CLARA: CONVEX LOW-RESOURCE ACCENT-ROBUST LANGUAGE DETECTION IN ASR

Anonymous authors
Paper under double-blind review

ABSTRACT

Globalization and multiculturalism have produced diverse speech varieties (such as Singaporean-accented English and numerous regional Mandarin dialects) that remain under-represented even within high-resource language data. As a result, spoken dialogue systems frequently misidentify the user's input language up to 49.33% of the time, degrading response accuracy regardless of language model capability. We propose a novel robust ASR framework for handling low-resource dialectal variance with minimal computational overhead, and lightweight training costs. Our Convex Language Detection (CLD) algorithm integrates a convex reformulation of a vanilla neural network (NN), and is solved efficiently with ADMM based methods in JAX. This provides human-level sub-500ms inference latency, strong convergence guarantees, reduced sample complexity, and practical possibilities for edge deployment use cases. As a motivating case study, CLD significantly improves transcription accuracy on mixed-dialect inputs when integrated with Whisper encoders. This demonstrates promising directions for principled statistical generalization in spoken dialogue systems for low-resource languages.

1 INTRODUCTION

Spoken language dialogue systems are increasingly ubiquitous across all cultures, countries, and applications. The common component in these systems is the critical Automatic Speech Recognition (ASR) model, which transcribes input user speech into text for downstream Large Language Models (LLMs) for processing. Without accurate transcription, even the most advanced LLMs cannot correctly interpret user intent or generate accurate responses. The widely adopted Whisper (Radford et al., 2022) ASR model series demonstrates strong ability to generalize to many datasets and domains in a zero-shot setting, yet often misidentifies the input language token due to user dialects and accents. This occurs since existing voice-transcription datasets typically do not annotate human accents, resulting in under-representation of regional dialects even within high-resource languages. For example, although the national language of Singapore is English (arguably the most dominant language in voice datasets), the unique and prevalent dialect of Singaporean accented English has led to the colloquial term "Singlish" (Wee, 2018). The intonation and prosody of Singlish is so distinct that it has been widely studied by linguists (Goh, 2016), Hoon (2003), Rubdy (2007), yet state-of-the-art ASR models often mistakenly transcribe it as a foreign language (such as Bahasa (Le Page, 1984) and Tamil (Rajan, 2018)).

1.1 MOTIVATIONS

In this paper, we aim to take a step towards democratizing the accessibility of spoken dialogue systems to robustly handle user speech input from multicultural backgrounds. For example, consider the scope defined by two resource-heavy languages: English and Mandarin, which are composed of numerous distinctive regional dialects. We introduce the novel Convex Language Detection (CLD) framework, which achieves global optimality in polynomial time, offers improved sample efficiency, and improved generalization bounds. As a result, the CLD architecture of only ten neurons is able to capture more signal with respect to user input sequences, than the standard linear layer in existing Whisper models. This efficiency is crucial, since end-to-end speech dialogue models require sub-500ms latency Meyer (2023) to preserve realistic human response time. We further optimize for

fast training and iteration by implementing our method in JAX Bradbury et al. (2021) and solving the optimization problem with principled ADMM based techniques (Feng et al., 2024) for Large Langauge Models (LLMs). To the best of our knowledge, this is the first time convex optimization reformulations have been practically applied to spoken language dialogue models.

1.2 CONTRIBUTIONS

We introduce and evaluate the effectiveness and practical feasibility of the novel Convex Language Detection (CLD) layer for low resource dialects in ASR systems:

- We find that even the most widely adopted ASR models do not classify accented dialects from large language datasets sufficiently accurately. We benchmark against the standard vanilla NN for language detection (which is currently commonly utilized in practical deployment) and demonstrate the improved classification accuracy, faster training efficiency, and higher memory efficiency of the CLD.
- We extend the work of Feng et al. (2024) from binary classification of text driven problems in LLMS, to *multiclass* problems in spoken dialogue systems. This further lifts the size constraints for cvxNN applications, since speech datasets are notoriously high-dimensional and slow to tractably work with. This extension is theoretically supported with statistical generalization and margin stability analysis.
- We introduce the novel CLD algorithm, and provide theoretically grounded analysis supporting its margin stability and statistical generalization.
- Our custom sourced and formatted dialect and low resource language datasets are provided for further ongoing research in spoken dialogue systems. In addition, our JAX code base is open source for replicability at , to help globally democratize voice assistants for all languages and persons.

The following paper is presented as follows: Section 2 outlines related work, Section 3 introduces the novel Convex Language Detection framework, Section 4 provides theoretical support, Section 5 gives main experimental results and discussion, finally Section 6 provides conclusions and directions for future work. For additional details and JAX code base, please see `https://anonymous.4open.science/r/CLD-845F/README.md`.

2 RELATED WORK

Foundational multilingual ASR models with up to 1550 million parameters and have been trained on 99+ languages Radford et al. (2022). However the vast majority of these models perform the best on English, with performance dropping significantly on lower resource languages Graham & Roll (2024). This has recently encouraged much work in the field of improving low-resource ASR performance. For example, the authors of Bansal et al. (2018), Khare et al. (2021), Stoian et al. (2020) propose using transfer learning via pretraining techniques to improve cross-lingual transfer. This requires expansive amounts of speech data in existing high-resource languages but with text transliterated to the target low-resource language. Essentially the mapping serves to encourage increased sharing between the output spaces of both languages, yet the success of pretraining is not well defined. The high-resource and low-resource language must share a certain amount of unclear ”basis similarity” in linguistics for this to be feasible. During the course of pretraining on extremely large datasets, the powerful base ASR model also experiences catastrophic forgetting, leading to overall deterioration in performance.

Even within high resource languages such as English and Mandarin, there exist many distinct dialects which state-of-the-art ASR models struggle to identify correctly. The recent works of Li et al. (2024), Weninger et al. (2019), and Wang et al. (2025) aim to implement prosody-assisted speech systems, or bidirectional Long-Short-Term Memory networks to better model acoustic context. With the the rise in popularity of spoken dialogue models, other researchers Reitmaier et al. (2022) have focused on more clearly identifying the challenges ASR models face with low-resource languages. These methods share the common weakness of being heavily dependent on large fine-tuning datasets

with a learning rate that is typically ten times smaller than standard supervised fine-tuning learning rates Wilson & Martinez (2001), Liu et al. (2024), de Zuazo et al. (2025).

Recently other researchers (Reitmaier et al., 2022) have focused on more clearly defining the challenges ASR models face with low-resource languages, such as Xhosa or Marathi. Limited training data is a dominant issue, and authors (Babirye et al., 2022) have worked on building partnerships to preserve and document valuable linguistic data by remotely engaging local participants to record themselves, identifying more recording opportunities, and categorizing challenges of ASR in deeply multicultural communities. This has uncovered valuable implications for collaborations across ASR and Human Computer Interface (HCI) that advance important discussions, while collecting more diverse speech datasets. Although promising, this approach also brings up new questions on the ethics of analyzing community voice recordings through platforms such as WhatsApp Barbosa & Milan (2019), and is slow to provide clearly annotated data from numerous low-resource languages.

Instead of relying heavily on pretraining, fine-tuning, or gathering more data: our key insight is that the existing seminal ASR models (such as Whisper (Radford et al., 2022)) have already been trained on 680,000 hours of speech-transcription data . Therefore we aim to creatively extend traditional resource intensive techniques by focusing on implementing a fast, efficient and practical language detection modification layer inside the Whisper architecture, which is capable of robustly and accurately mapping input dialects to respective languages. Since our method utilizes a convex reformulation of a multi-layer perceptron (MLP), we can achieve global optimality in polynomial time without incurring any additional latency at inference time.

3 Convex Language Detection Algorithm

In this section we introduce the Convex Language Detection (CLD) architecture for ASR in spoken dialogue systems. Section 3.1 provides preliminaries on two-layer ReLU networks, Section 3.2 introduces the equivalent convex optimization reformulated problem, Section 3.3 gives background on the high resource language-low resource dialect problem, and Section 3.4 presents its integration in the language detection problem to yield the CLD algorithm.

3.1 Preliminaries

The classic two-layer ReLU network is given by:

$$f(x) = \sum_{j=1}^{m} (\Theta_{1j} x)_+ \theta_{2j}, \tag{1}$$

Here $x \in \mathbb{R}^d$ represents the input, $\Theta_1 \in \mathbb{R}^{m \times d}$, $\theta_2 \in \mathbb{R}^m$ are weights of the first and last layers respectively, and $(\cdot)_+ = \max\{\cdot, 0\}$ is the ReLU activation function. Given targets $y \in \mathbb{R}^n$, the network in equation 1 seeks optimality by minimizing the non-convex loss function:

$$\min_{\Theta_1, \theta_2} \ell\left(f_{\Theta_1, \theta_2}(X), y\right) + \frac{\beta}{2} \sum_{j=1}^{m} \left(||\Theta_{1j}||_2^2 + (\theta_{2j})^2 \right), \tag{2}$$

where $\ell : \mathbb{R}^n \mapsto \mathbb{R}$ is the loss function, $X \in \mathbb{R}^{n \times d}$ is the data matrix, and $\beta \geq 0$ is the regularization strength. equation 2 presents a challenging non-convex optimization problem, with necessary iterations of hyperparameter grid-search for successful training. This is unprincipled and expensive to scale, especially in high-dimensional speech datasets that are significantly slower to train and more resource-intensive Sainath et al. (2013). Our goal is to maintain these expressive capabilities while still preserving the computational advantages of convex optimization.

3.2 Equivalent Convex Reformulation

Pilanci & Ergen (2020) have shown equation 2 admits a convex reformulation (cvxNN). Since the reformulation has the same optimal value as the original non-convex problem, provided $m \geq m^*$, for some $m \geq n+1$, no information is lost in equation 2. This is based on enumerating the actions of all possible ReLU activation patterns on data matrix X. These activation patterns act as separating

hyperplanes which multiply the rows of X by 0 or 1, and can be represented by diagonal matrices. For fixed X, the set of all possible ReLU activation patterns may be expressed as

$$\mathcal{D}_X = \left\{ D = \text{diag}\left(\mathbf{1}(Xv \geq 0)\right) : v \in \mathbb{R}^d \right\}.$$

The cardinality of $\mathcal{D}_X$ grows as $|\mathcal{D}_X| = \mathcal{O}\left(r(n/r)^r\right)$, where $r \coloneqq \text{rank}(X)$ Pilanci & Ergen (2020). Since the exponential size of $\mathcal{D}_X$ Pilanci & Ergen (2020) make its complete enumeration impractical, we work with a subset based on sampling P patterns from $\mathcal{D}_X$:

$$\min_{(v_i, w_i)_{i=1}^P} \ell\left(\sum_{i=1}^P D_i X(v_i - w_i), y\right) + \beta \sum_{i=1}^P ||v_i||_2 + ||w_i||_2 \quad \text{s.t. } v_i,\ w_i \in \mathcal{K}_i \quad \forall i \in [P]. \tag{3}$$

It can be shown under mild conditions that equation 3 still has the same optimal solution as equation 2 Mishkin et al. (2022). The recent work of Kim & Pilanci (2024) also proves that the difference is negligible even when they are not equal. Therefore, we can work with confidence with the tractable convex framework in equation 3.

3.3 Defining the Low Resource Dialects Problem

Although large-scale audio–text datasets have recently driven advances in ASR, most speech data remains expensive to collect and rare to annotate, especially for languages with varied dialects. Even the seminal corpora that underpins modern ASR systems (e.g., 680,000+ hours in Whisper training) overwhelmingly focus on a few dominant resource languages such as English and Mandarin, while leaving their many dialectal and accented forms unlabeled or ignored. This imbalance creates a critical gap: speakers with regional accents, from Singaporean English ("Singlish") to Shanghai Mandarin accents, currently have poor spoken dialogue systems experiences (analysis studies presented in Table 8) since even foundational ASR models frequently misidentify dialectical speech.

3.4 Integration with Spoken Dialogue Systems

Feng et al. (2024) has demonstrated the successful application of cvxNN on high-dimensional text-based language tasks. Therefore we aim to extend this approach on larger-scale spoken dialogue systems, by extracting the *hidden features* from the encoder of Whisper ASR. The Convex Language Detection (CLD) algorithm is formally presented below, where $\hat{y}$ represents the language label, $\hat{t}$ represents the decoded transcript, x is the input audio waveform, and $\{(x_i, y_i)\}_{i=1}^N$ represents the training set.

Algorithm 1 Convex Language Detection (CLD)

Require: Whisper encoder $\mathcal{E}$, decoder $\mathcal{D}$, penalty parameter ρ, regularization β

Training (offline):

for $i = 1$ to N **do**

  $h_i \leftarrow \mathcal{E}(x_i)$ ▷ Extract hidden states

end for

Train cvxNN on $\{(h_i, y_i)\}$ using ADMM with variables $(\mathbf{v}, \mathbf{w}, \mathbf{u})$

repeat

  $(\mathbf{v}, \mathbf{w}) \leftarrow \arg\min \ell\Big(\sum_{p=1}^P D_p H(\mathbf{v}_p - \mathbf{w}_p), y\Big) + \beta \sum_{p=1}^P \big(\|\mathbf{v}_p\|_2 + \|\mathbf{w}_p\|_2\big) + \frac{\rho}{2}\|\cdot\|_2^2$

  $\mathbf{u} \leftarrow \mathbf{u} + (\text{primal residual})$ ▷ Dual variable update

until convergence

Store trained convex detection head $\hat{f}_{\text{cvx}}$

Inference (online):

$h \leftarrow \mathcal{E}(x)$ ▷ Encoder stage

$\hat{y} \leftarrow \arg\max \hat{f}_{\text{cvx}}(h)$ ▷ Lightweight forward pass

Append $\hat{y}$ as initial language token to $\mathcal{D}$

$\hat{t} \leftarrow \mathcal{D}(x\,;\, \text{init token} = \hat{y})$ ▷ Decoder stage

return $(\hat{y}, \hat{t})$

4 CLD THEORETICAL ANALYSIS

4.1 OPTIMIZATION PERSPECTIVE AND STATISTICAL GENERALIZATION

This motivating application of CLD in robust ASR under dialectal variation demonstrates the practical advantages of our optimization framework in real world settings. In addition to being a valuable application-specific heuristic in spoken dialogue systems, CLD proves the advantages of principled convex optimization in the non-convex landscape of deep learning. The instantiation of ADMM naturally leads to efficient parallelization on multi-GPU platforms, while enabling low latency inference on par with human speed. From a statistical perspective, convex formulations offer principled benefits such as stable solutions under perturbations in data. Since convex estimators are less prone to variance caused by initialization randomness or nonconvex local minima, this results in improved generalization bounds. The convex reformulation also extracts more signal from limited dialectal data without requiring large-scale pretraining or extensive fine-tuning, and the lack of dependence on hyperparameter grid-search reduces variance across runs. This ensures consistent optimal performance rather than depending on fortuitous optimizer trajectories.

4.2 MARGIN STABILITY

Let $E : \mathbb{R}^T \to \mathbb{R}^d$ denote the ASR encoder, and $h = E(x) \in \mathbb{R}^d$ be the hidden features, then $f : \mathbb{R}^d \to \mathbb{R}^K$ is the CLD detection module trained by the convex program in Eq.3. We first formalize the margin and representation used in the following analysis.

Definition 1 (One-vs-Rest classification margin)**.** For $y \in \{1, \dots, K\}$ and logits $f(h) = (f_1(h), \dots, f_K(h))$, define the classification margin as

$$\mathrm{mar}(h, y) \;:=\; f_y(h) \;-\; \max_{k \neq y} f_k(h).$$

Proposition 1 (Two-layer representation for the CLD)**.** *Under Eq.3, the optimal detection head admits a finite two-layer ReLU representation*

$$f(h) \;=\; \sum_{j=1}^{m} a_j \, [u_j^\top h]_+, \qquad a_j \in \mathbb{R}^K,\; u_j \in \mathbb{R}^d,\; m \leq n + 1, \tag{4}$$

with the same objective value as the convex program (up to negligible approximation from activation-pattern sampling). Sec. 3.2.

Definition 2 (Variation norm)**.** For f represented as in equation 4, its variation norm is

$$\|f\|_{\mathrm{var}} \;:=\; \inf \Big\{ \sum_{j=1}^{m} \|a_j\|_2 \, \|u_j\|_2 \;\;:\;\; f(h) = \sum_{j=1}^{m} a_j [u_j^\top h]_+ \Big\}.$$

Lemma 1 (Logit Lipschitzness). If f admits a representation equation 4, then for any $h, h' \in \mathbb{R}^d$,

$$\|f(h) - f(h')\|_\infty \;\leq\; \|f\|_{\mathrm{var}} \, \|h - h'\|_2.$$

Proof. Write $f_k(h) = \sum_j a_{j,k} [u_j^\top h]_+$. Since $t \mapsto [t]_+$ is 1-Lipschitz, $|f_k(h) - f_k(h')| \leq \sum_j |a_{j,k}| \, |u_j^\top (h - h')| \leq \sum_j \|a_j\|_2 \|u_j\|_2 \, \|h - h'\|_2$. Maximizing over k and taking the infimum over representations yields the claim. □

Theorem 1 (Margin stability under hidden-feature perturbations)**.** *Let f be the detection head given by Eq. equation 3. For any $y \in \{1, \dots, K\}$ and any $\delta \in \mathbb{R}^d$,*

$$\mathrm{mar}(h + \delta, y) \;\geq\; \mathrm{mar}(h, y) \;-\; 2 \, \|f\|_{\mathrm{var}} \, \|\delta\|_2. \tag{5}$$

Consequently, if $\|\delta\|_2 < \mathrm{mar}(h, y)/(2\|f\|_{\mathrm{var}})$, the predicted class is unchanged.

Proof. Let $k^\star(h) = \arg\max_{k \neq y} f_k(h)$. Then

$$\mathrm{mar}(h + \delta, y) - \mathrm{mar}(h, y) = \big(f_y(h + \delta) - f_y(h)\big) - \big(\max_{k \neq y} f_k(h + \delta) - f_{k^\star(h)}(h)\big).$$

Each difference is $\leq \|f\|_{\mathrm{var}} \|\delta\|_2$ by Lemma 1, yielding equation 5. □

Corollary 2 (End-to-end stability). If the encoder E is L_E-Lipschitz, i.e., $\|E(x) - E(x')\|_2 \leq L_E\|x - x'\|_2$, then

$$\operatorname{mar}(E(x+\eta), y) \;\geq\; \operatorname{mar}(E(x), y) - 2\,\|f\|_{\mathrm{var}}\,L_E\,\|\eta\|_2,$$

and the predicted class is preserved whenever $\|\eta\|_2 < \operatorname{mar}(E(x), y)/(2\|f\|_{\mathrm{var}}L_E)$.

Proposition 2 (Variation-norm certificate from the convex penalty)**.** *Let $\{(v_p, w_p)\}_{p=1}^{P}$ denote the variables in Eq. equation 3. Then*

$$\|f\|_{\mathrm{var}} \;\leq\; \mathcal{B}_{\mathrm{cvx}} \quad \textit{with} \quad \mathcal{B}_{\mathrm{cvx}} := \sum_{p=1}^{P} \big(\|v_p\|_2 + \|w_p\|_2\big),$$

interpreting $\|\cdot\|_2$ blockwise (e.g., columnwise ℓ_2 with a sum across classes). Consequently,

$$\operatorname{mar}(h+\delta, y) \;\geq\; \operatorname{mar}(h, y) - 2\,\mathcal{B}_{\mathrm{cvx}}\,\|\delta\|_2.$$

If the nonconvex two-layer form (Eq. (2)) with penalty $\frac{\beta}{2}\sum_j(\|u_j\|_2^2 + \|a_j\|_2^2)$ is used instead, then by AM–GM, $\|f\|_{\mathrm{var}} \leq \frac{1}{2}\sum_j(\|u_j\|_2^2 + \|a_j\|_2^2)$, so larger β tightens the certified radius.

Remark 3 (Group structure). If logits share blocks (group-sparse outputs), one obtains $\|f(h) - f(h')\|_\infty \leq \sum_g w_g\|A_g\|_2\|U_g\|_2\,\|h - h'\|_2$, hence the margin bound with $\|f\|_{\mathrm{var}}$ replaced by the weighted group sum.

Therefore, the CLD module is Lipschitz-stable in hidden features with constant controlled by the convex regularizer in Eq. 3. Since bounded perturbations of $E(x)$ induce at most linear margin degradation, sufficiently large initial margin certifies label invariance. Appendix A provides further derivation details.

5 Main Experiments

In this section we present the main experimental results and discussion. We note that although the seminal Word Error Rate (WER) metric Jelinek (1997) often produces approximately the same results across different runs, the resulting human evaluation for different models dramatically varies across implementation. This is because evaluation metrics for spoken dialogue systems remain an *approximation* for human feedback. Section 5.1 provides details on datasets, in Section 5.6 we go head-to-head with real human feedback in a practical application scenario, and Section 4.1 provides discussion on optimization perspective and statistical generalization. Our baseline model is **(1)** vanilla Whisper-Small (244 Million parameters, referred to as WSP). In comparison we also benchmark against **(2)** vanilla Whisper-Small finetuned on our dataset (referred to as WSP-SFT), **(3)** a two-layer MLP for language detection (referred to as NN), and finally our **(4)** CLD algorithm (referred to as CVXNN).

5.1 Datasets

In this experiment, we compile a dataset of multilingual voice transcriptions across multiple languages and their respective accents. As a primary source of transcription data, we used the Common Voice (v23) Dataset (Ardila et al., 2019), but due to the lack of accent and regional variance, we supplement this with several other accent datasets. For one, we selected the Singaporean English (Singlish) dialect, which has previously shown high error rates during voice transcription (Fong et al., 2002). Through the Info-communications and Media Development Authority (IMDA) of Singapore, we were given direct access to the National Speech Corpus (NCS): the first Singapore English corpus. In addition, we use the Lahaja dataset, a benchmark comprising 12.5 hours of Hindi speech from 132 speakers across 83 Indian districts, for regional Hindi dialects (Javed et al., 2024). We then normalize and augment all audio files via the following techniques: Time stretching, volume gain, pitch shift, and recorded background noise (via MUSAN Snyder et al. (2015)), which are all used to simulate real-world variability and improve robustness. Primarily, our experiment is split into two parts:

Binary Experiment. For the binary classification experiment, we select English and Mandarin since despite being arguably the highest-resource languages in existing seminal datasets, they still

exhibit low accuracy in language prediction for accented speech due to the high variance of dialects and accents present (Weninger et al., 2019). In addition, we select different varieties of training sample sizes spanning from 100 to 10K per language to test the model's robustness in low-resource environments.

Multiclass Experiment. For the multiclass classification task, we select a total of five languages: English, Chinese, Indonesian, Malaysian, Hindi. We selected these languages due to their linguistic and geographical proximities, as well as the regional influences that certain dialects exert on one another—for instance, Singlish (Singaporean accented English) is often misidentified as Malay or Indonesian. This selection ensures the model's task is challenging and thereby provides a rigorous evaluation of its performance. In total, we selected 16,000 training samples across 5 languages and 24 accents with approximately 3200 samples per language and 666 samples per accent.

5.2 Low Resource Binary Experiment

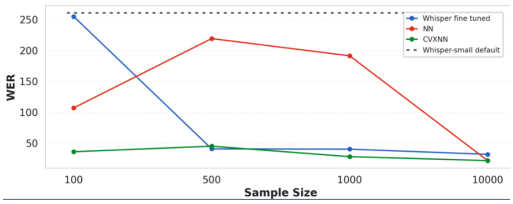


Figure 1: Word Error Rate (WER) vs. Sample Size

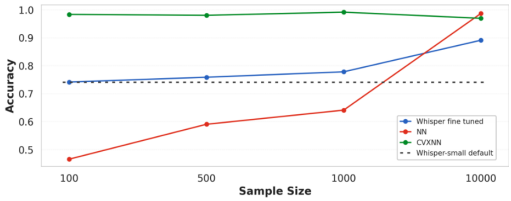


Figure 2: Language Detection Accuracy vs. Sample Size

We report the metrics of each model trained on the respective sample sizes in Figures 1 and 2. Both the vanilla NN and WSP-SFT showcase a similar correlation for lower WER and higher detection accuracy with sample size, requiring significant amounts of data for high performance, often which not available in low-resource dialects. However, our CLD model maintains consistent performance across all sample sizes, indicating a high sample efficiency and strong resilience against low-resource. In fact, in larger sample sizes, the CLD experiences a slight decrease in prediction accuracy, with peak detection accuracy occurring at a sample size of 1K. Therefore, the CLD model represents a promising direction, as it enables effective inclusion of accents and dialects with extremely limited speaker data and is an ideal application for low-resource data regimes.

5.3 CLD is FAST and Efficient

Table 1: Training Time and Compute Cost (TFLOPs) Across Models for the 1K Sample Size Dataset

Model	Training Time (s)	TFLOPs
WSP	1096.74	239,528
NN	840.30	183,521
cvxNN	64.45	14,075

Table 1 demonstrates that cvxNN achieves a training time of just 64.45 seconds—approximately 7.7% of the runtime of a standard vanilla NN, while requiring an order of magnitude fewer TFLOPs. This efficiency derives from the convex reformulation solved via ADMM and implemented in JAX, which enables highly parallelizable updates and rapid convergence. Unlike the vanilla NN which requires multiple passes and steps across the dataset for convergence with necessary hyperparameter grid search, the convex program uses a unique global optimum, thereby allowing us to solve directly to the global minima. Together, these properties establish CLD as both fast and efficient, offering a more practical alternative to conventional neural architectures for language detection.

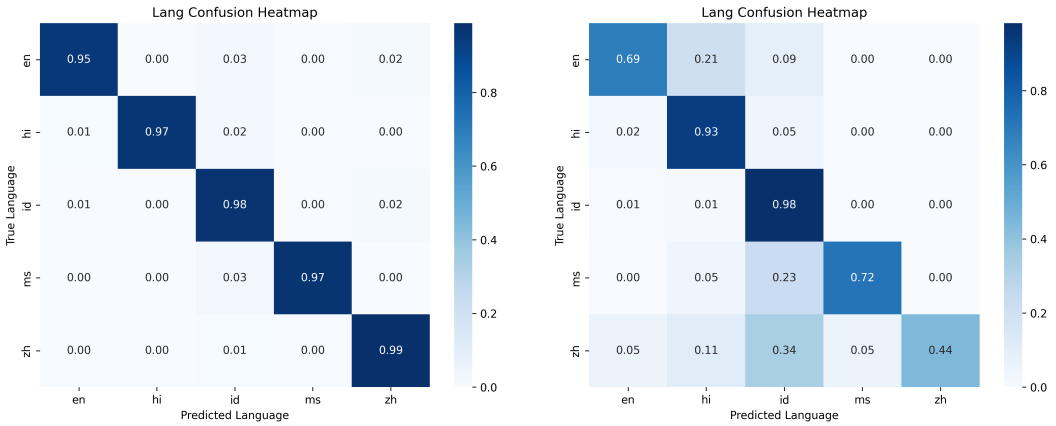
5.4 Multiclass Experiment

Table 2 reports the metrics on the scaled-up multiclass experiment and the confusion matrices for CVXNN and NN for in Figures 3a 3b respectievly. Compared to its performance in binary ex-

Model	Accuracy	WER	CER	F1
WSP	0.7154	139.37	73.85	0.7808
WSP-SFT	0.8033	35.85	22.30	0.8363
NN	0.7581	58.58	37.13	0.7492
CVXNN	**0.9715**	**31.74**	**17.84**	**0.9717**

Table 2: Performance comparison of models on the multiclass experiment, evaluated with accuracy, word error rate (WER), character error rate (CER), and F1 score.

periment, the NN model struggles to scale with multiple classes, lagging behind WSP-SFT in all metrics. In comparison, our CVXNN model still achieves a high accuracy of 97.15% and a F1 score of 0.9717 resulting in a WER of 31.74 and CER of 17.84, comparable with its performance in the previous binary task. This demonstrates that our model scales robustly to larger class counts, indicating strong potential for real-world language detection across diverse, in-the-wild conditions.



(a) Confusion matrix of true vs. predicted languages for CLD (ours)

(b) Confusion matrix of true vs. predicted languages for vanilla NN

Figure 3: Confusion matrices comparing CLD and NN model performance for language prediction

5.5 Regional Dialectical Accents Analysis

We showcase the language detection metrics of Vanilla Whisper (WSP), Finetuned Whisper (WSP-SFT), Traditional Neural Network Detection Head (NN), and our Convex Neural Network Detection head (cvxNN) across all 10 accents and the 10K, 1K, 500, 100 sample size in Table 4, Table 3, Table 5, Table 6 respectively. The respective accent names are listed in Table 7.

It is evident that traditional NN mostly performs well on English accents with consistently high to perfect accuracy rates, yet under performs on Mandarin accents, especially in lower sample sizes, achieving as low as 6.32% accuracy in zh-zh (Standard Mainland Mandarin) with a 500 sample size. This supports our argument where most traditional language models are English centric in a data perspective and thus achieve high performance exclusively in English, yet drops in performance on other less data-dominant and low-resource languages. In comparison, our novel CLD framework performs consistently high across all accents and languages with accuracy values all greater than 90%, demonstrating the model's robustness and low variance across diverse accents even with minimally available voice data.

5.6 Human Feedback and Discussion

Numerical tables of results and plots of WER are presented in Appendix B. Notably, in all cases varying runs on the same architecture (besides vanilla Whisper-Small) often produce similar WER. Therefore we perform analysis with real human testers based locally in Singapore and the Peoples Republic of China. Testers were instructed to assume the position of a general guest in a hospitality

Table 3: Performance Comparison for Accent Prediction Across Models @ 1K Sample Size

Accents	Samples Size	Correctly Predicted Samples				Accuracy			
		WSP	WSP-SFT	NN	CVXNN	WSP	WSP-SFT	NN	CVXNN
en-hi	190	18	175	190	185	0.9000	0.9211	**1.0000**	0.9737
en-my	215	12	115	215	214	0.5714	0.5349	**1.0000**	0.9953
en-sg	205	16	154	205	203	0.8000	0.7512	**1.0000**	0.9902
en-ur	189	17	180	189	187	0.9444	0.9524	**1.0000**	0.9894
en-us	204	17	195	204	203	0.9444	0.9559	**1.0000**	0.9951
zh-cdo	71	6	30	11	69	0.2727	0.4225	0.1549	**0.9718**
zh-cpx	216	3	82	47	216	0.2143	0.3796	0.2176	**1.0000**
zh-hk	184	18	143	53	182	0.6667	0.7772	0.2880	**0.9891**
zh-tw	181	15	181	33	181	0.8824	**1.0000**	0.1823	**1.0000**
zh-zh	205	20	192	45	204	0.8696	0.9366	0.2195	**0.9951**
Total	1860	122	1447	1192	**1844**	0.7066	0.7632	0.7062	**0.9900**

setting requesting an item. This ensures a precise and consistent conversational domain across all models. One example of vanilla Whisper-Small's output is below:

> **Concierge:** Hello Mr. Kevin Fong, this is Lucy at the front desk. How may I help you?
> **Guest:** Baru keadaan seperti seorang seorang seorang seperti seorang, seorang seorang berada di dalamnya.
> **Concierge:** I apologize, we'll send someone up right away. Do you need anything else?
> **Guest:** No, thank you.

Notably, although the guest was a local Singaporean person speaking naturally in his native English, the Whisper ASR model detected and transcribed this incorrectly into Bahasa. Experiments with the classic two-layer MLP model increased performance accuracy since errors became constrained between Singlish and mistakenly transcribed Mandarin characters (and vice versa). However a new type of error arose from the MLP detection head: local accents and dialects introduced errors such as the user speaking **'Both hot and cold settings'** to **'Both hood and coat setting'**. In contrast, our CLD algorithm produced the fastest and most accurate results: with both minimal word errors and the smallest numbers of wrong language detections.

6 Conclusion

In conclusion we conduct experiments with three variantes of ASR models in order to improve response accuracy in spoken dialogue systems for low-resource (yet high language) dialects. Vanilla Whisper-Small demonstrated the highest unstable WER, the vanilla NN constrained itself to the two languages we are interested in but introduced a different type of word error with frequent inaccurate transcriptions, and our Convex Language Detection algorithm received the highest human feedback in satisfaction with the lowest WER, which is particularly significant given its speed. Directions for future work include deeper analysis on the prosody of dialects within languages, and analysis on generalization possibilities for low resource dialects. Stronger theoretical interpretability of spoken dialogue models may also yield more valuable results, especially as scaling resources and training data become increasingly challenging.

Acknowledgments

During initial submission, we respectively omit the acknowledgments section in order to adhere to the the double-blind reviewing process.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. Building text and speech datasets for low resourced languages: A case of languages in east africa. In *3rd Workshop on African Natural Language Processing*, 2022.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*, 2018.

Sérgio Barbosa and Stefania Milan. Do not harm in private chat apps: Ethical issues for research on and with whatsapp. *Westminster Papers in Communication and Culture*, 14(1):49–65, 2019.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: Autograd and xla. *Astrophysics Source Code Library*, pp. ascl–2111, 2021.

Xabier de Zuazo, Eva Navas, Ibon Saratxaga, and Inma Hernáez Rioja. Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint arXiv:2503.23542*, 2025.

Miria Feng, Zachary Frangella, and Mert Pilanci. Cronos: Enhancing deep learning with scalable gpu accelerated convex neural networks. *arXiv preprint arXiv:2411.01088*, 2024.

Vivienne Fong, Lisa Lim, and Lionel Wee. "singlish": Used and abused. *Asian Englishes*, 5(1): 18–39, 2002.

Robbie BH Goh. The anatomy of singlish: globalisation, multiculturalism and the construction of the 'local'in singapore. *Journal of Multilingual and Multicultural Development*, 37(8):748–758, 2016.

Calbert Graham and Nathan Roll. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 2024.

Chng Huang Hoon. "you see me no up": Is singlish a problem? *Language Problems and Language Planning*, 27(1):45–62, 2003.

Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho Ittan George, Kaushal Santosh Bhogale, Deovrat Mehendale, and Mitesh M. Khapra. Lahaja: A robust multi-accent benchmark for evaluating hindi asr systems, 2024.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT press, 1997.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pp. 1529–1533, 2021.

Sungyoon Kim and Mert Pilanci. Convex relaxations of relu neural networks approximate global optima in polynomial time. In *International Conference on Machine Learning*, 2024.

Robert B Le Page. Retrospect and prognosis in malaysia and singapore. 1984.

Qiang Li, Qianyu Mai, Mandou Wang, and Mingjuan Ma. Chinese dialect speech recognition: a comprehensive survey. *Artificial Intelligence Review*, 57(2):25, 2024.

Yunpeng Liu, Xukui Yang, and Dan Qu. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29, 2024.

Antje S Meyer. Timing in conversation. *Journal of Cognition*, 6(1):20, 2023.

Aaron Mishkin, Arda Sahiner, and Mert Pilanci. Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions. In *International Conference on Machine Learning*, pp. 15770–15816. PMLR, 2022.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pp. 7695–7705. PMLR, 2020.

Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Rajeni Rajan. Tamil and tamils: A study of language and identity amongst the indian tamil community in singapore. *Curtin University, Perth, Western Australia*, 2018.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–17, 2022.

Rani Rubdy. Singlish in the school: An impediment or a resource? *Journal of Multilingual and Multicultural Development*, 28(4):308–324, 2007.

Tara N Sainath, Brian Kingsbury, Hagen Soltau, and Bhuvana Ramabhadran. Optimization techniques to improve training speed of deep neural networks for large speech tasks. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2267–2276, 2013.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. In *Proc. Interspeech*, pp. 27–31. ISCA, 2015.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7909–7913. IEEE, 2020.

Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F Chen, and AiTi Aw. Advancing singlish understanding: Bridging the gap with datasets and multimodal models. *arXiv preprint arXiv:2501.01034*, 2025.

Lionel Wee. *The Singlish controversy: Language, culture and identity in a globalizing world*. Cambridge University Press, 2018.

Felix Weninger, Yang Sun, Junho Park, Daniel Willett, and Puming Zhan. Deep learning based mandarin accent identification for accent robust asr. In *INTERSPEECH*, pp. 510–514, 2019.

D Randall Wilson and Tony R Martinez. The need for small learning rates on large problems. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 1, pp. 115–119. IEEE, 2001.

A Proof of Main Results and certificates for margin stability

This links between the convex program in Eq. (3) and the two-layer ReLU representation used in Theorem 1, together with computable certificates that translate directly into certified radii.

A.1 Activation patterns and pattern cones

For $u \in \mathbb{R}^d$ and a data matrix $H \in \mathbb{R}^{n\times d}$, define the training activation pattern

$$D(H,u) \;=\; \operatorname{diag}\big(\mathbf{1}\{Hu \geq 0\}\big) \in \{0,1\}^{n\times n}.$$

Given a fixed pattern $D \in \{0,1\}^{n\times n}$, define its associated *pattern cone*

$$\mathcal{K}(D) \;:=\; \big\{\, v \in \mathbb{R}^d : (2D - I)\, Hv \;\geq\; 0 \text{ (entrywise)} \big\}.$$

Then $v \in \mathcal{K}(D)$ if and only if $D(H, v) = D$ (up to measure-zero ties on $Hu = 0$). In particular, for $v \in \mathcal{K}(D)$ we have the identity

$$[Hv]_+ \;=\; D\,Hv, \tag{6}$$

where $[\cdot]_+$ is taken entrywise.

A.2 From the convex program to a two-layer ReLU

Recall the sampled-pattern convex model (Eq. (3)):

$$\min_{\{(v_i, w_i)\}_{i=1}^P} \ell\left(\sum_{i=1}^P D_i\, H\,(v_i - w_i),\, y\right) \;+\; \beta \sum_{i=1}^P \big(\|v_i\| + \|w_i\|\big) \quad \text{s.t.} \quad v_i, w_i \in \mathcal{K}(D_i). \tag{7}$$

Here $D_i \in \mathcal{D}_H$ are (sampled) activation patterns. In the multi-class case, take $v_i, w_i \in \mathbb{R}^{d\times K}$ with columns $(v_{i,1}, \ldots, v_{i,K})$ etc., and interpret $\|\cdot\|$ as either the block $\ell_{2,1}$ norm, $\|M\|_{2,1} = \sum_{k=1}^K \|M_{:,k}\|_2$, or the Frobenius norm.

Proposition 3 (Training-set representation)**.** *Let $\{(v_i, w_i)\}$ be any feasible point of equation 7. Then the training predictions equal those of a (vector-valued) two-layer ReLU network with at most $2PK$ hidden units:*

$$f(H) \;=\; \sum_{i=1}^P \sum_{k=1}^K \Big(e_k\,[Hv_{i,k}]_+ \;-\; e_k\,[Hw_{i,k}]_+\Big),$$

where e_k are the standard basis vectors in $\mathbb{R}^K$. Equivalently, this network has hidden weights $\{u_{i,k}^+, u_{i,k}^-\} = \{v_{i,k}, w_{i,k}\}$ and output weights $\{a_{i,k}^+, a_{i,k}^-\} = \{e_k, -e_k\}$.

A.3 Variation norm and a computable certificate

We use the standard two-layer ReLU variation norm:

$$\|f\|_{\mathrm{var}} := \inf\Big\{\sum_{j=1}^m \|a_j\|_2\,\|u_j\|_2 :\; f(h) = \sum_{j=1}^m a_j\,[u_j^\top h]_+,\; m \in \mathbb{N}\Big\}.$$

The following result turns any feasible solution of equation 7 into an *explicit upper bound* on $\|f\|_{\mathrm{var}}$, hence a Lipschitz certificate for the logits.

Theorem 4 (cvxNN $\Rightarrow$ variation-norm certificate)**.** *Let f be represented as in Proposition 3. Then*

$$\|f\|_{\mathrm{var}} \;\le\; \widehat{\mathcal{B}}_{\mathrm{cvx}}^{(2,1)} := \sum_{i=1}^P \big(\|v_i\|_{2,1} + \|w_i\|_{2,1}\big).$$

If equation 7 uses Frobenius penalties instead, then

$$\|f\|_{\mathrm{var}} \;\le\; \sqrt{K}\,\widehat{\mathcal{B}}_{\mathrm{cvx}}^{\mathrm{F}} \quad \textit{with} \quad \widehat{\mathcal{B}}_{\mathrm{cvx}}^{\mathrm{F}} := \sum_{i=1}^P \big(\|v_i\|_F + \|w_i\|_F\big).$$

Proof. Using the representation in Proposition 3, build a two-layer network whose hidden units are the *columns* $\{v_{i,k}\}_{i,k}$ and $\{w_{i,k}\}_{i,k}$ with output weights $\{+e_k\}$ and $\{-e_k\}$ respectively. For each unit (u, a) in this network, the atom cost is $\|a\|_2\|u\|_2 = \|u\|_2$ because $\|e_k\|_2 = 1$. Summing over units gives $\sum_{i,k}\big(\|v_{i,k}\|_2 + \|w_{i,k}\|_2\big) = \sum_i \big(\|v_i\|_{2,1} + \|w_i\|_{2,1}\big)$, which upper bounds $\|f\|_{\mathrm{var}}$ by definition. For Frobenius penalties, $\sum_k \|M_{:,k}\|_2 \le \sqrt{K}\|M\|_F$ yields the stated factor $\sqrt{K}$. □

By Lemma 1, $\|f(h) - f(h')\|_\infty \le \|f\|_{\mathrm{var}}\,\|h - h'\|_2$; combining with Theorem 1 yields the computable bounds

$$\mathrm{mar}(h+\delta, y) \;\ge\; \mathrm{mar}(h, y) - 2\,\widehat{\mathcal{B}}_{\mathrm{cvx}}^{(2,1)}\,\|\delta\|_2, \quad \mathrm{mar}(h+\delta, y) \;\ge\; \mathrm{mar}(h, y) - 2\,\sqrt{K}\,\widehat{\mathcal{B}}_{\mathrm{cvx}}^{\mathrm{F}}\,\|\delta\|_2,$$

depending on which penalty is used in Eq. (3). If the encoder E is L_E-Lipschitz, replace $\|\delta\|_2$ by $L_E\|x - x'\|_2$ to get the end-to-end certificate.

A.4 AM–GM LINK TO THE NONCONVEX ℓ_2^2 PENALTY

Consider the two-layer model $f(h) = \sum_{j=1}^{m} a_j [u_j^\top h]_+$ trained via the nonconvex penalty of Eq. (2): $(\beta/2)\sum_{j=1}^{m}(\|a_j\|_2^2 + \|u_j\|_2^2)$. By AM–GM, $2\|a_j\|_2\|u_j\|_2 \le \|a_j\|_2^2 + \|u_j\|_2^2$, hence

$$\|f\|_{\mathrm{var}} \;\le\; \frac{1}{2}\sum_{j=1}^{m}\big(\|a_j\|_2^2 + \|u_j\|_2^2\big) \;=\; \frac{1}{\beta}\,\frac{\beta}{2}\sum_{j=1}^{m}\big(\|a_j\|_2^2 + \|u_j\|_2^2\big).$$

Therefore any solution of Eq. (2) yields the certificate

$$\|f\|_{\mathrm{var}} \;\le\; \frac{1}{\beta}\,\mathcal{R}_{\ell_2^2} \quad\Rightarrow\quad \mathrm{mar}(h+\delta, y) \;\ge\; \mathrm{mar}(h, y) - \frac{2}{\beta}\,\mathcal{R}_{\ell_2^2}\,\|\delta\|_2.$$

Larger β tightens the bound linearly.

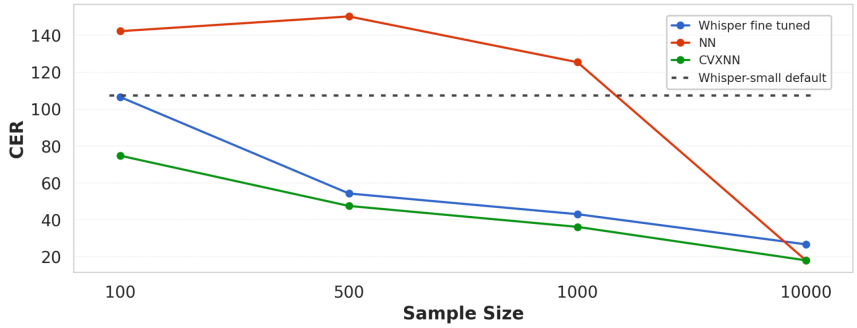
A.5 DETAILS FOR THE LOGIT LIPSCHITZ BOUND

Let $f(h) = \sum_j a_j [u_j^\top h]_+$. Since $t \mapsto [t]_+$ is 1-Lipschitz,

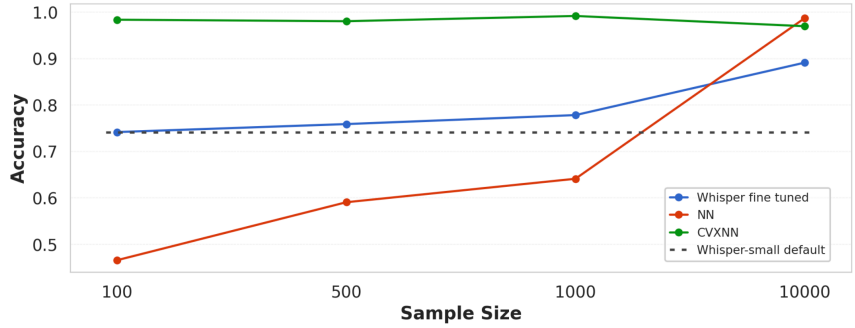
$$|f_k(h) - f_k(h')| = \Big|\sum_j a_{j,k}\big([u_j^\top h]_+ - [u_j^\top h']_+\big)\Big| \le \sum_j |a_{j,k}|\,|u_j^\top (h-h')| \le \sum_j \|a_j\|_2\|u_j\|_2\,\|h-h'\|_2.$$

Taking $\max_k$ and the infimum over all representations yields $\|f(h) - f(h')\|_\infty \le \|f\|_{\mathrm{var}}\|h - h'\|_2$, which is the Lemma used in Theorem 1.

B ADDITIONAL EMPIRICAL RESULTS



(a) Character Error Rate versus Sample Size



(b) Accuracy versus Sample Size

Figure 4: Error rates across different model configurations. Top panel shows word-level errors, middle panel shows character-level errors, bottom panel shows accuracy across all methods.

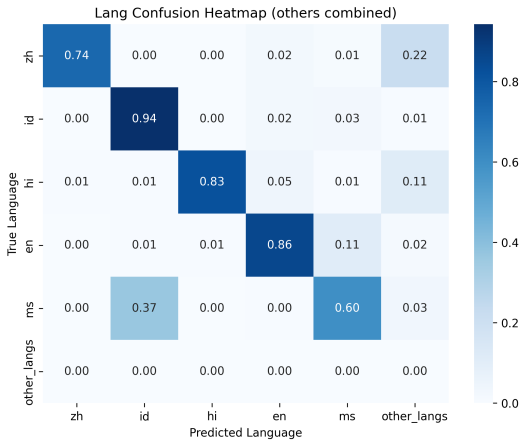


Figure 5: Confusion matrix of true vs. predicted languages for WSP-SFT

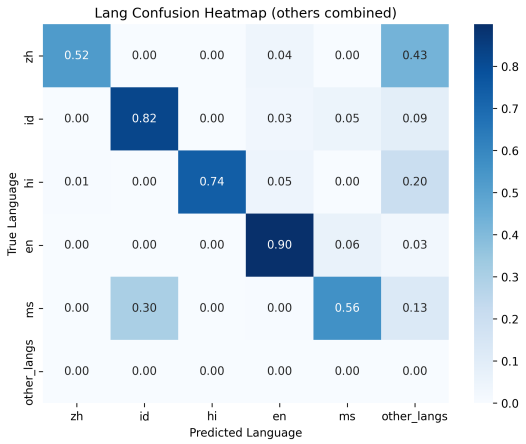


Figure 6: Confusion matrix of true vs. predicted languages for WSP

Table 4: Performance Comparison for Accent Prediction Across Models @ 10,000 Sample Size

Accents	**Samples Size**	**Correctly Predicted Samples**				**Accuracy**			
		WSP	**WSP-SFT**	**NN**	**CVXNN**	**WSP**	**WSP-SFT**	**NN**	**CVXNN**
en-hi	190	176	180	187	180	0.9263	0.9474	**0.9842**	0.9474
en-my	215	136	135	215	194	0.6326	0.6279	**1.0000**	0.9023
en-sg	205	166	166	204	195	0.8098	0.8098	**0.9951**	0.9512
en-ur	189	182	185	188	183	0.9630	0.9788	**0.9947**	0.9683
en-us	204	195	197	204	194	0.9559	0.9657	**1.0000**	0.9509
zh-cdo	71	7	50	70	71	0.0986	0.7042	0.9859	**1.0000**
zh-cpx	216	32	198	215	216	0.1481	0.9167	0.9954	**1.0000**
zh-hk	184	121	170	174	184	0.6576	0.9239	0.9457	**1.0000**
zh-tw	181	176	181	181	181	0.9724	**1.0000**	**1.0000**	**1.0000**
zh-zh	205	187	195	197	205	0.9122	0.9512	0.9610	**1.0000**
Total	1860	1378	1657	1815	1803	0.7077	0.8026	0.9162	**0.9721**

Table 5: Performance Comparison for Accent Prediction Across Models @ 500 Sample Size

Accents	Samples Size	Correctly Predicted Samples				Accuracy			
		WSP	WSP-SFT	NN	CVXNN	WSP	WSP-SFT	NN	CVXNN
en-hi	190	176	177	190	186	0.9263	0.9316	**1.0000**	0.9789
en-my	215	136	124	215	214	0.6326	0.5767	**1.0000**	0.9953
en-sg	205	166	162	205	205	0.8098	0.7902	**1.0000**	1.0000
en-ur	189	182	182	189	187	0.9630	0.9630	**1.0000**	0.9894
en-us	204	195	194	204	203	0.9559	0.9510	**1.0000**	0.9951
zh-cdo	71	7	15	21	63	0.0986	0.2113	0.2958	**0.8873**
zh-cpx	216	32	56	44	208	0.1481	0.2593	0.2037	**0.9630**
zh-hk	184	121	132	5	174	0.6576	0.7174	0.0272	**0.9457**
zh-tw	181	176	179	12	181	0.9724	**0.9890**	0.0663	1.0000
zh-zh	205	187	190	13	202	0.9122	0.9268	0.0634	**0.9854**
Total	1860	1378	1411	1088	1823	0.7077	0.7207	0.5576	**0.9695**

Table 6: Performance Comparison for Accent Prediction Across Models @ 100 Sample Size

Accents	Samples Size	Correctly Predicted Samples				Accuracy			
		WSP	WSP-SFT	NN	CVXNN	WSP	WSP-SFT	NN	CVXNN
en-hi	190	176	93	7	185	0.9263	0.9263	0.0368	**0.9737**
en-my	215	136	137	7	213	0.6326	0.6372	0.0326	**0.9907**
en-sg	205	166	166	3	203	0.8098	0.8098	0.0146	**0.9902**
en-ur	189	182	182	10	186	0.9630	**0.9630**	0.0529	0.9841
en-us	204	195	195	12	203	0.9559	0.9559	0.0588	**0.9951**
zh-cdo	71	7	7	67	71	0.0986	0.0986	0.9437	**1.0000**
zh-cpx	216	32	33	204	213	0.1481	0.1528	0.9444	**0.9861**
zh-hk	184	121	121	175	175	0.6576	0.6576	**0.9511**	0.9511
zh-tw	181	176	175	180	181	0.9724	0.9669	**0.9945**	1.0000
zh-zh	205	187	187	201	199	0.9122	0.9122	**0.9805**	0.9707
Total	1860	1378	1286	846	1829	0.7077	0.7102	0.3220	**0.9742**

Table 7: Mapping of Accent Codes to Accent Names

Code	Accent / Dialect Description
en-hi	Hindi-accented English
en-my	Malaysian-accented English
en-sg	Singaporean-accented English (Singlish)
en-ur	Pakistan-accented English
en-us	American English
zh-cdo	Min Dong Chinese / Fuzhou dialect Mandarin
zh-cpx	Pu-Xian Chinese
zh-hk	Hong Kong Cantonese
zh-tw	Taiwanese Mandarin
zh-zh	Standard Mainland Mandarin

C BASELINE WHISPER ASR ON LANGUAGE CLASSIFICATION ACCURACY

Table 8: Whisper ASR's default language detection accuracy by language and accent.

Language	Accent	Samples	Correct	Accuracy (%)
ID				
	Betawi	505	451	89.3
	Javanese	182	163	89.6
	Jawa Tengah	228	206	90.4
	Surakarta	221	200	90.5
	Bindeng	221	200	90.5
	Tionghoa	221	200	90.5
	Medhok	224	203	90.6
EN				
	Malaysian English	1000	614	61.4
	Filipino	1000	745	74.5
	Singaporean English	1000	752	75.2
	Zimbabwe	1000	831	83.1
	Southern African (South Africa, Namibia)	1000	831	83.1
	Welsh English	1000	918	91.8
	Scandinavian	1000	952	95.2
	Pakistan	1000	967	96.7
	India and South Asia (India, Sri Lanka)	1000	967	96.7
	Scottish English	1000	968	96.8
	Lancashire English	1000	970	97.0
	Liverpool English	1000	970	97.0
	England English	1000	982	98.2
	Australian English	1000	984	98.4
	United States English	1000	985	98.5
	New Zealand English	1000	987	98.7
	Hong Kong English	1000	988	98.8
	German English	1000	996	99.6
	Non native speaker	1000	996	99.6
	Irish English	1000	996	99.6
	Canadian English	1000	999	99.9
	Northern Irish	1000	999	99.9
	Low	1000	999	99.9
	Demure	1000	999	99.9
	Midwestern	1000	1000	100.0
	Transatlantic English	1000	1000	100.0
ZH				

Table 9: Human feedback for Vanilla Whisper-Small.

Input Source Language	Total Test Prompts	Wrong Language Transcribed
EN (5 human testers in Singapore)	595	59
ZH (10 human testers in South-East China)	300	148

Language	Accent	Samples	Correct	Accuracy (%)
	cdo	1000	103	10.3
	cpx	1000	111	11.1
	nan-tw	1000	545	54.5
	hk	1000	905	90.5
	zh	1000	910	91.0
	tw	1000	987	98.7
HI				
	Kashmiri	320	185	57.8
	Bodo	380	270	71.1
	Malayalam	365	271	74.2
	Punjabi	236	178	75.4
	Urdu	131	103	78.6
	Telugu	474	377	79.5
	Tamil	182	145	79.7
	Hindi	575	468	81.4
	Sindhi	141	116	82.3
	Odia	374	311	83.2
	Kannada	313	268	85.6
	Gujarati	299	257	86.0
	Assamese	262	226	86.3
	Konkani	422	364	86.3
	Nepali	359	311	86.6
	Bengali	418	366	87.6
	Dogri	219	194	88.6
	Maithili	287	255	88.9
	Marathi	395	366	92.7
MS				
	msi	1000	386	38.6
	ms	1000	934	93.4

D HUMAN FEEDBACK VALIDATION

Table 10: Human feedback for vanilla NN detection head.

Input Source Language	**Total Test Prompts**	**Wrong Language Transcribed**	**Word Errors in Transcription**
EN (same 5 humans)	450	22	81
ZH (same 10 humans)	450	5	14

Table 11: Human feedback for Convex Language Detection (CLD).

Input Source Language	**Total Test Prompts**	**Wrong Language Transcribed**	**Word Errors in Transcription**
EN (same 5 humans)	450	12	26
ZH (same 10 humans)	450	2	14