Investigating model performance in language identification beyond simple error statistics

Styles, Suzy J.; Chua, Victoria Y.H.; Woon, Fei Ting; Liu, Hexin; Perera, Leibny Paola Garcia; Khudanpur, Sanjeev; Khong, Andy W.H.; Dauwels, Justin

DOI
10.21437/Interspeech.2023-1707

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

Citation (APA)
Styles, S. J., Chua, V. Y. H., Woon, F. T., Liu, H., Perera, L. P. G., Khudanpur, S., Khong, A. W. H., & Dauwels, J. (2023). Investigating model performance in language identification: beyond simple error statistics. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2023-August*, 4129-4133. https://doi.org/10.21437/Interspeech.2023-1707

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Investigating model performance in language identification: beyond simple error statistics

Suzy J. Styles[1], *Victoria Y. H. Chua*[1], *Fei Ting Woon*[1], *Hexin Liu*[2], *Leibny Paola Garcia Perera*[3], *Sanjeev Khudanpur*[3], *Andy W. H. Khong*[2], *Justin Dauwels*[4]

[1]Psychology, School of Social Sciences, Nanyang Technological University, Singapore
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3]CLSP and HLT-COE, Johns Hopkins University, USA
[4]Department of Microelectronics, Delft University of Technology, Netherlands

suzy.styles@ntu.edu.sg, victoriachua@ntu.edu.sg, hexin.liu@ntu.edu.sg, lgarci27@jhu.edu

Abstract

Language development experts need tools that can automatically identify languages from fluent, conversational speech and provide reliable estimates of usage rates at the level of an individual recording. However, LID systems are typically evaluated on metrics such as equal error rate and balanced accuracy, applied at the level of an entire speech corpus. These overview metrics do not provide information about model performance at the level of individual speakers, recordings, or units of speech with different linguistic characteristics. Overview statistics may mask systematic errors in model performance for some subsets of the data, and consequently, have worse performance on data derived from some subsets of human speakers, creating a kind of algorithmic bias. Here, we investigate how well a number of LID systems perform on individual recordings and speech units with different linguistic properties in the MERLIon CCS Challenge featuring accented code-switched child-directed speech.

Index Terms: code-switching, child-directed speech, language identification, language diarization

1. Introduction

Linguistic variation is a fundamental characteristic of natural language. There is an increasing understanding that speech recognition and language identification (LID) systems often underperform for groups of people who speak non-standard varieties of a language, relative to speakers of the 'Standard' for that language. For example, larger error rates are observed for speakers of African American English versus speakers of Californian English [1] and for L2 English speakers who also speak a tonal language [2]. Systematic biases have also been found for individuals speaking non-Standard varieties of Estonian [3] and Arabic [4]. These findings align with the growing awareness that biases in automated classification accuracy present both technical [5], and ethical [6] challenges.

Existing LID systems have shown promising results on general LID tasks. Conventional methods such as i-vector and x-vector comprise a language encoder and a back-end classifier [7, 8]. Recent end-to-end LID systems integrate these two modules in a deep neural network and achieved high performance through the use of novel architectures, large data, and self-supervised learning speech representations [9, 10, 11, 12].

When evaluating LID systems, system-wide metrics such as equal error rate (EER) and balanced accuracy (BAcc) are often adopted. These metrics are useful in describing the overall accuracy of a LID system (i.e., how well did the system perform in terms of identifying language spoken across an entire corpus?). In some contexts, overview statistics can be used to estimate system-level opportunity costs for different models (for example, the cost of human labour required to supplement automated processing for different datasets). However, relying solely on system-wide metrics can be poorly suited to other use cases.

For many potential users of these systems, the end goal is to use automated metrics from individual recordings to derive secondary insights about individuals and communities. For example, in the field of language acquisition, metrics on how much an individual uses each of their languages in a given context is invaluable for understanding how multilingual skills are developing in the individual, how different combinations of linguistic input in different contexts are involved in the development of linguistic skills, and how different communities coordinate their linguistic interactions when they are with children [13]. Effective language use metrics are also critical in clinical contexts, where a difference in language *use* may be mistaken for a difference in cognitive or linguistic *ability* [14]. At present, creating this kind of metric from human data coding is incredibly labour intensive, meaning that this kind of research is slow, difficult to fund, and creates greater barriers for research in multilingual communities – communities that are already underrepresented in language development research [15]. In this context, developmental researchers have a high need for tools that automate LID and generate estimates of language use at the level of individual recordings. Recent work has shown that LID systems underperform on accented and code-switched speech [3, 9]. As systematic errors in such systems could underserve or disadvantage some speakers relative to others due to their language use patterns, it is important to investigate whether different models perform equally well across diverse speech samples within the corpus itself.

In this paper, we investigate the detailed characteristics of a group of LID systems submitted to the Multilingual Everyday Recordings - Language Identification on Child-Directed Code-Switched Speech (MERLIon CCS) challenge targeted at English-Mandarin code-switched, accented, child-directed spontaneous speech. We examine how these submitted systems perform on individual recordings from speakers in a contact language environment, whose English and Mandarin language varieties differ in terms of vocabulary and pronunciation and are underrepresented in publicly available speech corpora. Moreover, as the Challenge dataset features child-directed speech, the speech register is exaggerated and different from that of adult-directed speech characteristics typically represented in speech corpora.

The aim of this paper is to identify major difficulties that state-of-the-art LID systems present, and bridge the gap in speech research so that future models can serve end-users in related fields. In the following sections, we assess the performance of submitted systems in relation to the following variables: *proportion of Mandarin spoken* (Section 3), *presence of vernacular words* (Section 4), *degree of child-directed speech*

10.21437/Interspeech.2023-1707

(Section 5), *duration of language segments and rate of code-switching* (Section 6), and identify areas in which these systems collectively struggle.

2. Methods

The MERLIon CCS Challenge dataset [16] consists of audio recordings of parent-child shared storybook reading recorded over Zoom. It is divided into Development (for system development) and Evaluation (for test) partitions. Speakers in the dataset are parents to children under the age of 5. For LID purposes, ground-truth timestamps and language labels (English or Mandarin) are provided. The speakers in the dataset use the Singaporean variety of English and Mandarin, which features different pronunciation features from other well-documented varieties of English (e.g., US and UK English) and Mandarin (e.g., Putonghua) respectively. The Challenge dataset used for this investigation is described in more detail in the Challenge description paper [17].

There were two tracks in the challenge: in the closed track, systems were only allowed to train on monolingual English speech curated from Librispeech [18] and the Singapore National Speech Corpus [19], monolingual Mandarin speech curated from Aishell [20], and code-switching data from the LDC SEAME corpus [21]; in the open track, systems were allowed an additional 100 hours of open or proprietary training data. Pre-trained models that were publicly available were also allowed in the open track. The Challenge guidelines are detailed in the evaluation plan [22]. A total of 7 and 6 systems were submitted to the closed and open tracks of MERLIon CCS challenge respectively. As the LID task (English and Mandarin) is binary and the challenge dataset is highly imbalanced, equal error rate was selected as a primary evaluation metric [23].

Based on equal error rate, the top 3 models in the closed track were:

- **Multilingual ASR**: A ASR conformer model fine-tuned on Challenge Development set after removing the decoder [24] (Team Name: Speech@SRIB)
- **Codeswitched conformer**: A conformer model fine-tuned on code-switched utterances in LDC SEAME corpus [25] (SBT)
- **Ensemble 6-model**: An ensemble of 6 models fine-tuned on variable-length speech segments from Challenge Development set [26] (Lingua Lumos)

The top 3 open-track models were:

- **Whisper + Multilingual ASR**: Whisper with no fine-tuning combined with the multilingual ASR from closed track [24] (Speech@SRIB)
- **Fine-tuned wav2vec2.0**: A pre-trained wav2vec2.0 model fine-tuned on Librispeech phonetic attributes and Challenge Development set (UNSW Speech Processing)
- **Ensemble titanet-1**: An ensemble of 7 titanet-1 models fine-tuned on Challenge Development Set, curated segments[1] and LDC SEAME corpus [26] (Lingua Lumos)

In accordance to the MERLIon CCS Challenge, full model descriptions of all submitted systems can be found at the MERLIon CCS Challenge GitHub site[2].

We also include two baseline models for comparison: our **MERLIon CCS baseline** conformer model [17] trained on monolingual datasets of closed track and **Whisper baseline**, i.e., Whisper-large-v2 with no fine-tuning (industry benchmark). The submitted systems are described in more detail in the Challenge description paper [17]. In this paper, we focus on examining differences in performance of these 8 systems on the Challenge Evaluation set.

[1]Curated speech segments are 6-second segments from Mozilla Commonvoice as well as Singaporean English and Mandarin Youtube videos misidentified by ensemble 6-model in closed track.

[2]https://github.com/MERLIon-Challenge/merlion-ccs-2023

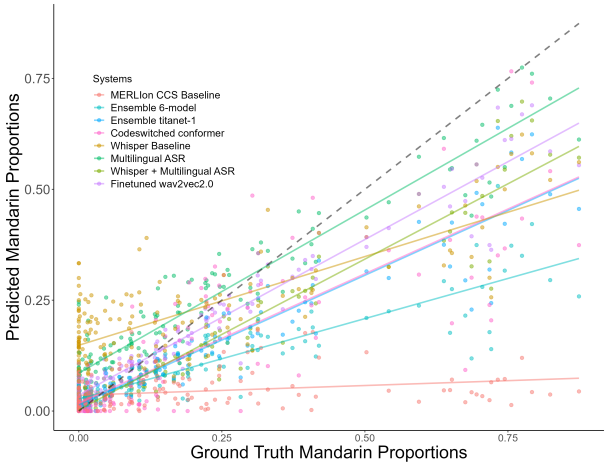

Figure 1: *Proportions of Mandarin spoken per individual recording as estimated by 8 systems. Grey dotted line indicates ground-truth proportions of Mandarin spoken for each individual recording.*

3. Language Proportion Estimations

For end-users working in language acquisition and clinical contexts, an important language identity measure is the proportion of speech in each language in an individual parent-child conversation. This means it is important to understand the accuracy of estimates in individual recordings as well as the estimate across a whole corpus. While the recordings were predominantly in English (Fig. 1), there were recordings in which speakers spoke mostly Mandarin. For each system, we map the proportions of English and Mandarin estimated against the ground-truth proportions for each recording. We present the outputs of all eight systems in Fig. 1. We observe similar trends for other systems submitted to the challenge. All models perform better on files with larger proportion of English than on files with a larger proportion of Mandarin. Performance is particularly poor when proportion of Mandarin is above 50%. Even pre-trained systems with significant amounts of Mandarin speech data collectively underperform for individuals who speak more Mandarin. Notably the system with the least systematic bias between conversations with high proportion of English and Mandarin, is the multilingual ASR without the pre-trained Whisper model.

4. Local Vernacular

The language context of Singapore can be understood as a high contact environment, resulting in lexical items, grammatical structures and accents unique to the region [27, 28, 29]. These speech features are more pronounced in informal spontaneous speech, rarely seen in formal discourse contexts and almost never occur in corpora of read speech. Vernacular features include: clause-final discourse markers that convey mood, attitude, solidarity or emphasis [30, 31], such as ‘lah’; local loan words known by a large number of Singaporeans and used in speech of all languages (e.g., ‘paiseh’ fr. Hokkien to denote embarrassment, ‘makan’ fr. Malay to mean eat).

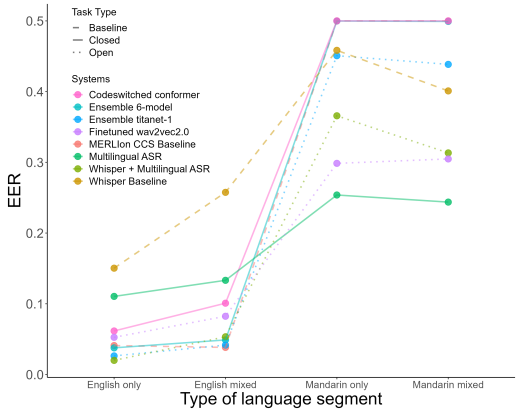

Figure 2: *EERs of each system on subsets of segments categorized by their language and presence of vernacular words. Mixed refers to segments containing English or Mandarin with vernacular words.*

Many lexical items and discourse markers are historic borrowings from regional varieties including Hokkien, Teochew, Cantonese, Hakka, Tamil and Malay. These vocabulary items can appear in both English and Mandarin segments (e.g., "don't have *lah*", "没有 mei3you2 *lah*" both of which mean "I don't have that - DECL. EMPH."). These unique features of local vernacular speech are under-documented and under-represented in publicly available speech corpora, and may present a substantial challenge to LID systems.

Fig. 2 shows the error rates for speech segments categorized by whether they contain only lexical items from Standard Singapore English (i.e., a lexicon largely consistent with Southern British English) and Standard Mandarin, or whether they also contain local vernacular items. English segments with vernacular words tend to be misidentified more often compared to English segments without vernacular words. On the other hand, we observe mixed results for Mandarin segments, with systems tending to perform *better* when local vernacular is present.

In particular, the system with the lowest equal error rate, the Whisper pre-trained model combined with a multilingual conformer ASR model (Whisper + Multilingual ASR), shows a marked improvement in identifying "mixed" Mandarin segments compared to Mandarin-only segments. As some of the vernacular items may have been loan words in other Chinese varieties present in Singapore (e.g., Hokkien 'paiseh' to denote embarrassment) or Mandarin (e.g., discourse markers like 'lah' 啦 and 'lor' 咯). These vernacular words may share similar phonetic attributes with the Mandarin variety represented in accessible Mandarin speech corpora (e.g., Beijing variety of Mandarin in Aishell). By contrast, although English speech is heavily represented in global speech corpora, these unique region-specific vernacular words may stand out as atypical of global English.

5. Child-Directed Speech

When communicating with young children, adults often systematically adopt a different speech style, altering the way they talk and sound. Such a register is often referred to as child-directed speech, or parentese. Child-directed speech is characterized by changes to word choice (simpler words, more repetitions), shorter syntactic structure, and unique acoustic features such as higher pitch, slower speech rate and hyperarticulation of vowel formants and tones [32]. Moreover, many features

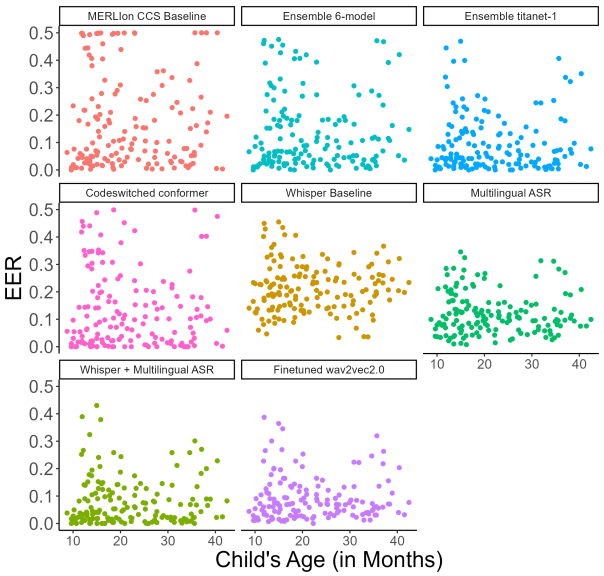

Figure 3: *Scatterplots showing the relationship between equal error rates and age of child per individual recording. Each plot corresponds to a language identification system. All correlations $r < 0.1$.*

of child-directed speech also occur in other low-intelligibility speech contexts including speech produced for foreigners [33], for artificial speech processing systems [34] and when audio clarity is reduced [32, 35]. Recent work has demonstrated that speech recognition and LID systems underperform when speech segments have prosodic or acoustic features that deviate from the features they are typically trained on [2, 3]. Thus, reducing errors on speech containing these features has implications beyond child-directed speech alone.

In the MERLIon CCS dataset [16], 85% of the speech is from parents of children under the age of 5. It has been found that as child age increases, parents employ less child-directed speech [36]. We use age of child at the time of recording as a proxy for the degree of child-directed speech in the recording. As such, we may expect to see a positive relationship between LID performance and age of child.

Contrary to what we expected, the age of child did not systematically affect the performance of LID systems, suggesting that features of child-directed speech present in our corpus do not disrupt model performance selectively for parents addressing children of different ages (Fig. 3).

6. Code-Switching and Segment Lengths

For multilingual speakers, there is increased variability and flexibility in the languages they may choose to use during spontaneous speech. As such, the length of speech segments in each language can vary across speakers in a spontaneous code-switched speech corpora compared to a monolingual speech corpus. Child-directed speech also has shorter sentences than adult-directed speech [37]. As language segment length becomes shorter, it can be increasingly difficult for models to extract sufficient acoustic features for LID, e.g., the average number of unique phonemes reduces exponentially with segment duration [38]. Across the 8 systems, we observe that systems tend to struggle with identifying the language in segments shorter than 2 seconds, as the majority of misidentified segments lie between 0.5 to 1.5 seconds. Most misidentified

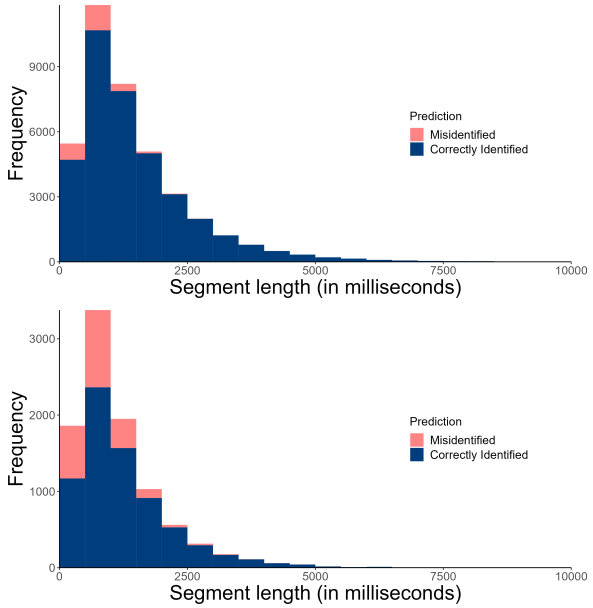

Figure 4: *Frequency distributions of lengths of TOP. all English segments (N = 39245) and BOTTOM. all Mandarin segments (N = 9540) that are correctly and incorrectly identified by system with lowest EER (Whisper + Multilingual ASR).*

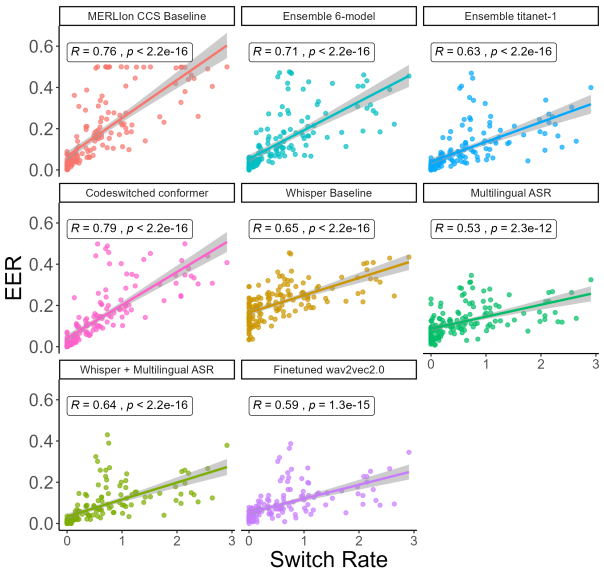

Figure 5: *Scatterplots showing the relationship between equal error rates and rate of code-switching per individual recording. Each plot corresponds to a language identification system.*

English segments range from 0.5 to 1 seconds (Fig. 4), while most misidentified Mandarin segments lie between 0.5 to 2 seconds (Fig. 4). Only the performance of the winning system is shown but the trends are the same for all other systems.

As the speech in our dataset is spontaneous, the rate of language switching also varies across recordings. For instance, two recordings that have equal proportions of English and Mandarin may differ in the number of language switches the speaker may choose to employ, either switching back and forth between languages or only switching to another language once. As the presence of language switches is closely related to length of speech segments in each language (i.e., more switches, shorter speech segments), it is important to investigate if LID systems struggle with speakers who employ more language switches. To that end, we computed the rate of code-switching in each recording. Rate of code-switching per minute is defined as the number of switches to another language between each speech segment divided by total minutes of speech per recording. For instance, a speaker who speaks in English, then Mandarin, back to English again, in one minute of speech, has a rate of 2 language switches per minute. As expected, we observe that LID systems make significantly more errors as the rate of code-switching increases (Fig. 5).

In general, all systems struggled with LID in short segments of speech, which was compounded by high switch rates. In addition, performance was poorer overall for predicting Mandarin segments, especially when segment length was short.

7. Conclusion and Future Directions

Overall, the results of the present study reveal systematic biases in the LID systems submitted to the MERLIon CCS Challenge. There are two main findings we wish to highlight in the present work. First, our findings provide foundational evidence that spoken corpora used in the training of large pre-trained systems under-represent the range of spontaneous speech from certain groups of speakers, specifically; speakers who speak less English, switch between languages more frequently, or use more vernacular words in their English. By contrast, LID systems are unaffected by features of child-directed speech produced for children of different ages. Second, we observe that the presence of vernacular words impacts LID accuracy of English and Mandarin differently. When English segments are mixed with vernacular words, systems tend to struggle with them, while the effect is not so pronounced for Mandarin LID. These words are underrepresented in most English speech corpora, contributing to its poorer performance. Conversely, region-specific words share more acoustic commonalities with Mandarin. In a contact language environment like Singapore, we recommend language models that are targeting code-switching not only train on monolingual resources but target common lexical items that occur between language pairs or multiple language groups.

Models that have been pre-trained on larger datasets do show performance advantages across the board, even though their pre-training datasets are not representative of local speech varieties. However, end users may wish to reduce biases by selecting a model with lower overall performance but less discrepancy in error rate across languages, at the level of individual recordings. Taken together, we demonstrate that additional investigations of multiple metrics alongside system-wide EER across the corpus could clarify how well LID systems are performing for different groups of speakers in a corpus. We recommend the adoption of additional metrics during system evaluation to ascertain the individual variation in error rates for automatic speech processing systems.

8. Acknowledgements

This work was partially supported by the National Research Foundation, Singapore, under the Science of Learning programme (NRF2016-SOL002-011), the Centre for Research and Development in Learning (CRADLE) at Nanyang Technological University (JHU IO 90071537), and the US National Science Foundation via CCRI Award #2120435.

9. References

[1] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[2] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, "Training and typological bias in ASR performance for world Englishes," in *Proc. Interspeech*, 2022, pp. 1273–1277.

[3] K. Kukk and T. Alumäe, "Improving Language Identification of Accented Speech," in *Proc. Interspeech*, 2022, pp. 1288–1292.

[4] C. Sabty, I. Mesabah, Ö. Çetinoğlu, and S. Abdennadher, "Language identification of intra-word code-switching for arabic–english," *Array*, vol. 12, p. 100104, 2021.

[5] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[6] A. Birhane, "Algorithmic injustice: a relational ethics approach," *Patterns*, vol. 2, no. 2, p. 100205, 2021.

[7] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. Interspeech.*, 2011.

[8] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018, pp. 105–111.

[9] H. Liu, L. P. G. Perera, A. W. H. Khong, S. J. Styles, and S. Khudanpur, "PHO-LID: A unified model incorporating acoustic-phonetic and phonotactic information for language identification," in *Proc. Interspeech*, 2022, pp. 2233–2237.

[10] H. Liu, L. P. G. Perera, A. W. H. Khong, E. S. Chng, S. J. Styles, and S. Khudanpur, "Efficient self-supervised learning representations for spoken language identification," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1296–1307, 2022.

[11] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, 2021, pp. 1509–1513.

[12] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[13] C. Scaff, M. Casillas, J. Stieglitz, and A. Cristia, "Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions," 2022.

[14] M. Uljarević, N. Katsos, K. Hudry, and J. L. Gibson, "Practitioner review: Multilingualism and neurodevelopmental disorders–an overview of recent research and discussion of clinical implications," *Journal of Child Psychology and Psychiatry*, vol. 57, no. 11, pp. 1205–1217, 2016.

[15] E. Kidd and R. Garcia, "How diverse is child language acquisition research?" *First Language*, vol. 42, no. 6, pp. 703–735, 2022.

[16] Y. H. V. Chua, L. P. Garcia Perera, S. Khudanpur, A. W. H. Khong, J. Dauwels, F. T. Woon, and S. J. Styles, "Development and Evaluation data for Multilingual Everyday Recordings - Language Identification on Code-Switched Child-Directed Speech (MERLIon CCS) Challenge," 2023, DR-NTU (Data), V1 https://doi.org/10.21979/N9/ANXS8Z.

[17] Y. H. V. Chua, H. Liu, L. P. Garcia Perera, F. T. Woon, J. Wong, X. Zhang, S. Khudanpur, A. W. H. Khong, J. Dauwels, and S. J. Styles, "MERLIon CCS Challenge: A English-Mandarin code-switching child-directed speech corpus for language identification and diarization," 2023. [Online]. Available: http://arxiv.org/abs/2305.18881

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[19] J. X. Koh, A. Mislan, K. Khoo, B. Ang, W. Ang, C. Ng, and Y. Tan, "Building the singapore english national speech corpus," *Malay*, vol. 20, no. 25.0, pp. 19–3, 2019.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, 2017, pp. 1–5.

[21] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "SEAME: a Mandarin-English code-switching speech corpus in South-East Asia," in *Proc. Interspeech*, 2010, pp. 1986–1989.

[22] L. P. Garcia Perera, Y. H. V. Chua, H. Liu, F. T. Woon, A. W. H. Khong, J. Dauwels, and S. J. Styles, "MERLIon CCS Challenge Evaluation Plan Version 1.2," 2023. [Online]. Available: https://arxiv.org/abs/2305.19493

[23] R. Duroselle, D. Jouvet, and I. Illina, "Metric learning loss functions to reduce domain mismatch in the x-vector space for language recognition," in *INTERSPEECH 2020*, 2020.

[24] K. Praveen, B. Radhakrishnan, K. M. Sabu, A. Pandey, and M. A. B. Shaik, "Language identification networks for multilingual everyday recordings," 2023, Accepted for Interspeech 2023.

[25] S. S. Li, C. Xiao, T. Li, and B. Odoom, "Simple yet effective code-switching language identification with multitask pre-training and transfer learning," 2023. [Online]. Available: http://arxiv.org/abs/2305.19759

[26] S. K. Gupta, S. Hiray, and P. Kukde, "Spoken language identification system for english-mandarin code-switching child-directed speech," 2023. [Online]. Available: https://arxiv.org/abs/2306.00736

[27] L. Lee, "The Tonal System of Singapore Mandarin," in *Proc. of NACCL*, 2010, pp. 345–362.

[28] D. Deterding, *Singapore English*. Edinburgh University Press, 2007.

[29] G. L. Sui, *Spiaking Singlish: A companion to how Singaporeans communicate*. Marshall Cavendish International Asia Pte Ltd, 2017.

[30] J. R. Leimgruber, J. J. Lim, W. D. W. Gonzales, and M. Hiramoto, "Ethnic and gender variation in the use of colloquial singapore english discourse particles," *English Language & Linguistics*, vol. 25, no. 3, pp. 601–620, 2021.

[31] D. Smakman and S. Wagenaar, "Discourse particles in colloquial singapore english," *World Englishes*, vol. 32, no. 3, pp. 308–324, 2013.

[32] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's new, pussycat? on talking to babies and animals," *Science (New York, N.Y.)*, vol. 296, p. 1435, 06 2002.

[33] M. Uther, M. Knoll, and D. Burnham, "Do you speak e-ng-l-i-sh? a comparison of foreigner- and infant-directed speech," *Speech Communication*, vol. 49, no. 1, pp. 2–7, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639306001385

[34] S. Perez-Kriz, G. Anderson, and J. Trafton, "Robot-directed speech: Using language to assess first-time users' conceptualizations of a robot," 03 2010, pp. 267–274.

[35] V. Hazan, M. Uther, and S. Granlund, "How does foreigner-directed speech differ from other forms of listener-directed clear speaking styles?" 08 2015.

[36] P. Foulkes, G. J. Docherty, and D. Watt, "Phonological variation in child-directed speech," *Language*, vol. 81, no. 1, pp. 177–206, 2005.

[37] G. Genovese, M. Spinelli, L. J. R. Lauro, T. Aureli, G. Castelletti, and M. Fasolo, "Infant-directed speech as a simplified but not simple register: a longitudinal study of lexical and syntactic features," *Journal of child language*, vol. 47, no. 1, pp. 22–44, 2020.

[38] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2018.