DECM: Evaluating Bilingual ASR Performance on a Code-switching/mixing Benchmark

Enes Yavuz Ugan, Ngoc-Quan Pham, *Alexander Waibel
Interactive Systems Lab,
KIT-Karlsruhe Insitut of Technology, Karlsruhe, Germany
*CMU-Carnegie Mellon University, Pittsburgh, USA
{enes.ugan, ngoc.pham}@kit.edu, *alexander.waibel@cmu.edu

Abstract

Automatic Speech Recognition has made significant progress, but challenges persist. Code-switched (CSW) Speech presents one such challenge, involving the mixing of multiple languages by a speaker. Even when multilingual ASR models are trained, each utterance on its own usually remains monolingual. We introduce an evaluation dataset for German-English CSW, with German as the matrix language and English as the embedded language. The dataset comprises spontaneous speech from diverse domains, enabling realistic CSW evaluation in German-English. It includes splits with varying degrees of CSW to facilitate specialized model analysis. As it is difficult to collect CSW data for all language pairs, the provision of such evaluation data, is crucial for developing and analyzing ASR models capable of generalizing across unseen pairs. Detailed data statistics are presented, and state-of-the-art (SOTA) multilingual models are evaluated showing challanges of CSW speech.

Keywords: speech recognition, code-switching, code-mixing, multilingual speech recognition, evaluation-corpus

1. Introduction

Code-switched (CSW) speech has long remained an important linguistic phenomenon Poplack (1980). Under the globalization effects, there exists a growing number of speakers that blend in the words of a second language (mostly English) in their mother tongues, as can be heard in modern German speakers. Even though speech recognition models can now be constructed to deal with multiple languages, their capabilities in dealing with CSW speech are still limited. This is because most training data is homogeneous, and each utterance is monolingual. These datasets, however, do not reflect the way certain English-speaking communities practice English. Take Germany as an example, English has been considered a second language being understood and spoken by the majority of people. As a result the vocabulary of English and German are gradually merged in many areas including business or information technology, where foreign English words are used in the same context with the German counterparts. The original English word "download" can be blended into German utterances such as "Kannst du die Datei downloaden." (Can you download the file.). This practice is often referred to as Denglisch and linguistically named code-switching or code-mixing (CM)[1]

In general, CSW can happen when the spoken utterances switch languages between sentences (inter-sentential CSW), or words in different languages can be mixed in the same sentences (intra-sentential CSW) Poplack (1980). The latter is often considered more challenging for speech recognition models, as the language models and acoustic models can be disrupted by the presence of peculiar foreign words. Different pronunciation, syntactical, and phonetical adaptations can add to the difficulty. CSW speech tends to happen more in natural conversations rather than scripted dialogues being used in recordings and requires participants to be proficient in multiple languages to contribute. This can explain the scarcity of CSW speech data, being able to cover only a limited range of language pairs and not publicly available.

The scope of this work, therefore, is to collect a natural testset for realistic German-English code-switched speech data, in which English is treated as the embedded language in the main language, German. The important feature of our collected dataset is the spontaneity of the data, to serve as a realistic and challenging benchmark for multilingual speech models, which is proven using the temporary high-performance ASR models. To enable a more in-depth analysis of ASR models, we also provide splits of different amounts of CSW in the utterances ($<$2%, 2%$<$9% and $>$9%). The amount of CSW was determined by the percentage of English words in the data. The data will be available by following instructions on our github page[2].

We evaluate several state-of-the-art ASR mod-

[1] On a definition level CSW and CM are viewed differently in different areas of linguistic studies Ezeh et al. (2022).

[2] https://github.com/enesyugan/DECM

LREC-COLING 2024, pages 4468–4475
20-25 May, 2024.

els on our test data revealing the difficulties when dealing with CSW speech.

2. Related Work

Due to the nature of CSW, there is only a very limited amount of such data present in general. Most datasets are concerned with the mixing of English with one of the following other languages: Mandarin Lyu et al. (2010), Cantonese Chan et al. (2005), different Indian languages Diwan et al. (2021), Spanish Deuchar (2008), Japanese Nakayama et al. (2018a), dialectal Arabic Hamed et al. (2022). There are some exception datasets like Amazouz et al. (2017) which cover CSW between Algerian and French. In the case of German-English CSW, there is one benchmark set published Khosravani et al. (2021). However, there are two major drawbacks to this data which is extracted from the Spoken Wikipedia Corpus Baumann et al. (2019). First of all the domain and style of the text is quite different from other speech which can be encountered in real-life scenarios like talks, presentations, lectures, or dialogues. Secondly, the data is read speech which makes it less realistic, than spontaneous speech.

Recently there has been a number of work focusing on CSW ASR which range from old approaches utilizing hybrid models for multilingual ASR such as in Schultz and Waibel (2001), or other more recent work utilizing Sequence-to-Sequence (S2S) models such as in Weller et al. (2022), Huber et al. (2022) or Connectionist Temporal Classification (CTC) based approaches such as in Seki et al. (2018),Song et al. (2022), Yan et al. (2023). Research in Liu et al. (2023), Shan et al. (2019), Li et al. (2019) incorporates some language information during training to improve the model. Whereas another line of work focuses on data augmentation, Liu and Cao (2021), Ugan et al. (2023), Shen and Guo (2022), Du et al. (2021),Nakayama et al. (2018b), Nakayama et al. (2019).

3. Code-switched Dataset

For our German-English CSW evaluation dataset, we collected 3.38 hours of YouTube videos from different topics that we suspected to have CSW present in them. The videos contained topics such as Denglisch itself, Finance, Computer Science, and Gaming.

In order to accelerate the transcript generation process, we broke down the process into two phases. First, the ASROT website [3] is used to generate initial segmentation and transcripts as well as make a first-pass correction across the transcripts. Such correction is done on true-cased words with the presence of punctuation. Additionally, English words as well as morphologically adapted English words are also annotated at the word level.

The second correction pass is done manually by annotators who are German native speakers being highly proficient in the English language. The ability to speak German as a mother tongue is important due to the German-specific ways of adapting English words. The transcription of the English words depends on the grammatical context. The words are treated as if they were originally German, which results in unseen English morphological forms. This transform can be rather intuitive, such as verbs being used in their past form are written in German rules rather than English, such as "gecancelt" instead of "canceled".

Another complication is how the German language has a high degree of freedom in making compound nouns. German and English words can be concatenated together as a single word, and in such cases, they can be annotated with a hyphen such as "Technik-Review". Writing them with hyphens allows for more flexible treatment in post-processing scripts if necessary. To enable a more fine-grained analysis of the CSW performance of ASR models we decided to split the data into three categories based on the amount of English words occurring in a transcript.

- low-CSW: 0.5-2.0%
- mid-CSW: 2-9%
- high-CSW: $>$9%

These numbers also correlate with the commonly used metrics such as Switching-Point Fraction (SPF) Pratapa et al. (2018) and Code-Mixing Index (CMI) Gambäck and Das (2014).

The total number of words is 44147 with 1964 words being present in the English dictionary. The number of words annotated as English amounted to 3348 which also includes abbreviations and morphologically adapted English words, which we refer to as Denglisch in the rest of this paper. Out of the 862 CSW utterances, 30 utterances start with an English word and 832 start with a German word. Our data contains 1353 transitions in which an English word is followed by a German word and 1339 transitions in the reverse direction. The maximum of transitions from English to German and vice versa in an utterance is Nine.

The different CSW-splits with their detailed information are given in Table 1. We can appreciate that each split contains speech of roughly an hour, with the high-CSW split having the most data. We also report the commonly used metrics of Switching-Point Fraction (SPF), and Code-Mixing Index (CMI).

[3] transcript-corrector.dataforlearningmachines.com

Split	# utts	duration	En-ratio	SPF	CMI
low-CSW	401	50.68 min	<2 %	0.02	0.013
mid-CSW	553	75.59 min	2-9 %	0.07	0.040
high-CSW	602	78.85 min	>9 %	0.16	0.113

Table 1: Statistics of the collected German-English CM dataset. Number of utterances, duration, ratio of English/Denglisch words in the data, Switching-Point Fraction, Code-Mixing Index

More detailed information about topics in the data can be found in the Appendix A.1 Table 7.

In Table 2 we show an example sentence for each split of the dataset. Transcipts in $\langle eng \ldots \rangle$ marking are Denglisch. The number 5 in the high-CSW example was said in English, and as such is also marked as Denglisch.

Split	example
low-CSW	Ja, das mit den <eng Bots> glaube ich nicht. Das hat mir noch nie einer zeigen können, dass diese <eng Bots> echt sind.
mid-CSW	Und das ist ein <eng Language Model>. Es schlägt schlägt mir im Endeffekt nur das nächste Wort vor.
high-CSW	Sie haben quasi die <eng Group-Stage> absolut <eng gespeedrunt> 2 schnelle 2 Nulls rausgeholt , dann ein 3 Null im <eng best of 5>.

Table 2: Example utterances from the dataset.

4. Recognition benchmarks

In this section, we demonstrate the challenges posed in this dataset using various speech recognizers, ranging from the models trained using publicly available data such as Commonvoice Ardila et al. (2019) or Librispeech Pratap et al. (2020) which are mostly limited to being controlled read speech to large scale models such as Whisper (Radford et al., 2022) and MMS (Pratap et al., 2023).

4.1. ASR Models

Here we divide the recognizers into two main categories: the CTC-based models (Graves et al., 2006) that relax the conditional dependencies between output tokens, and the encoder-decoder based models (Pham et al., 2019), with the intention to observe if the conditional language model nature of the latter struggles with the code-switched inputs.

For the CTC Model, we choose the massively multilingual (MMS) model (Pratap et al., 2023) which is a Transformer-based Wav2vec (Baevski et al., 2020) model fine-tuned on a large scale of data consisting of more than 1000 languages. The German output layer was selected for inference.

For the latter, we consider two different options: the open Whisper model (Radford et al., 2022) (with the Large configuration) being trained on a very large amount of data, especially with the sources potentially containing code-switched data. We evaluate this model with German decoding strategy, as well (WhisperDe).

For comparison, we also trained a model using publicly available resources for German, Italian, Portuguese, Dutch, Spanish, French and English using the data from Commonvoice (CV) (Ardila et al., 2019), Europarl (Koehn, 2005), Multilingual Librispeech (MLS) (Pratap et al., 2020), TedX (Salesky et al., 2021), in-house Lecture data. This model contains a wav2vec2 encoder and a MBART50 decoder, dubbed as **WMB**, which were shown to outperform models trained from scratch (Pham et al., 2022).

Training of the this model is done with PyTorch (Paszke et al., 2017). The batch size is set at ca. 2M samples (ca. 21 minutes of audio) and we used a warming-up schedule which increases the learning rate to 0.001 for the first 4000 steps, and then linearly decreases the learning rate over the next $100K$ updates. The model is language agnostic, during training there is no language cue given to the model. We measure the Word-Error-Rate (WER)s by comparing the outputs with the labels, with all punctuations being stripped off, and words are lower cased.

4.2. Comparison of CSW and monolingual testsets

First of all, we measure the WERs on standard and homogeneous testsets from CV and MLS. Both of the WMB and MMS models fall into the usable range of 10% error rate, with the WMB model exceling in CommonVoice but performs a bit worse in MLS.

Model	CV	MLS
MMS	13.07	8.75
WMB	1.68	10.16

Table 3: WER in percent on CV and MLS testsets

Compared to the standard testsets, our created dataset proves to be a challenge for the current ASR models, as demonstrated in Table 4. Here the lowest error rate achieved is 16.61% using the Whisper model, while the MMS model struggles at 28.49%. It is worth noting that even with the assistance of many times more training data, Whisper only surpasses our model by a relative margin of 12.9%.

There are two main challenges in this testset, causing the high error rates: the recording condition with a high variety of recording qualities in spontaneous settings, and the interference of the code-mixed words in the context.

The result suggests that the MMS model, a CTC based model, is highly overfit on read speech. Despite being better than our WMB model,an encoder-decoder model, in Multilingual Librispeech, MMS really falls behind in this setting with error rates reaching nearly 35% when the amount of code-switched elements is high. Does this suggest that having a conditional language model in the architecture is more important when the input is noisy and varied? This seems to be the case, when the performance of the encoder-decoder models are noticeably better, and the distance between our WMB and Whisper models are proportional in each CSW category.

By verifying the audio quality, it is recognizable that the audio in low-CSW is consisting of multiple speakers or dialogue, the data in mid-CSW is mainly monologues which might explain the unexpected performances of the models in the mid-CSW section, with the expectation of having higher error rates than the low-CSW. Further insights will be given in the following experiments section 4.3.

Model	low-CSW	mid-CSW	high-CSW	all-CSW
WhisperDe	16.58	11.30	25.03	17.71
Whisper	16.55	11.30	22.10	16.61
MMS	28.47	22.57	34.58	28.49
WMB	17.34	12.75	26.71	19.08

Table 4: WER in percent on our Evaluation set

4.3. CSW analysis

Here, we focus on the distinctive error rates of the German and Denglisch parts within the text, with the intention to identify whether the models struggle with the data in general or if they are able to transcribe the audio in German but struggle to do so in the Denglisch part. In order to obtain the separate WERs for each part, it is necessary to calculate the alignment between the model hypothesis and the target transcripts. Afterward, we count the substitutions, deletions, and insertions lying within the Denglisch part of the text or the German part and divide the values by the respective number of words in the target transcript.

Table 5 shows the WERs respective to the German parts of the text. The values closely align with those in Table 4. Given the prevalence of the matrix language and the predominantly German audio content, it is expected that these error rates reflect the overall WER on the dataset. Another interesting point we can appreciate is that the WhisperDe model with the German decoding prefix has almost the same performance as the Whisper model.

Model	low-CSW	mid-CSW	high-CSW	all-CSW
WhisperDe	16.36	10.74	19.03	15.02
Whisper	16.37	10.74	19.04	15.02
MMS	27.69	20.48	27.67	24.79
WMB	16.66	11.36	20.39	15.81

Table 5: WER in percent on German parts of the data

In Table 6 the error rates on the Denglisch parts of the utterances are depicted. First of all, we can see that the WERs on these parts are pretty high when compared to the German parts. The biggest difference on the overall testset is for the MMS model which has 24.79% WER in Table 5 and 72.78% WER on the Denglisch parts. Even the best-performing Whisper model has an increased WER from 15.02% to 34.68%. This shows that the models clearly struggle more when it comes to transcribing parts of the embedded language in the CSW scenario. Additionally, we can appreciate that the mid-CSW split, which is generally easier for the models to transcribe, has higher WERs on the Denglisch parts. This suggests, that even in generally easier-to-transcribe setups, the models still struggle with the embedded language. It is noteworthy that the WER disparity between the low-CSW and high-CSW segmentation, initially registering a modest 2.67% for WhisperDe and Whisper (Table 5), has now expanded to 29.49% and 10.25%, respectively.

Model	low-CSW	mid-CSW	high-CSW	all-CSW
WhisperDe	27.27	24.58	56.76	49.49
Whisper	24.43	24.75	37.67	34.68
MMS	69.32	78.43	71.72	72.78
WMB	44.89	47.83	60.47	57.41

Table 6: WER in percent on Denglisch parts of the data

An example prediction of the models is shown in the Appendix A.2 Table 8.

4.4. Effect of Model architectures

Language-aware models often have an edge in performance compared to language-agnostic counterparts, but do the former struggle with code-switching? For this question, the values in section 4.3 show that the WhisperDe model with a language specific decoding strategy has a similar performance to its language agnostic counterpart Whisper, however, it has a worse performance on the transcription of the embedded language. The WER for WhisperDe is worse than Whisper by 14.81%. These numbers suggest that keeping the model and the decoding language agnostic or letting the model implicitly determine the language

of the speech, yields better transcriptions than predetermining the matrix language.

On the other hand, we might also wonder if using a language-modeling-style approach in the encoder-decoder models is problematic for CSW since the model is less likely to output unfamiliar words given the context. The CTC model, in that case, might arguably be the better choice.

For this question, the values in our tables in section 4.3 show that the encoder-decoder-based models outperform the CTC model (MMS) with respect to WER in all scenarios. However, as the MMS model has fewer parameters as well as no explicit or implicit language model, such as the decoder in the other models, its performance even on the monolingual German parts is already pretty low. Although it performed quite well on read speech Table 3, this suggests that this model needs further training on more diverse data and possibly an external language model for better comparison of its actual CSW capabilities.

5. Conclusion

In this work, we present an evaluation dataset for German-English CSW, with German as the matrix and English as the embedded language. We provide word-level annotation of English words and describe detailed statistics of the data. Additionally, the data is annotated in three splits for more detailed analysis. We evaluate the data on SOTA models and show that they have significant problems with transcribing the embedded language (Table 6) when compared to the matrix language Table 5. Language specific decoding strategies hurt the performance on parts of the embedded language, which suggests the use and investigation of language agnostic models will yield better performing models on the task of CSW speech recognition. The data will be made available by providing download scripts of utilized videos, along with their corresponding segmentation and transcripts[4].

6. Ethics

The source of our data comes from videos uploaded on the Youtube platform, supported by Google LLC corporation. The published content only contains the segmentations and transcripts which account for insufficient original data in the videos, and the action of taking down videos, if necessary, is going to be handled by Google.

[4]https://github.com/enesyugan/DECM

7. Acknowledgement

The project on which this report is based was funded by the Federal Ministry of Education and Research (BMBF) of Germany under the numbers 01EF1803B (RELATER). Part of this work was supported by funding from the pilot program Core- Informatics of the Helmholtz Association (HGF).

8. Bibliographical References

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2017. Addressing code-switching in french/algerian arabic speech. In *Interspeech 2017*, pages 62–66.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Timo Baumann, Arne Köhn, and Felix Hennig. 2019. The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53:303–329.

Joyce YC Chan, PC Ching, and Tan Lee. 2005. Development of a cantonese-english code-mixing speech corpus. In *Ninth European Conference on Speech Communication and Technology*.

Margaret Deuchar. 2008. The miami corpus: Documentation file. *Bangortalk, bangortalk. org. uk/docs/Miami_doc. pdf*.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, et al. 2021. Multilingual and code-switching asr challenges for low resource indian languages. *arXiv preprint arXiv:2104.00235*.

Chenpeng Du, Hao Li, Yizhou Lu, Lan Wang, and Yanmin Qian. 2021. Data augmentation for end-to-end code-switching speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 194–200. IEEE.

Nnenna Gertrude Ezeh, Ifeoma Ann Umeh, Esther Chikaodi Anyanwu, et al. 2022. Code switching and code mixing in teaching and learning of english as a second language: Building on knowledge. *English Language Teaching*, 15(9):106–106.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th international conference on natural language processing, Goa, India*, pages 1–7.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2022. Investigations on speech recognition systems for low-resource dialectal arabic–english code-switching speech. *Computer Speech & Language*, 72:101278.

Christian Huber, Enes Yavuz Ugan, and Alexander Waibel. 2022. Code-switching without switching: Language agnostic end-to-end speech translation. *arXiv preprint arXiv:2210.01512*.

Abbas Khosravani, Philip N Garner, and Alexandros Lararidis. 2021. An evaluation benchmark for automatic speech recognition of german-english code-switching. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 811–816. IEEE.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Ke Li, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong. 2019. Towards code-switching asr for end-to-end ctc models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6076–6080. IEEE.

Guoyu Liu and Lixin Cao. 2021. Code-switch speech rescoring with monolingual data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6229–6233. IEEE.

Hexin Liu, Haihua Xu, Leibny Paola Garcia, Andy WH Khong, Yi He, and Sanjeev Khudanpur. 2023. Reducing language confusion for code-switching speech recognition with token-level language diarization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li. 2010. An analysis of a mandarin-english code-switching speech corpus: Seame. *Age*, 21:25–8.

Sahoko Nakayama, Takatomo Kano, Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. 2018a. Japanese-english code-switching speech data construction. In *2018 Oriental COCOSDA-International Conference on Speech Database and Assessments*, pages 67–71. IEEE.

Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2018b. Speech chain for semi-supervised learning of japanese-english code-switching asr and tts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 182–189. IEEE.

Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Zero-shot code-switching asr and tts with multilingual machine speech chain. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 964–971. IEEE.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019. Very Deep Self-Attention Networks for End-to-End Speech Recognition. In *Proc. Interspeech 2019*, pages 66–70.

Ngoc-Quan Pham, Alexander Waibel, and Jan Niehues. 2022. Adaptive multilingual speech recognition with pretrained models. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3879–3883. ISCA.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam

Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*.

Tanja Schultz and Alex Waibel. 2001. Experiments on cross-language acoustic modeling. In *INTERSPEECH*, pages 2721–2724.

Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey. 2018. An end-to-end language-tracking speech recognizer for mixed-language speech. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4919–4923. IEEE.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Investigating end-to-end speech recognition for mandarin-english code-switching. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6056–6060. IEEE.

Zhijie Shen and Wu Guo. 2022. An improved deliberation network with text pre-training for code-switching automatic speech recognition. *Proc. Interspeech 2022*, pages 3854–3858.

Tongtong Song, Qiang Xu, Meng Ge, Longbiao Wang, Hao Shi, Yongjie Lv, Yuqin Lin, and Jianwu Dang. 2022. Language-specific characteristic assistance for code-switching speech recognition. *arXiv preprint arXiv:2206.14580*.

Enes Yavuz Ugan, Christian Huber, Juan Hussain, and Alexander Waibel. 2023. Language-agnostic code-switching in sequence-to-sequence speech recognition.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. *arXiv preprint arXiv:2204.05076*.

Brian Yan, Matthew Wiesner, Ondřej Klejch, Preethi Jyothi, and Shinji Watanabe. 2023. Towards zero-shot code-switched speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A. Appendices

A.1. Topic distribution in the dataset

CSW Level	Style (Topic)	SPF	CMI	#utts	duration	#speakers
Low-CSW	Satire (artificial intelligence)	0.015	0.010	102	11.98 min	>3
	Podcast (political)	0.022	0.014	299	38.70 min	2
Mid-CSW	Podcast (mobile-phones)	0.072	0.044	262	35.5 min	1
	Podcast (artificial intelligence)	0.058	0.037	291	40.09 min	2
High-CSW	Educational comedy (Denglisch)	0.112	0.093	35	4.40 min	1
	Documentation (technology)	0.052	0.054	127	16.23 min	>3
	Satire (Denglisch)	0.208	0.155	30	4.63 min	>3
	Commentation (Esports)	0.199	0.129	410	53.60 min	>3

Table 7: Statistics and topics covered in the dataset

A.2. Example Hypothesis

Reference	sie haben quasi die group-stage absolut gespeedrunt 2 schnelle 2 nulls rausgeholt dann ein 3 null im best of 5
WhisperDe	sie haben quasi die groupstage absolut gespeedruns zwei schnelle 2-0 rausgeholt dann ein 3-0 im best of five
Whisper	sie haben quasi die groupstage absolut gespeedruns zwei schnelle 2-0 rausgeholt dann ein 3-0 im best of five
MMS	denn sie haben for se die group stage absolut gesfhet runs zwei schnelle zwei nulls rausgeholt dann in drei null im bester
WMB	sie haben die group stage absolute des feature ones zwei schnelle zwei nulls rausgeholt dann ein drei null im best of five

Table 8: Example utterances from the dataset. With an appropriate processing of the reference, green colored words could be considered correct due to the hyphen usage. Red depicted words are considered wrong.