CODE-SWITCHING DETECTION VIA LANGUAGE-SELECTIVE DIMENSIONS IN SPEECH FOUNDATION MODELS

Jie Chi, Zakaria Aldeneh, Masha Fedzechkina, Maureen de Seyssel, Natalie Schluter

Apple MLR

arXiv:2610.18726v1 [cs.CL] 13 Oct 2026

ABSTRACT

Code-switching (CS), the alternation between languages within an utterance, is common in multilingual speech but underrepresented in pretraining corpora. Increasing CS coverage requires finding relevant utterances in large speech collections so that they can be included in training data. One way to find such utterances is to transcribe the speech and detect language alternation in the text, but this depends on obtaining accurate multilingual transcriptions. We propose a lightweight method that detects CS directly from speech. It uses frame-level language information in pretrained speech encoders to determine whether both languages are present within an utterance. Specifically, we identify language-selective hidden-state dimensions using monolingual recordings and aggregate their frame-level activations into a CS prediction. With a small CS set used to tune a decision rule, we show that the Whisper-small encoder achieves a mean F1 of 0.798 across 14 language pairs in CS-FLEURS, compared with 0.776 for a CNN and 0.751 for a HuBERT linear baseline trained for monolingual language identification. Our findings suggest that pretrained speech encoders already encode the information needed to identify CS utterances, and they provide a lightweight tool for building more representative multilingual corpora.

***Index Terms*—** Code-switching Detection, Multilingual Speech Models, Language-selective Dimensions

1. INTRODUCTION

Multilingual speech corpora cover many languages but often contain few code-switched (CS) utterances. As a result, models trained on these corpora have limited exposure to mixed-language speech. Recent work shows that training or adapting ASR systems with CS data improves recognition of such speech [1, 2, 3]. A recent study introducing CS-YODAS, a dataset of naturally occurring CS speech mined from the large multilingual YODAS speech corpus, also reports gains in spoken language identification when this speech is added to training [4].

Existing corpora such as SEAME [5], the Miami corpora [6], and the South African soap-opera corpus [7] provide natural CS speech, but the amount available for training is limited by their size, language-pair coverage, and the proportion of recordings that contain a switch. For example, only 8% of utterances in the Miami corpus are code-switched. Even in SEAME, where some interview questions were designed to elicit code-switching, the reported proportion is 52% [8]. Therefore, larger multilingual CS benchmarks have relied mainly on read or synthetic speech [9, 10, 11]. Mining CS utterances from larger speech collections could provide more natural examples across languages and domains.

Mining CS utterances from a large speech collection requires identifying which recordings contain CS. CS-YODAS [4] shows that natural CS speech can be mined from web audio using a pipeline that uses ASR transcripts, a 14B-parameter language model, and bilingual human feedback. Having access to a model that can detect CS directly from speech can be beneficial as a pre-filter before running more expensive parts of the pipeline. We therefore investigate whether CS can be detected directly from speech with a lightweight binary detector.

Prior work has shown that frame-level hidden representations of pretrained speech models encode information about spoken language [12]. In this work, we ask whether this frame-level information can be used to detect whether code-switching has occurred within a given utterance. Our approach proceeds in two steps: we first identify hidden-state dimensions in a pretrained speech encoder whose activations differ between the target languages, using a small set of monolingual recordings; then, at test time, we monitor these same dimensions to predict whether a switch in language occurs within a given utterance. We study the utility of this approach and show that it yields an effective CS detector.

2. RELATED WORK

CS can be detected from speech without first transcribing it. Systems trained on labelled monolingual and CS speech have been evaluated for utterance-level detection on three English–Indian-language pairs, as well as for frame-level language identification [13]. CS language identification has also been studied on English–Mandarin child-directed speech using CNN–GRU models and ASR-assisted pretraining [14]. These systems are developed and evaluated for specific language pairs and speech settings. Applying a trained CS detector to a new pair may require labelled CS examples, which can be difficult to obtain before mining begins.

Language diarization offers another way to mine samples with CS by assigning language labels to temporal regions of speech. An end-to-end approach adapted from speaker diarization trains on CS utterances with language labels at 200 ms intervals, but reports a substantial performance drop from synthetic to recorded speech, partly attributed to differences in segment duration [15]. Another approach fine-tunes speech representations on 14.3 hours of annotated South African soap-opera speech spanning English and four other languages; approximately 4.5 hours of CS speech are in its training set [16]. The DISPLACE challenges evaluate diarization on multilingual conversational speech [17, 18], with one strong system fine-tuning its language representations on in-domain data [19]. These results show that language boundaries can be recovered from speech, while also illustrating the role of annotated CS data and recording conditions. We examine whether likely CS utterances can instead be identified using dimensions selected from monolingual speech and a small labelled CS set for tuning. We test this approach across language pairs and on conversational speech from a second domain.

3. METHOD

In this section, we describe how we identify language-selective dimensions in a pretrained speech model and use this information to detect CS. We first measure how each dimension's activations differ between two languages in monolingual speech. We then test whether the same dimensions are selective across language pairs. Finally, we use their frame-level activations to predict whether an utterance contains CS. We use monolingual speech to select dimensions and estimate language centroids, and code-switched speech to evaluate the detector and tune its decision rule. The corpora, language pairs, and data splits used for each step are described in Section 4

3.1. Language-Selective Dimensions

If language information in a multi-lingual speech model is concentrated in a small number of dimensions, those dimensions may provide a compact signal for detecting both languages within an utterance. We therefore measure how strongly each dimension separates monolingual speech from two languages. Let $\mathbf{h}_t^{(l)} \in \mathbb{R}^d$ denote the hidden state at frame t after Transformer block l. We analyse its individual dimensions, corresponding to coordinates of the encoder residual stream.[1] For a language pair (A, B), we rank dimensions using the absolute standardised difference between their mean activations, Cohen's d:

$$d_i = \frac{|\mu_i^A - \mu_i^B|}{\sqrt{((\sigma_i^A)^2 + (\sigma_i^B)^2)\,/2}}, \tag{1}$$

where μ_i^A and σ_i^A are the mean and standard deviation of dimension i across frames from language A, and likewise for B. A larger d_i indicates a larger difference in mean activation relative to within-language variation. We use the maximum d_i in each layer to choose one layer per encoder, then rank dimensions within that layer for each language pair. We do not compare d_i values across models.

3.2. Cross-Pair Recurrence

A dimension that is selective for one language pair may not be selective for others. We therefore measure how often the same dimensions rank highly across pairs. At the selected layer, we rank dimensions by Cohen's d for each pair and record the top K.

For dimension i, we define its *cross-pair recurrence* as

$$U_i = \frac{1}{|\mathcal{P}|} \sum_{(A,B)\in\mathcal{P}} \mathbb{I}\left[\mathrm{rank}_{A,B}(i) \le K\right], \tag{2}$$

where $\mathcal{P}$ is the set of language pairs and K is the number of dimensions in the top 1%, rounded to the nearest integer. We report $\max_i U_i$, the recurrence of the most frequently selected dimension. Since we take the maximum over all dimensions, the chance level is higher than K/d. We estimate it by Monte Carlo simulation, randomly selecting K dimensions for each pair and computing the maximum recurrence each time.

[1] We also evaluate FFN intermediate activations, taken after the nonlinearity and before the second feed-forward projection; e.g., on Whisper-small, FFN intermediate activations are 3072-dimensional while residual-stream activations are 768-dimensional. The residual stream gives better code-switching detection, so we report residual-stream results throughout.

3.3. Code-Switching Detection

We use the selected dimensions to obtain a language score for each frame. Monolingual speech establishes the typical activations for languages A and B along a language axis; frames closer to each language's centroid on this axis receive scores in opposite directions. We then test whether the utterance contains sufficient local evidence for each language.

For a language pair (A, B), we estimate language centroids $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ from monolingual speech. Given a set of selected dimensions $\mathcal{K}$, we define a unit-length axis between the centroids:

$$\mathbf{v} = \frac{(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)[\mathcal{K}]}{\|(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)[\mathcal{K}]\|_2}, \tag{3}$$

and project their midpoint onto this axis:

$$c_{\mathrm{mid}} = \frac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)[\mathcal{K}] \cdot \mathbf{v}. \tag{4}$$

The score for frame t is

$$s_t = \mathbf{h}_t[\mathcal{K}] \cdot \mathbf{v} - c_{\mathrm{mid}}, \tag{5}$$

where positive scores point towards the A centroid and negative scores towards the B centroid.

We use a 0.5-second sliding window with a 0.1-second stride. Within each window, p_A and p_B are the fractions of frames with positive and negative scores, respectively. We classify an utterance as code-switched if at least n windows satisfy $p_A \ge \theta$ and at least n windows satisfy $p_B \ge \theta$. These need not be consecutive windows. We tune θ and n on a small labelled set, as described in Section 4.

4. EXPERIMENTAL SETUP

4.1. Data

We use three speech corpora. **SEAME** [5] contains approximately 192 hours of spontaneous Mandarin–English speech from 156 speakers in Singapore and Malaysia. We use monolingual training utterances to select encoder layers, estimate English and Mandarin centroids, and train the three baselines. We evaluate CS detection on its two official development sets, *dev_man* (Mandarin-dominant) and *dev_sge* (Singapore English-dominant).

FLEURS [20] contains read speech in 102 languages. We use monolingual training utterances from English and the 14 non-English languages in our CS-FLEURS evaluation. These 15 languages form 105 pairs for the cross-pair recurrence analysis. The same training utterances are used for pair-specific dimension selection, centroid estimation, and baseline training.

CS-FLEURS [11] contains read and synthetic CS speech, covering 113 language pairs and 52 languages in total. We evaluate only on its read-speech portion, using 14 English–X pairs.

For both evaluation corpora, CS utterances are used to tune the detection rule and compute CS detection F1. Neither corpus contributes CS utterances to dimension selection or centroid estimation.

4.2. Pretrained Models

We study three pretrained multilingual speech encoders with different training objectives and language coverage. **Whisper-small** [21] is trained with supervised multitask learning that includes predicting a language token. It covers approximately 100 languages and has a hidden size of $d = 768$. **mHuBERT-147** [22] uses self-supervised

masked prediction on speech from 147 languages, without an explicit language identification objective ($d = 768$). **MMS-1B** [23] uses wav2vec2.0-style contrastive pretraining on 1,406 languages ($d = 1{,}280$).

4.3. Layer and Dimension Selection

For each model, we select one encoder layer using monolingual SEAME training speech, so the layer is not chosen to maximise cross-pair recurrence on FLEURS. We sample 300 English and 300 Mandarin utterances, compute Cohen's d for every dimension at every layer, and select the layer with the highest peak selectivity. Three independent samples of 300 utterances per language selected the same layer.

We keep this layer fixed for the remaining analyses. To measure cross-pair recurrence, we sample 300 training utterances from each of the 15 FLEURS languages and compute Cohen's d for all 105 pairs. We then rank dimensions by their recurrence score (Eq. 2).

For CS detection, we compare two ways of selecting K dimensions. **Recurring dimensions** uses the dimensions with the highest cross-pair recurrence, with the same indices for every target pair. **Pair-specific dimensions** uses the dimensions with the highest Cohen's d for each target pair, selected from its monolingual training speech. We set K=8 for the two models with d=768 and K=13 for MMS-1B with d=1,280. For both strategies, language centroids are estimated from monolingual speech in the corpus used for the corresponding evaluation.

4.4. Baselines

We compare with three classifiers trained to identify the target languages from monolingual speech. Each produces frame-level language predictions at inference, which we aggregate into a CS decision using the same sliding-window rule as our method.

CNN is a four-layer 1D convolutional network operating on 80-dimensional log-Mel features extracted with a 25 ms window and 10 ms shift. Each layer has 256 channels and a kernel size of 5, followed by batch normalization and ReLU. The model is trained with utterance-level language labels using cross-entropy and Adam (lr=10^{-3}) for 30 epochs. **HuBERT linear** is a two-class linear head trained on mean-pooled representations from frozen HuBERT-base (d=768), using the same optimizer and number of epochs. At inference, the head is applied to each frame. **Full-layer probe** uses the same training and inference procedure with frozen Whisper-small layer 8 representations (d=768). It tests whether a classifier trained on the full layer performs better than using a small set of selected dimensions.[2]

For CS-FLEURS, we train each baseline separately for each language pair on monolingual FLEURS training utterances. For SEAME, we train them on monolingual English and Mandarin utterances from its training set. We balance the two languages by subsampling each to $\min(n_A, n_B)$ utterances. On SEAME, this leaves over 22,000 training utterances per language for each baseline, compared with 300 per language for dimension selection.

4.5. Code-Switching Evaluation

We report utterance-level F1 for CS detection. We select the threshold θ and minimum window count n by grid search over $\theta \in \{0.05, 0.10, \ldots, 0.95\}$ and $n \in \{1, \ldots, 5\}$. We report two settings. **Oracle** selects the parameters that maximise F1 on the full evaluation set. **Calibration** tests parameter selection with limited labelled data. For CS-FLEURS, we select (θ, n) on a random 10% of evaluation utterances and evaluate on the remaining 90%. For SEAME, we select (θ, n) on 200 randomly sampled CS training utterances and an equal number of monolingual training utterances, then evaluate on the development sets. We repeat each calibration procedure 200 times and report mean F1 and standard deviation.

[2]We also evaluated a linear head on mHuBERT-147 representations, but it performed worse than the HuBERT-base. Both HuBERT linear and full-layer probe achieve near-perfect monolingual LID accuracy on FLEURS

5. RESULTS

We first examine whether language-selective dimensions recur across pairs and recording domains. If they do, the same dimensions may provide language information for CS detection beyond the pair on which they were identified. We then test this possibility through utterance-level CS detection, per-pair results, and transfer to held-out pairs.

5.1. Language-Selective Dimensions Recur Across Pairs

We first ask whether dimensions selected for one language pair also rank highly for others. If so, the same dimensions may support CS detection across pairs, rather than requiring an unrelated set for each one. Across the 105 FLEURS pairs, the most frequently selected dimension ranks in the top 1% for 55.2% of pairs in Whisper-small, 26.7% in mHuBERT-147, and 16.2% in MMS-1B. The corresponding Monte Carlo chance levels are 5.3% ($\pm$0.7%) for the 768-dimensional models and 5.5% ($\pm$0.7%) for MMS-1B. Recurrence is highest in Whisper-small, although differences in objective, language coverage, and training data prevent us from attributing this result to one aspect of pretraining.

5.2. Some Recurring Dimensions Remain Selective Across Domains

We examine the recurring Whisper-small dimensions more closely. At layer8, dimensions 75, 332, and 704 rank in the top eight for 55.2%, 51.4%, and 51.4% of the 105 pairs, respectively; dimensions 398 and 466 also recur frequently. This raises a further question for CS detection: *do dimensions selected from read speech remain selective in conversational speech?*

For English–Mandarin, per-dimension Cohen's d rankings on FLEURS and SEAME have a weak Spearman correlation of ρ=0.251. Three of the eight most recurring dimensions nevertheless rank in the top eight in both corpora. Dimension 332 ranks first in SEAME, while dimensions 75 and 398 rank in the top seven in both. Other dimensions transfer less consistently: dimension 704 ranks 24th in FLEURS and 26th in SEAME for this pair, whereas dimension 466 falls from 118th to 741st. Thus, some highly recurring dimensions remain selective across domains, but recurrence alone does not guarantee domain stability.

5.3. Selected Dimensions Detect CS in Read and Conversational Speech

Table 1 reports utterance-level CS detection F1. On CS-FLEURS, Whisper-small reaches the highest mean oracle F1, 0.798, using eight pair-specific dimensions. It exceeds the CNN (0.776), HuBERT linear (0.751), and full-layer Whisper-small probe (0.786). mHuBERT-147 and MMS-1B reach 0.756 and 0.762, respectively.

Table 1. Utterance-level CS detection F1. Oracle parameters are selected on the evaluation set; calibrated parameters are selected from limited labelled data. Calibrated results show mean and standard deviation over 200 repetitions.

			CS-FLEURS			SEAME_combined		
				Calibrated			Calibrated	
Method	Model	K	Oracle	Mean	Std	Oracle	Mean	Std
Pair-specific, $K = 1\%$	Whisper-small	8	**0.798**	0.784	.021	0.734	0.733	.008
Pair-specific, $K = 1\%$	mHuBERT-147	8	0.756	0.737	.022	0.732	0.731	.009
Pair-specific, $K = 1\%$	MMS-1B	13	0.762	0.745	.020	**0.758**	**0.755**	.006
Recurring dims. (all pairs)	Whisper-small	8	0.790	0.774	.020	0.736	0.734	.011
Recurring dims. (all pairs)	mHuBERT-147	8	0.697	0.681	.021	0.737	0.735	.009
Recurring dims. (all pairs)	MMS-1B	13	0.735	0.720	.019	0.733	0.732	.009
Recurring dims., leave-X-out	Whisper-small	8	0.797	0.783	.020		N/A	
Recurring dims., leave-both-out	Whisper-small	8	0.795	0.781	.020		N/A	
CNN classifier	—	—	0.776	0.760	.017	0.753	0.740	.006
HuBERT linear classifier	HuBERT	—	0.751	0.733	.019	0.772	0.769	.009
Full-layer probe	Whisper-small	—	0.786	0.768	.019	0.721	0.720	.008

With θ and n selected from 10% of labelled CS-FLEURS utterances, Whisper-small retains mean F1 of 0.784 ± 0.021, compared with 0.798 under oracle selection. This calibration subset is small for most pairs. For example, Italian–English has only 176 CS utterances in total, so a random 10% sample contains about 18 CS examples on average.

For SEAME, Table 1 reports F1 on the combined *dev_man* and *dev_sge* sets. MMS-1B has the highest oracle F1 among the dimension methods (0.758), while HuBERT linear scores highest overall (0.772). The calibrated scores are close to the corresponding oracle scores. In the split-wise results, F1 is always lower on *dev_sge* than on *dev_man*

5.4. CS Detection Varies Across Language Pairs

Table 2 reports oracle F1 for each CS-FLEURS pair. Whisper-small ranges from 0.754 on German–English to 0.870 on Telugu–English and has the highest score on 11 of the 14 pairs. mHuBERT-147 leads on Portuguese–English (0.862) and Spanish–English (0.890), while MMS-1B leads on Telugu–English with the highest single-pair score (0.918). Whisper-small has the highest mean F1, but the strongest model differs by pair.

5.5. Dimension Selection Transfers to Held-Out Pairs

We test whether dimension indices must be selected using the target pair. For each of the 14 CS-FLEURS pairs, we rank dimensions by recurrence over a subset of the 105 FLEURS pairs using Whisper-small. **Leave-X-out** excludes the non-English target language, leaving 91 pairs for dimension selection. **Leave-both-out** excludes both target languages, leaving 78 pairs. In both conditions, we estimate centroids from monolingual speech in the target languages; only the dimension indices are selected without them.

Leave-X-out and leave-both-out achieve mean oracle F1 of 0.797 and 0.795, compared with 0.798 for pair-specific selection. Their calibrated scores are also similar (0.783 and 0.781, compared with 0.784). On these 14 pairs, selecting dimensions without the target pair has little effect on CS detection F1, provided that monolingual speech from the target languages is available for centroid estimation.

Table 2. Utterance-level CS detection F1 for each CS-FLEURS language pair using pair-specific dimensions and oracle parameter selection.

Pair	Whisper-small	mHuBERT-147	MMS-1B
ara-eng	0.832	0.806	0.824
ces-eng	0.793	0.709	0.674
cmn-eng	0.762	0.693	0.728
deu-eng	0.754	0.672	0.677
fra-eng	0.792	0.712	0.727
hin-eng	0.846	0.762	0.835
ita-eng	0.779	0.722	0.735
jpn-eng	0.785	0.743	0.773
kor-eng	0.800	0.693	0.785
por-eng	0.774	**0.862**	0.800
rus-eng	0.827	0.770	0.785
slk-eng	0.781	0.701	0.704
spa-eng	0.775	**0.890**	0.703
tel-eng	0.870	0.855	**0.918**
Mean	**0.798**	0.756	0.762

6. CONCLUSION

We present a lightweight CS detector intended for selecting candidate utterances from speech corpora. It identifies language-selective dimensions using only monolingual speech. With approximately 1% of one Whisper-small encoder layer, it achieves oracle F1 of 0.798 on CS-FLEURS, exceeding the CNN and HuBERT linear baselines. Its F1 on conversational SEAME speech is 0.734. The most frequently selected Whisper-small dimension ranks in the top 1% for 55.2% of language pairs, compared with 5.3% under random selection. Excluding both target languages when selecting dimension indices lowers CS-FLEURS F1 by 0.003, while target-language monolingual speech is still used to estimate centroids. However, these results are based on benchmark datasets; we have not yet tested the detector as part of a mining pipeline on unlabelled, in-the-wild audio. It also requires a small labelled CS set to tune its decision rule. Future work will test candidate selection in that setting and investigate calibration with no CS labels.

7. REFERENCES

[1] Hexin Liu, Xiangyu Zhang, Haoyang Zhang, Leibny Paola Garcia-Perera, Andy W. H. Khong, Eng Siong Chng, and Shinji Watanabe, "Aligning speech to languages to enhance code-switching speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4712–4725, 2025.

[2] Tuan Nguyen and Huy Dat Tran, "Asyncswitch: Asynchronous text-speech adaptation for code-switched asr," *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2025.

[3] Yue Heng Yeo, Yuchen Hu, Shreyas Gopal, Yizhou Peng, Hexin Liu, and Chng Eng Siong, "Improving code-switching speech recognition with tts data augmentation," *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 992–997, 2025.

[4] Brian Yan, Qingzheng Wang, Matthew Wiesner, Anuj Diwan, Olga Iakovenko, Alex Polok, Injy Hamed, Shuichiro Shimizu, Iris Emerman, Thomas Hain, David R. Mortensen, Peter Viechnicki, and Shinji Watanabe, "CS-YODAS: A mined dataset of in-the-wild code-switched speech," in *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, Eds., Palma de Mallorca, Spain, May 2026, pp. 5776–5784, ELRA Language Resource Association.

[5] Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li, "SEAME: A mandarin–english code-switching speech corpus in south-east asia," in *Proceedings of Interspeech*, 2010, pp. 1986–1989.

[6] Margaret Deuchar, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto, and Diana Carter, "Building bilingual corpora," in *Advances in the Study of Bilingualism*, Enlli Môn Thomas and Ineke Mennen, Eds., pp. 93–111. Multilingual Matters, Bristol, 2014.

[7] Ewald van der Westhuizen and Thomas Niesler, "A first South African corpus of multilingual code-switched soap opera speech," in *Proceedings of LREC*, Miyazaki, Japan, May 2018, European Language Resources Association (ELRA).

[8] Jie Chi and Peter Bell, "Analyzing the role of part-of-speech in code-switching: A corpus-based study," in *Findings of ACL: EACL 2024*, Yvette Graham and Matthew Purver, Eds., St. Julian's, Malta, Mar. 2024, pp. 1712–1721, Association for Computational Linguistics.

[9] Anuj Diwan et al., "MUCS 2021: Multilingual and code-switching ASR challenges for low-resource indian languages," in *Proceedings of Interspeech*, 2021, pp. 2446–2450.

[10] Amir Hussein, Dorsa Zeinali, Ondřej Klejch, Matthew Wiesner, Brian Yan, Shammur Chowdhury, Ahmed Ali, Shinji Watanabe, and Sanjeev Khudanpur, "Speech collage: Code-switched audio generation by collaging monolingual corpora," in *Proceedings of ICASSP*, 2024, pp. 12006–12010.

[11] Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, Shinji Watanabe, Chih-Chen Chen, Zhen Wu, Karim Benharrak, Anuj Diwan, Samuele Cornell, Eunjung Yeo, Kwanghee Choi, Carlos Carvalho, and Karen Rosero, "CS-FLEURS: A massively multilingual and code-switched speech dataset," in *Proceedings of Interspeech*, 2025, pp. 743–747.

[12] Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski, "Probing phoneme, language and speaker information in unsupervised speech representations," in *Proceedings of Interspeech*, 2022.

[13] P. Rangan, Santosh Teki, and Hemant Misra, "Exploiting spectral augmentation for code-switched spoken language identification," *arXiv preprint arXiv:2010.07130*, 2020.

[14] Shun-Sheng Li, Cheng Xiao, Tian Li, and Benjamin Odoom, "Simple yet effective code-switching language identification with multitask pre-training and transfer learning," *arXiv preprint arXiv:2305.19759*, 2023.

[15] Jagabandhu Mishra and S. R. Mahadeva Prasanna, "Implicit self-supervised language representation for spoken language diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3393–3407, 2024.

[16] Geoffrey Frost, Emily Morris, Joshua Jansen van Vüren, and Thomas Niesler, "Fine-tuned self-supervised speech representations for language diarization in multilingual code-switched speech," in *Artificial Intelligence Research*, Anban Pillay, Edgar Jembere, and Aurona Gerber, Eds., Cham, 2022, vol. 1734 of *Communications in Computer and Information Science*, pp. 246–259, Springer.

[17] Shikha Baghel, Shreyas Ramoji, Sidharth, Ranjana H, Prachi Singh, Somil Jain, Pratik Roy Chowdhuri, Kaustubh Kulkarni, Swapnil Padhi, Deepu Vijayasenan, and Sriram Ganapathy, "The DISPLACE Challenge 2023 - DIarization of SPeaker and LAnguage in Conversational Environments," in *Interspeech 2023*, 2023, pp. 3562–3566.

[18] Shareef Babu Kalluri et al., "The Second DISPLACE Challenge: DIarization of SPeaker and LAnguage in Conversational Environments," in *Interspeech 2024*, 2024, pp. 1630–1634.

[19] Joonas Kalda, Tanel Alumae, Martin Lebourdais, Hervé Bredin, Séverin Baroudi, and Ricard Marxer, "TalTech-IRIT-LIS Speaker and Language Diarization Systems for DISPLACE 2024," in *Interspeech 2024*, 2024, pp. 1635–1639.

[20] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 798–805.

[21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[22] Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu, "mHuBERT-147: A Compact Multilingual HuBERT Model," in *Interspeech 2024*, 2024, pp. 3939–3943.

[23] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, "Scaling speech technology to 1,000+ languages," *J. Mach. Learn. Res.*, vol. 25, no. 1, Jan. 2024.