Convex Low-resource Accent-Robust Language Detection in Speech Recognition

MIRIA FENG, WILLIAM TAN, AND MERT PILANCI

43rd Conference on International Conference on Machine Learning
(ICML 2026)

Preliminaries and Method

Spoken Dialogue Models

- Becoming globally ubiquitous, with applications ranging from Siri, Amazon Echo, to LLMs.
- ASR is a shared component across these systems, but speech data is more scarce than text data.
- 380 Million people speak English as their first or second language, and globalization continues to produce diverse speech varieties / new linguistic patterns.

Convex Language Detection (CLD) and ASR

- Pilanci and Ergen[1] developed exact representations of training two-layer ReLU NN with a single convex program
- We recast the low-resource training problem as a convex program.

Algorithm 1 CLD Training (Offline)

Input: Whisper Encoder $\mathcal{E}$, Dataset $\mathcal{D}_{\text{train}}$, parameters ρ, β

Output: Trained Convex Detection Head $\hat{f}_{\text{cvx}}$

for $i \leftarrow 1$ **to** N **do**

  $h_i \leftarrow \mathcal{E}(x_i)$

Initialize ADMM variables $(\mathbf{v}, \mathbf{w}, \mathbf{u})$.

repeat

  $(\mathbf{v}, \mathbf{w}) \leftarrow \arg\min_{\mathbf{v},\mathbf{w}} \left[\ell\left(\sum_{p=1}^{P} D_p X(\mathbf{v}_p - \mathbf{w}_p), y \right) + \beta \sum_{p=1}^{P} (\|\mathbf{v}_p\|_2 + \|\mathbf{w}_p\|_2) + \frac{\rho}{2}\|v\|_2^2 + \|w\|_2^2 \right]$

  $\mathbf{u} \leftarrow \mathbf{u} + \rho \cdot (\mathbf{v} - \mathbf{w})$

until *convergence*

return *Store trained weights as* $\hat{f}_{cvx}$

[1] Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks (2020)

Four Main Points of Analysis

(1) Variation-norm certificate. For hidden features $h = E(x)$ and CLD logits $f(h)$, any feasible solution $\{(v_p, w_p)\}_{p=1}^{P}$ of the convex program (Eq. 1) upper-bounds the variation norm:

$$\|f\|_{\text{var}} \le \mathcal{B}_{\text{cvx}} := \sum_{p=1}^{P} \left(\|v_p\|_2 + \|w_p\|_2\right).$$

(2) Exact logit-Lipschitz constant. This bound makes CLD provably Lipschitz-stable in encoder space:

$$\|f(h) - f(h')\|_\infty \le \mathcal{B}_{\text{cvx}} \|h - h'\|_2.$$

(3) Certified margin stability. For the one-vs-rest margin $\text{mar}(h, y) = f_y(h) - \max_{k \neq y} f_k(h)$, any hidden-feature perturbation δ satisfies

$$\text{mar}(h + \delta, y) \ge \text{mar}(h, y) - 2\,\mathcal{B}_{\text{cvx}} \|\delta\|_2.$$

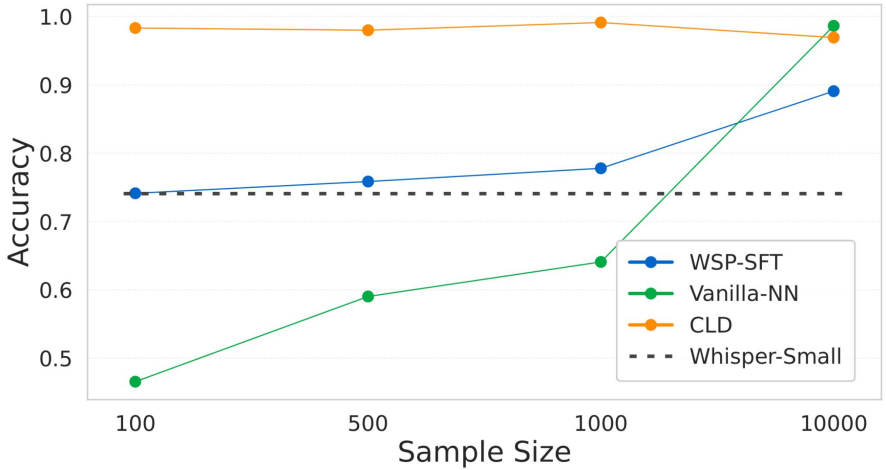
(4) Certified radius of label invariance. The predicted language is provably unchanged within the certified feature-space radius

$$\|\delta\|_2 < r_h(h, y) := \frac{\text{mar}(h, y)}{2\,\mathcal{B}_{\text{cvx}}}.$$

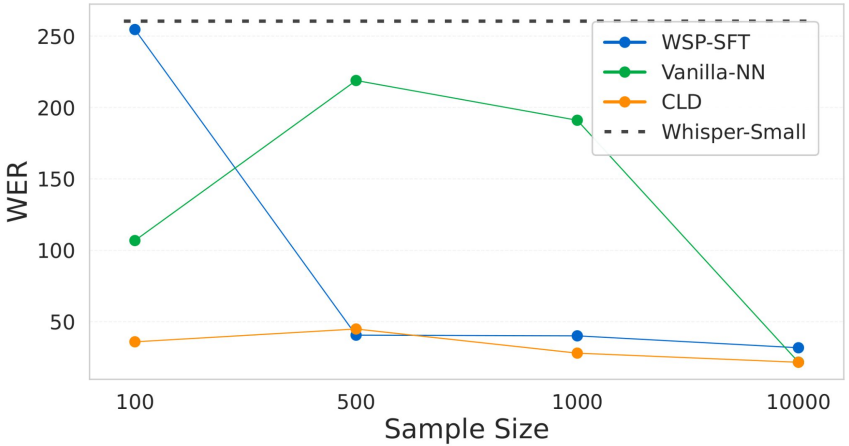
Summary: larger margins ⇒ larger certified radius; stronger convex penalty ⇒ smaller $\mathcal{B}_{\text{cvx}}$, tightening the guarantee. Feature-space certificates are our primary stability measure; end-to-end audio radii ($r_x = r_h / L_E$) are reported as conservative diagnostics, since global encoder Lipschitz bounds L_E are pessimistic for deep Transformers.

Please see Section 4, and Appendix A of the paper for Theorems and associated proofs.

Binary Experiment



a) EN-ZH binary classification accuracy vs. training sample sizes



b) EN-ZH binary classification WER vs. training sample sizes

Abbreviations: Fine-tuned whisper (WSP-SFT)

Multiclass Experiment

	Detection Accuracy			WER			CER		
Language Classifier	WSP	WSP-L	MMS-1B	WSP	WSP-L	MMS-1B	WSP	WSP-L	MMS-1B
Default	0.7154	0.8033	0.6701	139.37	40.41	51.88	73.85	21.80	27.61
KNN	0.6123	0.7145	0.4981	145.21	44.89	57.34	81.05	29.12	32.76
Linear SVM	0.9392	0.9501	0.5653	48.74	39.36	50.73	28.28	23.68	26.07
Kernel SVM	0.9431	0.9582	0.5701	46.52	37.91	49.12	26.14	22.05	25.88
NN	0.7737	0.9605	0.8612	53.84	29.25	48.26	34.52	15.99	23.64
CLD (ours)	**0.9715**	**0.9806**	**0.9702**	**31.74**	**28.60**	**45.27**	**17.84**	**15.37**	**21.58**

CLD multiclass detection accuracy, WER, CER benchmarked across language classifiers and ASR models

Evaluated on 5 languages: English, Chinese, Indonesian, Malaysian, and Hindi
Model abbreviations: Whisper-small (WSP), Whisper-large-v3 (WSP-L), and MMS-1B

Conclusions and Future Work

- We propose **Convex Language Detection (CLD),** a fast sample-efficient algorithm for robust spoken language classification within low-resource accented data regimes. Available as a package at [2].

- **Robustness**: performant in low-resource regimes (~100 samples), and empirically outperforms competitors at 97-99% accuracy

- **Speed**: 13× fewer training TFLOPs of a comparable non-convex MLP, enabled by GPU acceleration and parallelism with JAX-ADMM

- **Future work**: end-to-end differentiable CLD pipeline by unrolling the ADMM iterations or applying implicit differentiation through the KKT conditions of the convex program. Particularly amiable with our JAX codebase!

[2] Pip package at https://pypi.org/project/jaxcld/