OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Convex Neural Networks For Robust ASR Language Detection

Miria Feng MIRIA00@STANFORD.EDU
Mert Pilanci PILANCI@STANFORD.EDU
Stanford University, California

Abstract

Globalization and multiculturalism have produced numerous diverse dialects, such as Singaporean-accented English and regional Mandarin speech. These speech variants remain significantly underrepresented even in high-resource language datasets. Consequently, standard spoken dialogue systems frequently misidentify the user's input language, compromising response accuracy regardless of downstream language model capability. To address this, we propose a robust ASR framework capable of handling dialectal variance with minimal computational overhead and lightweight training costs. Our Convex Language Detection (CLD) framework integrates a convex neural network that guarantees global optimality in polynomial time. This is solved efficiently using ADMM in JAX, achieving sub-500ms inference latency. CLD offers strong convergence guarantees, stability across runs, and reduced sample complexity. As a motivating case study, CLD significantly improves transcription accuracy on bilingual inputs when integrated with Whisper encoders. These results enable more inclusive multilingual interactions and highlight promising directions for convex optimization methods in spoken dialogue systems.

1. Introduction

Spoken language dialogue systems are increasingly ubiquitous across all cultures, countries, and applications. The mainstream adoption of video-conferencing, Siri voice assistant [1], live transcription, and voice navigation has driven much research into advancing state-of-the-art Automatic Speech Recognition (ASR) models [43]. ASR is the crucial and common component in these systems, and serves to transcribe input user speech into text for downstream large language models (LLMs) to process. Without accurate transcription, even the most advanced LLMs cannot correctly interpret user intent or generate accurate responses.

The widely adopted Whisper [29] ASR model series demonstrates strong ability to generalize to many datasets and domains in a zero-shot setting, yet frequently misidentifies the input language token due to user dialects and accents [15]. This occurs since ASR performance varies directly as a function of speaker characteristics such as dialect [26], gender [41], and cultural background [24]. However, addressing this performance variability is critical for building communication systems that are inclusive and accessible to a global audience [20]. For example, although the national language of Singapore is English (the most dominant language in voice datasets [7]), the unique and prevalent dialect of Singaporean accented English has led to the colloquial term "Singlish" [39]. The intonation and prosody of Singlish is so distinct that it has been widely studied by linguists [14], [16], [31], yet state-of-the-art ASR models often mistakenly transcribe Singaporean native English into an incorrect language (such as Bahasa [21]).

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. The scope of this work is 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 speech sequences, than the standard linear detection layer in existing Whisper models. This efficiency is crucial, since end-to-end speech dialogue models require sub-500ms latency [25] to preserve realistic human response time. We further optimize for fast training and iteration by implementing our method in JAX [8] and solving the optimization problem with ADMM based techniques [12].

The following paper is presented as follows: Section 2 outlines related work, Section 3 introduces our Convex Language Detection method, Section 4 provides main experimental results and theoretical analysis, finally Section 5 summarizes conclusions and directions for future work.

2. Related Work

Multilingual ASR models with up to 1550 Million parameters and have been trained on 99+ languages [29]. However the vast majority of these models perform the best on English, with performance dropping significantly on lower resource languages [15]. This has recently encouraged much work in the field of improving low-resource ASR performance. For example, the authors of [5], [18], [34] 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 successful. During the course of pretraining on extremely large datasets, the powerful base ASR model also experiences catastrophic forgetting [13], 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 [22], [40], and [38] 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 [30] have focused on more clearly identifying the challenges ASR models face with low-resource languages. These methods all 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 [42], [23], [10].

Instead of relying heavily on pretraining, fine-tuning, or gathering more data: our key insight is that the existing Whisper ASR models have already been trained on 680 000+ hours of speech-transcription data [29]. Therefore instead of relying on traditional resource intensive techniques, we focus on implementing a fast and efficient language detection modification layer embedded inside the Whisper architecture that 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. Additional discussion on related work continues in Appendix B.

3. Convex Language Detection Algorithm

In this section we introduce the Convex Language Detection (CLD) architecture for ASR models. Section 3.1 provides preliminaries on two-layer ReLU networks, Section 3.2 introduces the equivalent convex optimization reformulated neural network (cvxNN), and Section 3.3 present its integration within the language detection framework.

3.1. Two-layer ReLU Networks

Let $x \in \mathbb{R}^d$ represent the input, $\Theta_1 \in \mathbb{R}^{m\times d}$, $\theta_2 \in \mathbb{R}^m$ represent weights of the first and last layers respectively, and $(\cdot)_+ = \max\{\cdot, 0\}$ represent the ReLU activation function. The classic two-layer ReLU network is then given by:

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

Given targets $y \in \mathbb{R}^n$, the network in (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. (2) presents a challenging non-convex optimization problem, with necessary iterations of hyperparameter grid-search for successful training. This approach becomes exceedingly expensive as we scale to high-dimensional audio and speech datasets, which are inherently slower to train and more resource-intensive [32]. Therefore our goal is to maintain these expressive capabilities while still preserving the computational advantages of convex optimization.

3.2. Equivalent Convex Reformulation

Given that $m \geq m^*$, for some $m \geq n+1$, (2) yields a convex reformulation with same optimal value as the original non-convex problem [28]. This is based on enumerating the actions of all possible ReLU activation patterns on data matrix X, which act as separating hyperplanes represented by diagonal matrices. For fixed X, the set of all possible ReLU activation patterns may then be expressed as:

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

The cardinality of $\mathcal{D}_X$ grows as $|\mathcal{D}_X| = \mathcal{O}\left(r(n/r)^r\right)$, where $r := \text{rank}(X)$ [28]. Since the exponential size of $\mathcal{D}_X$ [28] make its complete enumeration impractical, we work with a subset based on sampling P patterns from $\mathcal{D}_X$:

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

It can be shown under mild conditions that (3) has the same optimal solution as (2) [27]. The recent work of [19] also proves that the difference is negligible even when they are not equal. Therefore we work with the tractable convex framework in (3) and no information is lost.

3.3. Integration with Spoken Language Systems

Recent work of [12] has demonstrated the successful application of cvxNN on high-dimensional text-based LLMs. 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)

Whisper encoder $\mathcal{E}$; decoder $\mathcal{D}$

Training (offline):
for $i \leftarrow 1$ **to** N **do**
  $h_i \leftarrow \mathcal{E}(x_i)$ `// Extract hidden states`
end
Train cvxNN on $\{(h_i, y_i)\}$ using ADMM with variables $(\mathbf{v}, \mathbf{w}, \mathbf{u})$ and penalty ρ
while *not converged* **do**
  $(\mathbf{v}, \mathbf{w}) \leftarrow \arg\min\ \ell\Big(\sum_{p=1}^{P} D_p H(\mathbf{v}_p - \mathbf{w}_p), y\Big) + \beta \sum_{p=1}^{P} (\|\mathbf{v}_p\|_2 + \|\mathbf{w}_p\|_2) + \frac{\rho}{2}\| \cdot \|_2^2$
  $\mathbf{u} \leftarrow \mathbf{u} +$ (primal residual)
end
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 the 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. Main Experiments and Analysis

This section presents our experimental results and analysis. We observe that while the standard Word Error Rate (WER) metric [17] often yields similar scores across configurations, human evaluation reveals dramatic differences in model performance. This discrepancy arises since automated metrics for spoken dialogue systems serve only as an *approximation* of true user experience. Appendix A.2 details our datasets, Section 4.1 presents a theoretical analysis of optimization and statistical generalization, and Section 4.2 summarizes our evaluation results. We benchmark the CLD algorithm against two baselines: the unmodified Whisper-Small (244M parameters) and a standard ASR pipeline augmented with a trained two-layer MLP for bilingual language detection.

4.1. Optimization and Statistical Generalization

To formalize the optimization and generalization properties of the CLD module, we demonstrate that increasing the number of sampled activation patterns P in Eq 3 incurs only a logarithmic cost in sample complexity. Consequently, we can over-parameterize P to maximize expressivity while controlling capacity through the encoder radius R and the effective norm bound B (governed by the parameter β). This provides a theoretical guarantee that CLD maintains controlled capacity and resists overfitting, even in low-sample regimes. Appendix C gives the proof of Theorem 1 below.

Recall that CLD module is trained by solving the convex reformulation in Eq 3. Let $(x_i, y_i)_{i=1}^N$ denote the training samples and $h_i = E(x_i) \in \mathbb{R}^d$ denote the corresponding Whisper encoder states. Let $H \in \mathbb{R}^{N\times d}$ be the matrix whose i-th row is $h_i^\top$. The convex CLD module (with P sampled activation patterns) can be written as:

$$f_{v,w}(h) \;=\; \sum_{p=1}^{P} \big\langle v_p - w_p,\, D_p h \big\rangle, \quad \text{with parameters} \quad (v_p, w_p) \in K_p \subset \mathbb{R}^d.$$

Formally, during training the parameters $\{(v_p, w_p)\}_{p=1}^P$ are obtained by solving the convex problem:

$$\min_{(v_p,w_p)_{p=1}^P} \; \ell\Big(\sum_{p=1}^{P} D_p H(v_p - w_p),\, y\Big) \;+\; \beta \sum_{p=1}^{P} \big(\|v_p\|_2 + \|w_p\|_2\big),$$

where $y \in \mathbb{R}^N$ is the vector of language labels and $\beta > 0$ is the regularization parameter. For a predictor $f_{v,w}$, define the population and empirical risks as:

$$L(f_{v,w}) \;:=\; \mathbb{E}_{(h,y)}\big[\ell(f_{v,w}(h), y)\big], \qquad \hat{L}_N(f_{v,w}) \;:=\; \frac{1}{N}\sum_{i=1}^{N} \ell\big(f_{v,w}(h_i), y_i\big).$$

Theorem 1 (Sample complexity of CLD) *Assume that the encoder states are uniformly bounded as $\|h_i\|_2 \le R$ for all $i \in [N]$, and consider the class of CLD predictors*

$$\mathcal{F}_B \;=\; \Big\{ f_{v,w} \;:\; f_{v,w}(h) = \sum_{p=1}^{P} \langle v_p - w_p, D_p h\rangle,\; \sum_{p=1}^{P} \big(\|v_p\|_2 + \|w_p\|_2\big) \le B \Big\}.$$

Let $\ell(\cdot,\cdot)$ be convex in its first argument, 1-Lipschitz in that argument, and bounded in $[0,1]$. Then there exists a universal constant $C > 0$ such that, with probability at least $1-\delta$ over the training set draw,

$$L(f_{v,w}) - \hat{L}_N(f_{v,w}) \;\le\; C\,\frac{RB}{\sqrt{N}}\left(\sqrt{\log P} + \sqrt{\log(1/\delta)}\right)$$

simultaneously for all $f_{v,w} \in \mathcal{F}_B$.

The bound depends logarithmically on the number of sampled activation patterns P. The main complexity term is primarily controlled by the encoder radius R, the sample size N, and the effective norm bound B. Since standard KKT conditions imply $B = \mathcal{O}(1/\beta)$, increasing the regularization parameter β directly tightens the generalization bound without altering the convex landscape. These theoretical properties translate into two distinct practical advantages for spoken dialogue systems:

1. **Global Optimality:** Since the CLD training objective is convex, any local minimum is global. Unlike traditional non-convex MLPs, which are sensitive to initialization and optimizer trajectories, CLD converges to a global optimal solution (up to numerical precision). This supports the observed empirical stability and performance in Section 4.2.

2. **Sample Efficiency:** Theorem 1 guarantees that sample complexity scales with $\sqrt{\log P}$. This implies we can increase the expressiveness of the CLD module (by increasing P) without requiring a proportionally large increase in expensive labeled audio data. This is critical for speech and regional dialect applications, where annotated corpora are scarce and challenging to curate. In our setting, the encoder radius R is implicitly bounded by the Whisper architecture.

4.2. Human Feedback and Discussion

Numerical results are presented in Appendix A.1. Notably, in all cases varying runs on the same architecture (despite varying configurations) often produce similar WER. Therefore we perform evaluation and validate our results 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 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 human evaluator was a local Singaporean person speaking naturally in his native English, the Whisper ASR model detected and transcribed this incorrectly into Bahasa. Experiments with concatenating the trained MLP model for bilingual language detection increased performance accuracy, as errors became constrained between English 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 transcribing 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 WER and the smallest numbers of wrong language detections. Table 1 presents numerical results of real human evaluation across three models.

5. Conclusion and Future Work

We introduce the Convex Language Detection (CLD) framework, a theoretically grounded approach for robust language identification in bilingual spoken dialogue systems. CLD proves the practical advantages of principled convex optimization in the non-convex landscape of deep learning. The motivating application of CLD for bilingual language detection under dialectal variation demonstrates the practical advantages of our optimization framework in real world settings. We achieve superior sample efficiency and guaranteed global optimality, as demonstrated in CLD's consistent results while mitigating expensive hyperparameter grid-search. Future work will explore scaling CLD to larger speech frameworks across challenging and realistic multi-dialect multi-cultural settings, as well as deeper evaluations to assess computational efficiency.

5.1. Acknowledgments

We thank Lucy Woof for support throughout, and Andrew Maas for many insightful discussions. We also thank Kevin Nam for feedback on an early draft of this paper, and the contribution of FCS Solutions to the human feedback evaluation.

This work was supported in part by National Science Foundation (NSF) CAREER Award under Grant CCF-2236829; in part by the U.S. Army Research Office Early Career Award under Grant W911NF-21-1-0242; in part by the Office of Naval Research under Grant N00014-24-1-2164. In addition, Miria Feng was supported in part by the Stanford Graduate Fellowship.

References

[1] Siri - apple voice assistant. https://www.apple.com/siri/. Accessed: 2025-11-25.

[2] Infocomm media development authority (imda) – singapore's digital future. https://www.imda.gov.sg/. Accessed: 2025-11-26.

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Reuben Henretty, Michael Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222. European Language Resources Association (ELRA), 2020. URL https://commonvoice.mozilla.org.

[4] 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.

[5] 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.

[6] 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.

[7] David E. Blasi et al. The dominance of english in llm training data. *Kili Technology Blog*, 2024. URL https://kili-technology.com/large-language-models-llms/9-open-sourced-datasets-for-training-large-language-models. English accounts for approximately 46% to over 90% of training tokens in major LLM datasets such as Common Crawl and GPT-3's corpus, greatly outweighing other languages.

[8] 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*, pages ascl–2111, 2021.

[9] Alexis Conneau, Jean Maillard, Mathis Riviere, Kushal Lakhotia, Shigeki Karita, Julien Rostaing, Jade Copet, Yulia Tsvetkov, Abdelrahman Mohamed, Gabriel Synnaeve, et al. Fleurs: Few-shot learning evaluation of universal representations of speech. *Transactions of the*

Association for Computational Linguistics, 11:149–168, 2023. doi: 10.1162/tacl_a_00536. URL https://aclanthology.org/2023.tacl-1.9.

[10] 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.

[11] FCS Solutions. Hotel operations management software. https://www.fcshub.com/en, 2025. Accessed: 2025-11-26.

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

[13] Robert M. French. Catastrophic forgetting in connectionist networks. In *Trends in cognitive sciences*, volume 3, pages 128–135. Elsevier, 1999.

[14] 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.

[15] C. Graham and N. Roll. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2):025206, 2024. doi: 10.1121/10.0024876.

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

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

[18] 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*, pages 1529–1533, 2021.

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

[20] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689, 2020.

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

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

[23] 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.

[24] Joshua L Martin and Kevin Tang. Understanding racial disparities in automatic speech recognition: The case of habitual" be". In *Interspeech*, pages 626–630, 2020.

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

[26] Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6462–6468, 2020.

[27] 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*, pages 15770–15816. PMLR, 2022.

[28] 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*, pages 7695–7705. PMLR, 2020.

[29] 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.

[30] 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*, pages 1–17, 2022.

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

[32] 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.

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

[34] 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)*, pages 7909–7913. IEEE, 2020.

[35] Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

[36] Vladimir N. Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.

[37] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

[38] 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.

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

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

[41] Barbara Wheatley and Joseph Picone. Voice across america: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, 1(2):45–63, 1991.

[42] 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, pages 115–119. IEEE, 2001.

[43] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE, 2018.

Appendix A. Metrics and Datasets

A.1. Evaluation Metrics

This section provides tables of experimental results corresponding to Section 4. Real human blind evaluation was conducted with five human participants in Singapore speaking native English (EN) and ten human participants in People's Republic of China speaking native Mandarin (ZH). All persons were consenting members of the [11] team. With respect to individual privacy, please contact the authors directly for individual names of human samples and correspondence. The CLD-augmented ASR model consistently received the highest human satisfaction with the lowest error rates.

Method	**Total Test Prompts**	**Wrong Language Transcribed**	**Word Error Rate in Transcription (WER)**
Default			
EN	595	59	–
ZH	300	148	–
MLP			
EN	450	22	81
ZH	450	5	14
CLD (ours)			
EN	450	12	26
ZH	450	**2**	**14**

Table 1: Human evaluation across three models: Whisper-small using its *Default* automatic language detection layer, *MLP*-augmented Whisper-small for enhanced language detection, and the *CLD*-augmented Whisper-small architecture.

A.2. Datasets

Although multilingual voice-transcript datasets exist, such as Mozilla Common Voice [3] and Google Fleur [9], there are few datasets for regional dialects and accents. Therefore in order to produce the most accurate results for our experimentation, we utilize the direct dataset from the Info-communications and Media Development Authority (IMDA) of Singapore [2]. Through IMDA, we were given access to National Speech Corpus (NCS): the first Singapore English corpus which aims to become a valuable resource for researchers and developers working on AI technology in Singapore and South East Asia. After cleaning the 2TB NCS dataset into matching voice-transcript pairs, we were left with 2576 training data samples. We augment these via the following techniques: Time stretching, volume gain, pitch shift, and recorded background noise (via MUSAN [33]) are used to simulate real-world variability and improve robustness. This yields approximately 4500 training data samples for Singlish. We then match this with an equivalent 4500 training data samples for Mandarin from the Common Voice (v16) dataset.

Appendix B. Related Work

Other researchers [30] have focused on more clearly identifying the challenges ASR models face with low-resource languages, such as Xhosa or Marathi. Limited training data is a dominant issue, and authors [4] have worked on building partnerships to preserve and document linguistics 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 [6], and is slow to provide clearly annotated data from numerous low-resource languages.

Appendix C. Proof of Main Results

In this section we formally provide the proof of Theorem 1. Recall that for a predictor f we write

$$L(f) := \mathbb{E}_{(h,y)}\big[\ell(f(h), y)\big], \qquad \hat{L}_N(f) := \frac{1}{N}\sum_{i=1}^{N} \ell\big(f(h_i), y_i\big),$$

and let $\mathcal{F}_B$ denote the class of CLD predictors defined in Theorem 1.

Lipschitz contraction. Let $S = \{(h_i, y_i)\}_{i=1}^N$ denote the training sample and define

$$\Phi(S) \;:=\; \sup_{f \in \mathcal{F}_B} \big(L(f) - \hat{L}_N(f)\big).$$

We first bound its expectation in terms of the empirical Rademacher complexity of $\mathcal{F}_B$. For a fixed sample $(h_1, \ldots, h_N)$, the empirical Rademacher complexity of $\mathcal{F}_B$ is

$$R_N(\mathcal{F}_B) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_B} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i f(h_i) \;\middle|\; h_1, \ldots, h_N \right],$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are independent Rademacher variables taking values in $\{\pm 1\}$.

By the standard symmetrization lemma for bounded losses [36], we have

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S \left[\sup_{f \in \mathcal{F}_B} \big(L(f) - \hat{L}_N(f)\big) \right] \;\le\; 2\,\mathbb{E}_S\, R_N(\ell \circ \mathcal{F}_B),$$

where $\ell \circ \mathcal{F}_B := \{(h, y) \mapsto \ell(f(h), y) : f \in \mathcal{F}_B\}$.

We apply the Lipschitz contraction inequality. Since $\ell(\cdot, y)$ is 1-Lipschitz in its first argument and bounded in $[0, 1]$, the contraction property of Rademacher averages implies

$$R_N(\ell \circ \mathcal{F}_B) \;\le\; R_N(\mathcal{F}_B).$$

Combining the two yields

$$\mathbb{E}_S[\Phi(S)] \;\le\; C_1\, \mathbb{E}_S\big[R_N(\mathcal{F}_B)\big] \tag{4}$$

for some universal numerical constant $C_1 > 0$. Thus, to control the expected generalization gap it suffices to bound $R_N(\mathcal{F}_B)$.

Rademacher complexity of the CLD class. We now bound $R_N(\mathcal{F}_B)$ for the specific CLD architecture. Fix a realization of the encoder states $\{h_i\}_{i=1}^N$ and the activation pattern matrices $\{D_p\}_{p=1}^P$. For any $(v_p, w_p)_{p=1}^P$ we therefore write

$$f_{v,w}(h) = \sum_{p=1}^{P} \langle v_p - w_p,\, D_p h \rangle, \qquad \sum_{p=1}^{P} \big(\|v_p\|_2 + \|w_p\|_2\big) \le B,$$

and let $u_p := v_p - w_p$. By the triangle inequality,

$$\sum_{p=1}^{P} \|u_p\|_2 \;\le\; \sum_{p=1}^{P} \big(\|v_p\|_2 + \|w_p\|_2\big) \;\le\; B,$$

thus the constraint on (v_p, w_p) induces the constraint $\sum_{p=1}^P \|u_p\|_2 \le B$.

Conditioned on $(h_i)_{i=1}^N$, the empirical Rademacher complexity of $\mathcal{F}_B$ is

$$\begin{aligned} R_N(\mathcal{F}_B) &= \mathbb{E}_\varepsilon \left[\sup_{\sum_p (\|v_p\|_2 + \|w_p\|_2) \le B} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \sum_{p=1}^{P} \langle v_p - w_p,\, D_p h_i \rangle \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{\sum_p \|u_p\|_2 \le B} \frac{1}{N} \sum_{p=1}^{P} \left\langle u_p,\, \sum_{i=1}^{N} \varepsilon_i D_p h_i \right\rangle \right]. \end{aligned}$$

For each pattern p define the random vector

$$A_p := \sum_{i=1}^{N} \varepsilon_i D_p h_i \in \mathbb{R}^d.$$

Then we can rewrite

$$R_N(\mathcal{F}_B) = \frac{1}{N}\, \mathbb{E}_\varepsilon \left[\sup_{\sum_p \|u_p\|_2 \le B} \sum_{p=1}^{P} \langle u_p, A_p \rangle \right].$$

For any feasible family $\{u_p\}_{p=1}^P$ we have, by Cauchy–Schwarz and the ℓ_1/ℓ_∞ Hölder inequality,

$$\sum_{p=1}^{P} \langle u_p, A_p \rangle \;\le\; \sum_{p=1}^{P} \|u_p\|_2 \, \|A_p\|_2 \;\le\; \Big(\sum_{p=1}^{P} \|u_p\|_2 \Big) \max_{1 \le p \le P} \|A_p\|_2 \;\le\; B \max_{1 \le p \le P} \|A_p\|_2.$$

Therefore we can conclude

$$R_N(\mathcal{F}_B) \;\le\; \frac{B}{N}\, \mathbb{E}_\varepsilon \Big[\max_{1 \le p \le P} \|A_p\|_2 \Big]. \tag{5}$$

By assumption, $\|h_i\|_2 \le R$ for all i and each D_p is a diagonal matrix with entries in $\{0, 1\}$. Therefore

$$\|D_p h_i\|_2 \;\le\; \|h_i\|_2 \;\le\; R \qquad \text{for all } i, p.$$

Define $z_{i,p} := D_p h_i$. Then $A_p = \sum_{i=1}^N \varepsilon_i z_{i,p}$ is a sum of independent, mean-zero random vectors with $\|z_{i,p}\|_2 \le R$.

For any fixed p and any unit vector $u \in \mathbb{S}^{d-1}$,

$$\langle u, A_p \rangle = \sum_{i=1}^N \varepsilon_i \langle u, z_{i,p} \rangle.$$

Each term $\varepsilon_i \langle u, z_{i,p} \rangle$ is mean-zero and bounded in absolute value by R. A standard moment generating function bound (or Hoeffding's inequality) implies that $\langle u, A_p \rangle$ is sub-Gaussian with parameter at most $R\sqrt{N}$, i.e.

$$\mathbb{P}\big(|\langle u, A_p \rangle| \ge t\big) \;\le\; 2\exp\!\left(-\frac{t^2}{2NR^2}\right) \qquad \text{for all } t > 0.$$

By covering the unit sphere with a finite ε-net and using the fact that $\|A_p\|_2 = \sup_{\|u\|_2=1} \langle u, A_p \rangle$, we obtain the classical bound for maxima of finitely many sub-Gaussian vectors [37], [35]:

$$\mathbb{E}_\varepsilon\Big[\max_{1 \le p \le P} \|A_p\|_2\Big] \;\le\; C_2\, R\, \sqrt{N \log P}, \tag{6}$$

for some universal constant $C_2 > 0$. Substituting (6) into (5) yields the bound

$$R_N(\mathcal{F}_B) \;\le\; C_2\, \frac{RB}{\sqrt{N}}\, \sqrt{\log P}. \tag{7}$$

Since the right-hand side depends on (h_i) only through the uniform norm bound $\|h_i\|_2 \le R$, (7) holds for every sample S.

Combining (4) and (7) gives

$$\mathbb{E}_S[\Phi(S)] \;\le\; C\, \frac{RB}{\sqrt{N}}\, \sqrt{\log P} \tag{8}$$

for some universal $C > 0$.

Utilizing McDiarmid's inequality. Consider the function

$$\Phi(S) = \sup_{f \in \mathcal{F}_B} \big(L(f) - \hat{L}_N(f)\big)$$

as a function of the N independent random variables $(h_1, y_1), \ldots, (h_N, y_N)$. If we change a single sample (h_j, y_j) to (h'_j, y'_j) while keeping all other points fixed, then for any fixed f we have

$$\big|\hat{L}_N(f; S) - \hat{L}_N(f; S')\big| = \frac{1}{N}\, \big|\ell(f(h_j), y_j) - \ell(f(h'_j), y'_j)\big| \;\le\; \frac{1}{N},$$

because ℓ takes values in $[0, 1]$. Taking the supremum over $f \in \mathcal{F}_B$ and noting that $L(f)$ does not depend on the empirical sample, we obtain

$$|\Phi(S) - \Phi(S')| \;\le\; \frac{1}{N}.$$

Thus $\Phi(S)$ satisfies a bounded-differences condition with parameters $c_i = 1/N$ for all i. McDiarmid's inequality yields, for any $t > 0$,

$$\mathbb{P}\Big(\Phi(S) - \mathbb{E}_S[\Phi(S)] \geq t\Big) \;\leq\; \exp\left(-\frac{2t^2}{\sum_{i=1}^{N} c_i^2}\right) = \exp\big(-2Nt^2\big)\,.$$

Setting $t = \sqrt{\frac{1}{2N}\log(1/\delta)}$ and rearranging, we obtain that with probability at least $1-\delta$,

$$\Phi(S) \;\leq\; \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\log(1/\delta)}{2N}}.$$

Combining this with (8) gives

$$\sup_{f \in \mathcal{F}_B} \big(L(f) - \hat{L}_N(f)\big) \;\leq\; C\,\frac{RB}{\sqrt{N}}\,\sqrt{\log P} \;+\; \sqrt{\frac{\log(1/\delta)}{2N}}.$$

Since the bound holds uniformly over $f \in \mathcal{F}_B$, it holds in particular for any fixed CLD predictor $f_{v,w} \in \mathcal{F}_B$. Absorbing constants into C and collecting terms yields the stated form of Theorem 1, namely

$$L(f_{v,w}) - \hat{L}_N(f_{v,w}) \;\leq\; C\,\frac{RB}{\sqrt{N}}\left(\sqrt{\log P} \;+\; \sqrt{\log(1/\delta)}\right),$$

up to a change of the universal constant $C > 0$. This completes the proof of Theorem 1. □