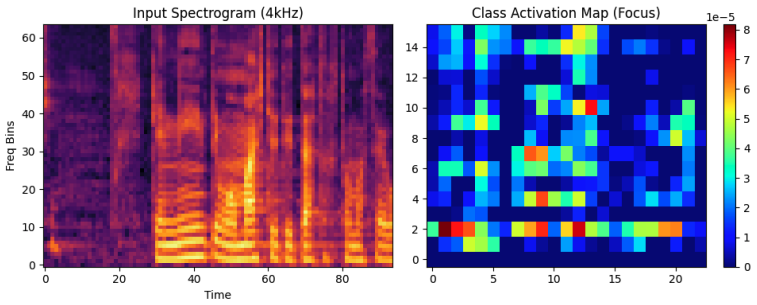

Fig. 3. GradCAM analysis of a Nahuatl sample. High activation (red/yellow) spatially aligns with formant transitions; silence and noise bands are ignored.

and static noise bands show near-zero activation, confirming that high accuracy is driven by linguistic pattern matching rather than background noise detection.

5.6 Practical Implications and Methodological Constraints

These findings have direct implications for deploying Indigenous Language Technologies: models can run on low-cost edge devices (e.g., Raspberry Pi) using 4 kHz audio. Naturalistic field recordings captured in classrooms are sufficient if processed with narrowband filtering and instance normalization, removing the need for expensive studio time.

However, several methodological constraints remain. First, while our dataset is balanced acoustically, it is not strictly parallel in linguistic content. We acknowledge that comparing conversational field interviews (Nahuatl) to scripted broadcast radio (Spanish) introduces speech style as a significant confounding variable. Consequently, the near-perfect 99.51% accuracy likely reflects a combination of true phonetic discrimination and learned differences in discourse cadence. While our proposed 4 kHz narrowband filter successfully mitigates the *acoustic* domain gap by removing environmental artifacts, it cannot completely eliminate this *stylistic* domain gap. Nevertheless, the filter serves as a critical step toward isolating essential phonetic features in highly heterogeneous datasets.

Second, the segmentation of audio into fixed 3-second windows, while standard practice, may dilute the presence of distinct phonetic units if boundaries are not aligned. Lastly, the variant modeled is highly specific and may differ phonologically from Huasteca or Guerrero Nahuatl, limiting immediate generalizability. The model also currently lacks a Non-Speech class, required for processing unsegmented audio in production environments.

6 Conclusion

We demonstrated that a CRNN trained on 4 kHz narrowband Log-Mel Spectrograms with Instance Normalization achieves superior performance (99.51%

accuracy, $F_1 \approx 99.96\%$) for Nahuatl-Spanish LID. Addressing RQ1, spectral bandwidth reduction to 4 kHz *enhanced* rather than degraded robustness by eliminating high-frequency environmental artifacts while preserving discriminative phonetic structures. Addressing RQ2, Instance Normalization accelerated convergence by 2–4 epochs and reduced loss volatility, demonstrating its effectiveness as a per-sample amplitude normalizer under high inter-recording variance. Addressing RQ3, GradCAM confirms that spatial decision boundaries align with formant transitions rather than noise, validating the linguistic basis of the model's predictions. Group K-Fold cross-validation with strict speaker isolation ensures metrics reflect true language identification rather than speaker recognition.

To address the limitations of the current study, future work will extend this pipeline in four critical directions: (i) incorporating strict text-to-audio phonetic alignment or baseline pronunciation dictionaries to isolate precise linguistic units during classification; (ii) evaluating modern, self-supervised audio representations such as WavLM, which have demonstrated superior robustness to acoustic mismatch and may reduce reliance on handcrafted spectral features; (iii) evaluating zero-shot generalization to unseen Nahuatl dialects, utilizing parallel subsets of the TonalliCorpus [13] to serve as a natural testbed for cross-variant and cross-lingual transfer experiments without the confounding variable of speaking style; and (iv) compressing the model via weight quantization and pruning for offline deployment on edge hardware in communities without internet access, in alignment with Green AI principles [15].

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. Amith, J.D., et al.: Zacatlán-Ahuacatlán-Tepetzintla (Puebla) Nahuatl speech corpus. OpenSLR (2023), `https://www.openslr.org`
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). vol. 33, pp. 12449–12460 (2020)
3. Bartz, C., Herold, T., Yang, H., Meinel, C.: Language identification using deep convolutional recurrent neural networks. In: Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science. vol. 10639, pp. 880–889. Springer (2017). `https://doi.org/10.1007/978-3-319-70136-3_93`
4. Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., Dehak, R.: Language recognition via I-vectors and dimensionality reduction. In: Proceedings of Interspeech. pp. 857–860. Florence, Italy (2011)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). `https://doi.org/10.1109/CVPR.2016.90`

6. Hernandez Mena, C.D., Herrera Camacho, A.: CIEMPIESS: A new open-sourced Mexican Spanish radio corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 371–375. European Language Resources Association, Reykjavik, Iceland (2014)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
8. Instituto Nacional de Estadística y Geografía: Censo de población y vivienda 2020. INEGI (2020), https://www.inegi.org.mx/programas/ccpv/2020/
9. International Telecommunication Union: Recommendation G.711: Pulse code modulation (PCM) of voice frequencies. Tech. rep., ITU-T, Geneva, Switzerland (1972)
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the 7th International Conference on Learning Representations (ICLR) (2019), https://arxiv.org/abs/1711.05101
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). pp. 55–69. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018)
12. O'Shaughnessy, D.: Speech Communications: Human and Machine. IEEE Press, New York, NY, 2 edn. (2000)
13. 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 In press
14. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning (ICML). vol. 202, pp. 28492–28518. PMLR (2023)
15. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green AI. Communications of the ACM **63**(12), 54–63 (2020). https://doi.org/10.1145/3381831
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626. Venice, Italy (2017)
17. Shi, J., Amith, J.D., Chang, X., Dalmia, S., Yan, B., Watanabe, S.: Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP). pp. 53–63. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.americasnlp-1.7
18. Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., Khudanpur, S.: Spoken language recognition using X-vectors. In: Proceedings of Odyssey 2018 – The Speaker and Language Recognition Workshop. pp. 105–111 (2018). https://doi.org/10.21437/Odyssey.2018-15
19. Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller, J.R.: Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). pp. 89–92. Denver, Colorado, USA (2002). https://doi.org/10.21437/ICSLP.2002-74
20. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2017)