MERLIon CCS Challenge

A English-Mandarin code-switching child-directed speech corpus for language identification and diarization

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

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

MERLIon CCS Challenge: A English-Mandarin code-switching child-directed speech corpus for language identification and diarization

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

[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

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

Abstract

To enhance the reliability and robustness of language identification (LID) and language diarization (LD) systems for heterogeneous populations and scenarios, there is a need for speech processing models to be trained on datasets that feature diverse language registers and speech patterns. We present the MERLIon CCS challenge, featuring a first-of-its-kind Zoom video call dataset of parent-child shared book reading, of over 30 hours with over 300 recordings, annotated by multilingual transcribers using a high-fidelity linguistic transcription protocol. The audio corpus features spontaneous and in-the-wild English-Mandarin code-switching, child-directed speech in non-standard accents with diverse language-mixing patterns recorded in a variety of home environments. This report describes the corpus, as well as LID and LD results for our baseline and several systems submitted to the MERLIon CCS challenge using the corpus.

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

1. Introduction

In recent years, language identification and diarization have progressed significantly, resulting in models that achieve low error rates across different languages [1, 2]. However, most speech processing technologies have been developed for performance on a narrow range of audio data – adult voices, speaking or reading a single language in a particular variety, recorded in a controlled acoustic environment [3, 4]. However, spontaneous speech in-the-wild is rarely so controlled (see CHiME data [5]). In addition, speaking in more than one language, switching rapidly between languages, using a variety of English other than American or British English, and adopting different speech registers when speaking to different parties, are common for most speakers in the globalized world. Recent work has found that language identification systems struggle to cope with different varieties of English [6, 7, 8, 9], as well as spontaneous in-the-wild speech containing more than one language and extremely short language spans [10, 11, 12]. Hence, there is a need to develop models that can handle spontaneous code-switched speech across different populations of speakers.

While speech technologies have been mainly trained on adult voices, effective technology requires robustness across speech registers (e.g., when the same speaker adopts a different style and pitch for a different audience). For instance, child-directed speech by adults comprises a range of acoustic features that differentiate it from adult-directed speech, such as overall higher fundamental frequency, expanded pitch range, hyper-articulation of vowel formants and tones, slower speech rate and lengthened speech segments. These features present critical challenges for automatic language identification and language diarization. Speech diarization systems when employed

for parent-child conversations often underestimate the number of child vocalizations, child-adult, and adult-adult conversation turns [13]. Moreover, many features of child-directed speech also occur in other low-intelligibility speech contexts including speech produced for foreigners [14], for artificial speech processing systems [15] and when audio clarity is reduced [16, 17]. Hence reducing errors on speech containing these features will have impacts beyond child-directed speech alone. Notwithstanding the above, with the increase in the ubiquity of videocalls, conversations recorded on videocall platforms across a variety of home environments (user devices, connection speeds, and noise types) are understudied in the field of speech processing.

To enhance the inclusiveness, reliability, and robustness of language identification and diarization systems for heterogeneous populations and scenarios, there is a need for systems to be trained on datasets that feature different registers and speech patterns (e.g., US/UK English and Mainland Chinese), and spontaneous codeswitching. To that end, we present a unique audio dataset of Zoom videocalls featuring a Southeast Asian variety of English-Mandarin code-switched child-directed spontaneous speech with diverse language-mixing patterns across the speakers recorded in a variety of home environments, curated for the Multilingual Everyday Recordings - Language Identification on Code-Switched Child-Directed Speech (MERLIon CCS) challenge[1]. In this report, we describe the MERLIon CCS challenge, with a detailed description of the dataset, baseline experiments and the results of systems submitted to the challenge.

2. Dataset Description

The MERLIon CCS Challenge dataset [18] results from the collaboration between researchers investigating children's language development in multilingual scenarios and speech engineers working on spontaneous and code-switched speech.

2.1. Data Collection

The audio data collection was part of the Talk Together Study, a large developmental study conducted in Singapore [19]. The study is approved by the Nanyang Technological Institutional IRB Board (IRB-2018-10-001). In the study, parents were required to narrate an onscreen wordless picture book [20] to their children over the Zoom video-conferencing software. All participating parents gave consent for the video calls (including their child's voice) to be recorded and audio to be released. Each recorded session lasted between 3 and 35 minutes, of which 2 to 25 minutes have been manually annotated by a team of multilingual transcribers [21]. The MERLIon CCS Chal-

---

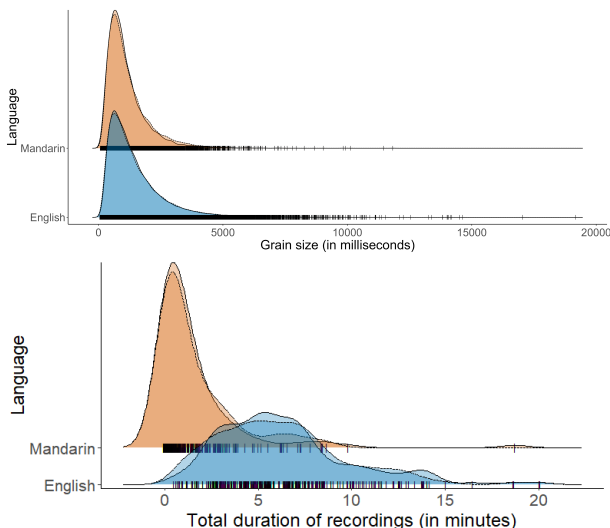[1]https://github.com/MERLIon-Challenge/merlion-ccs-2023

Figure 1: *Density plots comparing English and Mandarin audio recordings in Development (Solid line) and Evaluation (Dashed line) sets, with strip plots indicating distribution. Top. Grain size over Development (N = 9983 Mandarin grains; 40287 English grains) and Evaluation (N = 9766 Mandarin grains; 39473 English grains) sets. Bottom. Total duration of each language in recordings over Development ($N = 151$) and Evaluation ($N = 154$) sets.*

lenge dataset includes only audio recordings.

The entire dataset (including development and evaluation datasets) contains 305 Zoom audio recordings of 112 parent-child pairs [18]. 103 of the parent-child pairs were recorded at least twice on separate occasions, with a maximum of three recordings for each pair. The dataset includes over 25 hours of English child-directed speech and over 5 hours of Mandarin Chinese child-directed speech. Around 10% of the speech in the dataset is from children. Almost all adult voices in the dataset are female, speaking to children under the age of 5.

Zoom calls were conducted in the homes of participating families on a variety of internet-enabled personal electronic devices including laptops, tablets, and mobile phones.[2] Environmental background noise and voices of other family members are also presented in the recordings.

2.2. Languages and Accents

Adults in our dataset use the Singaporean variety of English [22], which features different pronunciations from Standard US, Standard UK, and other well-documented varieties of English, and the Singaporean variety of Mandarin Chinese, which features different pronunciations from the standard variety of Mandarin (Putonghua), and other well-documented Chinese varieties [23]. The Singaporean varieties also feature some unique vocabulary and grammar [22]. The dataset includes frequent *code-switching* within and between utterances.

Only 61 recordings (20%) feature one language throughout. For parents who used both languages, the proportion of Mandarin spoken overall ranged from 0.85% to 80.7%. The utterances are short (mean of 1.4 seconds for English and 1.1 seconds for Mandarin). A breakdown of the composition of files in the Development and Evaluation sets can be found in Table 2.

---

[2]Although some metadata is available for some recordings, the Zoom platform does not contain verifiable data about user services.

Table 1: *Training datasets for the closed track*

Corpus	Hours	Language
LibriSpeech	100	English
AISHELL	200	Mandarin
NSC	101	English
SEAME	192	English-Mandarin Code-Switching

2.3. Human Annotation

As part of the main study, each audio recording was manually annotated by transcribers in ELAN, using the in-house transcription protocol [21]. As part of this protocol, transcribers were instructed to segment and annotate all utterances (i.e., any sound/voice produced by the speaker via their vocal apparatus). The start and end of an utterance are defined by intonation patterns and pauses [21]. In the transcription protocol, a subdivision of an utterance, due to code-switching to a different language, is known as a "grain". Onsets and offsets of different languages are marked. In addition, each utterance or subdivision is labelled with boundaries for non-linguistic communicative acts including vocal sounds (e.g., humming) and non-vocal sounds (e.g., clapping).

Transcribers were given instructions to place the start-stop boundaries carefully, taking note to include sounds at the edge of words (such as fricatives like /s/ at the end). In the event of a language change, they were asked to include all word boundaries in each language. When overlap between speakers occurs, transcribers are instructed to identify the start and end of each speaker turn to the best of their abilities. To minimize transcription errors, each file in the dataset has been crosschecked by at least one senior member of the transcription team.

3. MERLIon CCS Challenge

The Evaluation dataset is selected to be a representative subset of the Development dataset, where features such as ratio of English to Mandarin per recording are well-matched as shown in Fig. 1 and Table 2. To reduce overfitting to individual parent-child pairs, the pairs in the Evaluation dataset do not appear in the Development dataset. Voices of some research assistants who were initiating the videocall repeat in both Development (4.1%) and Evaluation (4.5%) datasets.

3.1. Open and Closed Datasets for Training

For the MERLIon CCS Challenge, we created two tracks which differ in terms of the volume of data systems can be trained on. We refer to these tracks as open versus closed respectively.

In the closed track, to facilitate replicable research outcomes and for systems to be broadly comparable across research communities, we specified only training on pre-selected partitions of three open-access monolingual speech corpora and one LDC corpus (Table 1)[3]. The motivation behind using a limited amount of monolingual training data is to determine the lower limit of error rates when dealing with complex code-switching speech with a fixed set of monolingual data resources for training. In many contexts, monolingual speech is better represented in existing speech corpora and may be less challenging to collect and annotate. In the open track, all systems were allowed an additional maximum 100 hours of any publicly available or proprietary data. Pretrained models that are publicly available were also allowed in the open track.

---

[3]Access   to   pre-selected   partitions   at

Table 2: *Data distribution in Development and Evaluation sets*

	Dev	Eval
Total duration (hh:mm:ss)	28:36:28	28:47:14
Recordings N	151	154
Parent-child pairs N	56	56
Recordings with one language N	28	33
Total English speech (hh:mm:ss)	16:09:27	15:57:02
Total Mandarin speech (hh:mm:ss)	03:04:05	03:11:04
Total other language speech (hh:mm:ss)	00:00:14	00:00:19
English segments N	40287	39473
Mandarin segments N	9983	9766
Median length of English segments (ms)	1125	1120
Median length of Mandarin segments (ms)	900	930
Mean length of English segments (ms)	1443.84	1454.73
Mean length of Mandarin segments (ms)	1106.41	1173.87
Mean Proportion of English / Total	0.85	0.83
Median Proportion of English / Total	0.95	0.94

3.2. Task 1: Language Identification

For language identification, the goal is to detect and label the language spoken automatically in a given audio segment marked by ground-truth timestamps of an audio recording. During development, systems were provided with audio recordings where ground-truth language labels have been annotated with timestamps. Each audio segment had a unique language label of either English or Mandarin spoken by the parent, child or research assistant. During evaluation, audio recordings were provided with timestamps of audio segments, but the language labels of these audio segments were not provided.

As the language identification task is binary, the primary and secondary evaluation metrics were system-level equal error rate (EER) and balanced accuracy, respectively. Only English and Mandarin segments within each recording were scored during evaluation. Overlapping language segments with different language labels were not scored. For instance, for an English speech segment overlapping with another Mandarin speech segment, both speech segments were excluded from scoring. By contrast, English and Mandarin segments that overlap with non-speech segments were scored. In the event of overlapping speech, if the languages of all speakers are the same, the audio segment was scored.

3.3. Task 2: Language Diarization

The goal of language diarization is to detect spans of each language spoken in each audio recording automatically with no pre-established timestamps for speech or language. During development, systems were provided with audio recordings with ground-truth timestamps and language labels. During evaluation, only the audio recordings were provided. To ensure a fair evaluation of conversational code-switching speech, language overlaps of different language labels were evaluated. In the event of overlapping speech, where languages were the same, both speech segments were evaluated. When a speech segment occurs with another speaker's non-speech vocalizations, only the region which contains the evaluated language segment (i.e., English or Mandarin) was considered for evaluation. For operational reasons in the primary study, for each recording, there were some regions where no annotations have been performed, i.e., there may be speech unannotated for timestamps and language labels. Timestamps of these regions in the Development set were provided as well. These regions were excluded from the test audio during evaluation.

https://sites.google.com/view/merlion-ccs-challenge/datasets.

Table 3: *Baseline system results*

Metric	Dev	Eval
Equal Error Rate (EER)	22.1%	21.7%
Balanced Accuracy (BAC)	50.3%	50.9%
Language Diarzation Error Rate	86.6%	84.0%
English LDER	83.9%	80.5%
Mandarin LDER	99.8%	101.2%

The primary and secondary evaluation metrics were system-level language diarization error rate (LDER) and individual language error rates (LER), respectively. The LDER is based on the speaker diarization error rate used in speech diarization system evaluations [24]. The LDER is computed as the sum of:

- Language error – percentage of scored time for which the wrong language tag is assigned to a speech region.
- False alarm speech – percentage of scored time for which a nonspeech region is incorrectly marked as English or Mandarin speech.
- Missed speech – percentage of scored time for which English or Mandarin speech is incorrectly marked as nonspeech.

For each system, the individual LERs for English and Mandarin are computed as the percentage of scored time for which a speech region containing the target language is incorrectly marked as non-speech or another language.

3.4. MERLIon CCS Baseline System

The baseline system is an end-to-end conformer model. The conformer model has shown higher performance in both language identification and speech recognition tasks [12, 25]. The baseline model consists of four conformer encoder layers followed by a statistics pooling layer and three linear layers with ReLU activation in the first two linear layers. All self-attention encoder layers have eight attention heads with input and output dimensions being 512, and the inner layer of the position-wise feed-forward network is of dimensionality 2048. The 39-dimensional mel-frequency cepstral coefficients (MFCCs) features comprising 13-dim MFCCs and their first- and second-order deviations are extracted for each speech signal before being fed into the conformer encoder layers. The statistics pooling layer then generates a 1024-dimensional output which is finally projected by three linear layers, comprising 1024, 512, and 2 output nodes, to the number of target languages.

The baseline system is trained on the pre-selected partitions of three monolingual speech corpora (Table 1). The speech signals in these datasets were segmented into a maximum of 3 s prior to the feature extraction stage. The model was trained for five epochs with batch size 32 and updated with a learning rate that warms up from 0 to 10^{-4} in 5000 steps followed by the cosine annealing decay. For language diarization, an energy-based voice activity detection is performed to identify the silent parts. Each speech signal is partitioned into speech clips after removing silences before performing language identification on these clips. We assume no code-switch exists in each speech clip. The performance of our baseline system is summarized in Table 3.

4. Results and Discussion

The MERLIon CCS Challenge was outlined in the evaluation plan [26] and all submitted system outputs were evaluated via CodaLab [27]. We compare our baseline system language iden-

tification results to 7 systems submitted to the closed track and 6 submitted to the open track (Fig. 2). No systems were submitted to the language diarization task. All submitted models made improvements over the baseline system in terms of EER, BAC or both. The three best performing models in the Open track outperformed all models in the Closed track in terms of EER, of which two of these were adaptations of systems submitted to the closed track.

The first place winner in the Open track achieved an EER of 9.50% and a BAC of 78.9% [28]. Submitted by Speech@SRIB (Samsung R&D Institute India-Bangalore), the system had an ensemble learning method combining a multilingual ASR network with the pre-trained Whisper-large-v2 [29]. The multilingual ASR network is a transformer-based model trained on the monolingual datasets in Table 1 and the Development set with RNN-T loss. This model is adapted to generate the language tag before generating the linguistic tokens. The ASR network is then trained on a subset of the Development set with weighted cross-entropy loss applied to the output layer. The final system combines language posteriors from the multilingual ASR model and Whisper model. In the closed track, the fine-tuned multilingual ASR system alone ranked first, with an EER of 13.9% and 81.7% BAC.

The University of New South Wales (UNSW) Signal Processing team submitted the second placed system in the Open track. They fine-tuned the wav2vec2-large-robust model [30] with phonological attributes from the Librispeech partition (Table 1) extracted by the CMU dictionary. To handle short speech segments, truncated utterances of 3 seconds from the Development set were used for downstream fine-tuning in the language identification phase. The system achieved an EER of 10.6% and 81.3% BAC.

The third place system in the Open track was submitted by Lingua Lumos team from RingCentral [31] and utilized the pre-trained titanet-1 model [32] in the Nvidia NeMo toolkit [33] and handles the problem of varying speech segment lengths and accented speech with an additional curated dataset. Six models based on the TitaNet architecture were trained on varying segment lengths from the closed training dataset with combinations of loss functions to increase model variability, forming an ensemble system. Then, speech segments from Singaporean English and Mandarin YouTube videos and speech of less than 6 seconds from the Mozilla CommonVoice dataset [34] were selected. Speech segments misclassified by this ensemble system formed a curated dataset of 44 hours. Finally, fine-tuning of 7 different titanet-1 models was performed using the LDC SEAME corpus, the Development set, and the curated dataset of misidentified segments. The final ensemble 7-model, achieves an EER of 11.1% and 76.0% BAC. In the Closed track, the ensemble 6-model system alone ranked third, with an EER of 15.6% and 66.0% BAC.

In the Closed track, in addition to the first and third placing systems already described, the second place system pre-trained a bilingual Conformer ASR model on monolingual datasets in Table 1, then fine-tuned on the Development set and segments derived from code-switched utterances in LDC SEAME corpus [35]. This system, submitted by SBT from John Hopkins University, achieved an EER of 15.5% and 70.1% BAC.

The top three systems across both tracks incorporated pre-trained models. The challenge saw improvements to developing language models that handle complex code-switched accented speech. The best performing system achieved an error rate below 10% and improvements of 12.2% over baseline EER and 30.4% over baseline BAC. In line with the rules of the Chal-
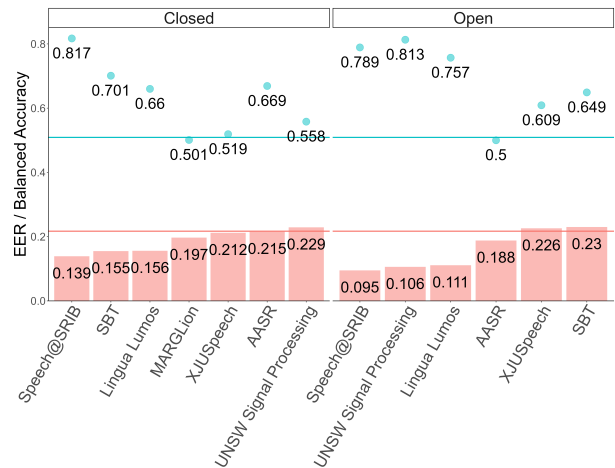

Figure 2: *Equal error rates (bars) and balanced accuracy (dots) of language identification systems submitted to closed and open tracks compared to baseline EER and BAC (lines).*

lenge, full model descriptions of all submitted systems can be found at the MERLIon CCS Challenge GitHub site.

5. Conclusion

Tests performed on the MERLIon CCS evaluation dataset suggest that large pre-trained models can be useful for language identification tasks that feature code-switching speech in different accents and registers when used in combination with models that are trained or curated via a small amount of data that match the speech characteristics of the targeted dataset. Our results highlight the possibility that at least 30 hours of fine-tuning data is beneficial for large pre-trained models to adapt. We further investigate the impact of significantly short speech segments (e.g., less than 3 seconds), presence of local vernacular (e.g., discourse markers like "lah" [22]), and rate of code-switching on language identification performance in a separate paper [36].

It should be acknowledged that language diarization remains a complex task, particularly for recordings across diverse recording environments and noise sources. Such requirements give rise to a robust speech detection system that can accurately generate speech onset and offsets before language identification can occur. This also accounts for why we received sign-ups for our language diarization task but no submissions. As feedback from participants suggests, significant difficulties in identifying the on and offset timings of the speech markers affected language diarization performance. For future research, there is a need to benchmark existing speech detection and diarization tools on our datasets so as to ascertain its errors and impact on subsequent language diarization performance.

6. Acknowledgements

7. References

[1] S. O. Sadjadi, T. Kheyrkhah, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "Performance Analysis of the 2017 NIST Language Recognition Evaluation," in *Proc. Interspeech*, 2018, pp. 1798–1802.

[2] 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.

[3] 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.

[4] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The AMI Meeting Corpus," 2005.

[5] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. Interspeech 2018*, 2018, pp. 1561–1565.

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

[7] 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.

[8] Y. Zhang, Y. Zhang, B. Halpern, T. Patel, and O. Scharenborg, "Mitigating Bias Against Non-Native Accents," in *Proc. Interspeech 2022*, 2022, pp. 3168–3172.

[9] H. Liu, H. Xu, L. P. Garcia, A. W. H. Khong, Y. He, and S. Khudanpur, "Reducing language confusion for code-switching speech recognition with token-level language diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.)*, 2023, pp. 1–5.

[10] H. Liu, L. P. G. Perera, X. Zhang, J. Dauwels, A. W. H. Khong, S. Khudanpur, and S. J. Styles, "End-to-End Language Diarization for Bilingual Code-Switching Speech," in *Proc. Interspeech*, 2021, pp. 1489–1493.

[11] 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.

[12] H. Liu, L. P. G. Perera, A. W. H. Khong, J. Dauwels, S. J. Styles, and S. Khudanpur, "Enhancing Language Identification Using Dual-Mode Model with Knowledge Distillation," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2022, pp. 248–254.

[13] A. Cristia, M. Lavechin, C. Scaff, C. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, and E. Bergelson, "A Thorough Evaluation of the Language Environment Analysis (lena) System," *Behavior Research Methods*, vol. 53, no. 2, pp. 467–486, 2021.

[14] M. Uther, M. Knoll, and D. Burnham, "Do You Speak E-N-G-L-I-S-H? A Comparison of Foreigner- and Infant-Directed Speech," *Speech Communication*, vol. 49, no. 1, pp. 2–7, 2007.

[15] 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.

[16] V. Hazan, M. Uther, and S. Granlund, "How Does Foreigner-Directed Speech Differ From Other Forms of Listener-Directed Clear Speaking Styles?" 08 2015.

[17] 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.

[18] 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.

[19] F. T. Woon, E. C. Yogarrajah, S. Fong, N. S. M. Salleh, S. Sundaray, and S. J. Styles, "Creating a Corpus of Multilingual Parent-child Speech Remotely: Lessons Learned in a Large-scale Onscreen Picturebook Sharing Task," *Frontiers in Psychology*, vol. 12, 2021.

[20] S. J. Styles, "Little Orangutan: What a Scary Storm," 2020, DR-NTU (Data), V1, https://doi.org/10.21979/N9/MJMFXV.

[21] A. Loh *et al.*, "BLIP Lab ELAN Language Annotation (BELA) Transcription Conventions." https://blipntu.github.io/belacon/, 2021.

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

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

[24] N. Ryant, K. W. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Y. Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Proc. Interspeech*, 2019.

[25] Q. Wang, Y. Yu, J. Pelecanos, Y. Huang, and I. L. Moreno, "Attentive Temporal Pooling for Conformer-Based Streaming Language Identification in Long-Form Speech," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 255–262.

[26] 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

[27] A. Pavao, I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu, "Codalab competitions: An open source platform to organize scientific challenges," *Technical report*, 2022. [Online]. Available: https://hal.inria.fr/hal-03629462v1

[28] 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.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[31] 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

[32] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural Model for Speaker Representation with 1D Depth-wise Separable Convolutions and Global Context," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8102–8106.

[33] O. Kuchaiev *et al.*, "Nemo: A Toolkit for Building AI Applications Using Neural Modules," *CoRR*, vol. abs/1909.09577, 2019. [Online]. Available: http://arxiv.org/abs/1909.09577

[34] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[35] 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

[36] S. J. Styles, Y. H. V. Chua, F. T. Woon, H. Liu, L. P. Garcia Perera, S. Khudanpur, A. W. H. Khong, and J. Dauwels, "Investigating model performance in language identification: beyond simple error statistics," 2023. [Online]. Available: http://arxiv.org/abs/2305.18925