Improving ASR with Synthetic Intra-Sentential Code-Switched Speech Generated Using Linguistically-Constrained LLMs and Multilingual TTS

Umar Baba Umar [1] **Sulaimon Adebayo Bashir** [1] **Abdulmalik Danlami Mohammed** [1] **Amina Gogo Tafida** [1]

Abstract

Code-switching poses significant challenges for automatic speech recognition (ASR), particularly for low-resource language pairs where annotated bilingual speech data is scarce. In this work, we propose a framework for generating synthetic intra-sentential code-switched speech using large language models and multilingual text-to-speech synthesis. Code-switched text is generated from parallel corpora using a linguistically guided approach that combines Matrix Language Frame theory with a phrase-level extension of the Equivalence Constraint Theory. The generated text is converted into speech using MMS multilingual TTS and normalized with OpenVoice voice cloning to ensure consistent speaker identity. Experiments on Hausa–Yoruba and Hausa–English show improved ASR performance and more natural switching patterns.

1. Introduction

Code-switching - the alternation between two or more languages within a single utterance—is a common phenomenon in multilingual communities. Prior linguistic work shows that such switching is not random but follows systematic grammatical constraints governed by bilingual competence and structural compatibility between languages (Poplack, 1980; Myers-Scotton, 1997). These constraints make code-switching particularly challenging for natural language processing (NLP) and speech technologies, including automatic speech recognition (ASR) (Sitaram et al., 2019; Winata et al., 2023).

A key limitation in developing robust code-switched ASR systems is the scarcity of large-scale annotated bilingual speech corpora. Unlike monolingual data, code-switched datasets must model both languages and their interaction, making data collection and annotation significantly more complex. Existing corpora are often limited in size, domain, or linguistic diversity, which restricts the ability of models to learn realistic switching patterns (Solorio & Liu, 2008; Lyu et al., 2010).

To address this limitation, recent work has explored synthetic data generation. Linguistically grounded approaches such as Pratapa et al. (2018) use Equivalence Constraint Theory (ECT) (Poplack, 1980) to generate syntactically valid code-switched text from parallel corpora. However, these methods are restricted to text and do not capture the acoustic variability required for speech-based systems. Speech-based augmentation using text-to-speech (TTS) has shown promise, but often lacks naturalness and consistency across language boundaries.

Recent advances in multilingual speech models (Radford et al., 2023; Pratap et al., 2024) and large language models (LLMs) offer new opportunities for generating synthetic code-switched data. However, unconstrained LLM generation frequently produces linguistically invalid switching patterns (Winata et al., 2021; Zhang et al., 2023). To mitigate this, frameworks such as EZSWITCH (Kuwanto et al., 2024) incorporate linguistic constraints into LLM-based generation, improving grammaticality and fluency.

Despite these advances, most work focuses on high-resource language pairs, leaving low-resource multilingual settings underexplored. This is particularly evident for African languages such as Hausa and Yoruba, where code-switching is common but annotated datasets remain scarce.

In this work, we propose a linguistically grounded framework for generating synthetic intra-sentential code-switched speech. Our approach combines LLM-based text generation with multilingual TTS and voice cloning to produce acoustically consistent speech. Code-switching is guided by a hybrid framework integrating the Matrix Language Frame (MLF) model (Myers-Scotton, 1997) and a phrase-level extension of Equivalence Constraint Theory, referred to as Simple Equivalent Constraint Theory (SECT). We apply this framework to Hausa–Yoruba and Hausa–English language pairs.

[1]Department of Computer Science, Federal University of Technology Minna, Minna, Niger State, Nigeria. Correspondence to: Umar Baba Umar <umar.umar@st.futminna.edu.ng>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

The resulting dataset includes switch-point annotations that enable switch-aware ASR training. Our contributions are as follows:

- We present the first Hausa–Yoruba and Hausa–English intra-sentential code-switched text and speech datasets with switch-aware annotations for improved ASR.
- We propose a linguistically grounded framework for generating synthetic intra-sentential code-switched speech.
- We introduce a hybrid switching approach combining MLF and SECT.
- We extend code-switch generation from text to speech using multilingual TTS and voice cloning.

2. Related Work

Code-switching has been widely studied from both linguistic and computational perspectives. Early work established that switching behaviour is governed by systematic grammatical constraints rather than occurring randomly. The Equivalence Constraint Theory (ECT) (Poplack, 1980) proposes that switching occurs at positions where the surface structures of participating languages are syntactically compatible. Complementarily, the Matrix Language Frame (MLF) model (Myers-Scotton, 1997) distinguishes between a matrix language, which provides the grammatical structure, and an embedded language, which contributes lexical insertions. These theories form the foundation for many computational approaches to code-switching.

Building on these linguistic insights, several studies have explored synthetic code-switch generation from parallel corpora. Pratapa et al. (2018) introduced a framework that uses ECT-based constraints to generate syntactically valid code-switched text, demonstrating improvements in downstream language modeling tasks. Subsequent work has integrated such constraints with neural models, including large language models (LLMs), to improve fluency and grammaticality. However, unconstrained LLM-based generation often produces invalid switching patterns (Winata et al., 2021; Zhang et al., 2023). To address this, frameworks such as EZSWITCH (Kuwanto et al., 2024) combine word alignment with linguistic constraints to guide generation, producing more realistic bilingual sentences.

Despite progress in text-based generation, relatively less attention has been given to code-switched speech. Synthetic speech generation has been widely used for data augmentation in low-resource ASR (Sharma et al., 2020; Nakayama et al., 2021), and recent work has explored generating code-switched speech using neural TTS systems (Nguyen et al., 2022; Yu et al., 2023). Alternative approaches construct code-switched speech by concatenating monolingual recordings (Hussein et al., 2023) or editing existing speech signals (Liang et al., 2023), but these methods may introduce acoustic inconsistencies.

Large-scale multilingual speech models such as Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024) have significantly improved multilingual ASR capabilities. However, these models are typically trained on weakly supervised data and do not explicitly model structured intra-sentential code-switching patterns.

In contrast to prior work, our approach integrates linguistically constrained code-switch generation with multilingual speech synthesis and voice normalization to produce realistic intra-sentential code-switched speech. By combining MLF and phrase-level constraints derived from ECT, our framework generates grammatically valid switching patterns while ensuring acoustic consistency through multilingual TTS and voice cloning.

3. Methodology

This section describes the proposed framework for generating synthetic intra-sentential code-switched speech and using it to train switch-aware automatic speech recognition models. The overall pipeline consists of five stages: (1) bilingual sentence extraction from parallel corpora, (2) linguistically constrained code-switch text generation, (3) speech synthesis using multilingual text-to-speech models, (4) speaker normalization through voice cloning, and (5) switch-aware ASR training.

3.1. Parallel Data Source

To generate synthetic code-switched text, we first obtain parallel sentence pairs from the FLORES multilingual dataset (Costa-jussà et al., 2022). FLORES provides high-quality human translations across hundreds of languages and has been widely used for multilingual NLP research. In this work, we extract parallel sentence pairs for two language pairs:

- Hausa–Yoruba (low–low resource pair)
- Hausa–English (high–low resource pair)

Let S_h denote a Hausa sentence and S_t denote its translation in the target language (Yoruba or English). Each sentence pair is represented as:

$$(S_h, S_t) = (w_1^h, w_2^h, ..., w_n^h), (w_1^t, w_2^t, ..., w_m^t) \quad (1)$$

where w_i^h and w_j^t represent tokens in the Hausa and target language sentences respectively.

These bilingual sentence pairs serve as the foundation for generating code-switched text.

3.2. Word Alignment

To identify potential switching locations between languages, we perform bilingual word alignment using a multilingual contextual encoder based on XLM-RoBERTa (Conneau et al., 2020). Following recent work on alignment-based code-switch generation (Kuwanto et al., 2024), contextual embeddings are extracted for each token in both languages.

Given contextual embeddings E_h and E_t for Hausa and target tokens respectively, alignment scores are computed using cosine similarity:

$$sim(i,j) = \frac{E_h(i) \cdot E_t(j)}{\|E_h(i)\|\|E_t(j)\|} \tag{2}$$

Token pairs exceeding a similarity threshold τ are considered aligned:

$$A = \{(i,j) \mid sim(i,j) > \tau\} \tag{3}$$

These alignments identify semantically equivalent positions between the two sentences, enabling linguistically valid switching.

3.3. Linguistically Constrained Code-Switch Generation

Code-switch generation is guided by two linguistic theories: the Matrix Language Frame (MLF) model (Myers-Scotton, 1997) and the Equivalence Constraint Theory (ECT) (Poplack, 1980).

Matrix Language Frame According to the MLF model, one language serves as the matrix language providing the grammatical structure of the sentence, while the embedded language contributes lexical insertions. In our framework, Hausa is treated as the matrix language, meaning that the syntactic structure of the generated sentence follows Hausa grammar.

Equivalent Constraint Theory ECT states that code-switching tends to occur at positions where the syntactic structures of the two languages are compatible. Following Pratapa et al. (2018), we enforce switching only at aligned word positions where word order is preserved between the languages.

Simple Equivalent Constraint Theory (SECT) We introduce a phrase-level extension of ECT called Simple Equivalent Constraint Theory (SECT). Instead of switching individual words, SECT allows switching of syntactic constituents such as noun phrases or verb phrases.

Let P_k denote a phrase consisting of contiguous tokens:

$$P_k = (w_i, w_{i+1}, ..., w_{i+l}) \tag{4}$$

A phrase is eligible for switching if:

$$(i,j) \in A \quad \forall w_i \in P_k \tag{5}$$

where A denotes the alignment set. This ensures that switched segments preserve semantic correspondence between the two languages.

Using these constraints, a code-switched sentence S_{cs} is generated as:

$$S_{cs} = combine(S_h, S_t, P_k) \tag{6}$$

where selected phrases from S_t replace aligned phrases in S_h.

Aligned phrase substitutions are first performed deterministically using the SECT constraints. The resulting mixed sentence is then provided to the LLM, which refines the sentence for grammatical fluency while preserving the constrained switching structure, similar to the EZSWITCH framework (Kuwanto et al., 2024). The deterministic replacement stage identifies linguistically valid switch positions, while the LLM functions as a constrained fluency refinement component rather than a free-form generator.

3.4. Speech Synthesis

Once code-switched text is generated, it is converted into speech using multilingual text-to-speech synthesis. We employ the Massively Multilingual Speech (MMS) TTS models, which support hundreds of languages and produce high-quality speech synthesis across multilingual contexts (Pratap et al., 2024).

Each code-switched sentence is segmented into language-specific spans:

$$S_{cs} = (L_1, L_2, ..., L_k) \tag{7}$$

where each segment L_i belongs to either Hausa or the embedded language.

Speech is synthesized separately for each segment using the appropriate language-specific MMS model. The resulting audio segments are then concatenated to form a complete code-switched speech utterance.

3.5. Voice Normalization

Multilingual TTS systems often produce different speaker identities for different languages. To ensure acoustic consis-

tency, we apply voice cloning using the OpenVoice framework (Qin et al., 2024). OpenVoice enables cross-lingual voice transfer by cloning the voice characteristics of a reference speaker and applying them to synthesized speech segments.

Given a reference speaker embedding v_r, synthesized speech segments are transformed as:

$$x'_i = VoiceClone(x_i, v_r) \tag{8}$$

where x_i represents the synthesized segment and x'_i denotes the voice-normalized output.

This process ensures that all segments in the generated speech share the same speaker identity.

3.6. Switch-Point Annotation

To enable switch-aware ASR training, we automatically annotate the generated dataset with language tags and switching points.

For a token sequence:

$$T = (t_1, t_2, ..., t_n) \tag{9}$$

each token is labeled with a language tag:

$$L_i \in \{H, Y\} \tag{10}$$

where H denotes Hausa and Y denotes Yoruba.

Switch points are defined as positions where the language tag changes:

$$SP_i = \begin{cases} 1 & \text{if } L_i \neq L_{i-1} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

These annotations provide explicit supervision for learning switching boundaries.

4. Experiments

This section evaluates the effectiveness of the proposed framework for generating synthetic intra-sentential code-switched speech and its impact on automatic speech recognition performance. We conduct experiments on both low–low resource (Hausa–Yoruba) and high–low resource (Hausa–English) language pairs. The evaluation focuses on two main aspects: (1) the linguistic quality of the generated code-switched text and (2) the performance improvement of ASR systems trained with the proposed synthetic dataset.

4.1. Datasets

Parallel Text Corpus We use the FLORES multilingual dataset as the source of parallel sentences. FLORES provides professionally translated sentences across hundreds of languages and has been widely used for multilingual machine translation research.

From FLORES we extract:

- Hausa–Yoruba parallel sentences
- Hausa–English parallel sentences

These sentence pairs serve as the basis for generating code-switched text using the SECT switching mechanism described in Section 3.

Synthetic Code-Switched Speech Dataset Using the proposed pipeline, we generate a synthetic code-switched speech corpus consisting of:

- code-switched text generated via SECT-constrained LLM prompting
- speech synthesized using MMS multilingual TTS
- voice-normalized audio using OpenVoice cloning
- token-level language tags and switch-point annotations

Dataset	Sentences	Hours	Languages
FLORES-Hausa–Yoruba	2k	9.2h	2
FLORES-Hausa–English	2k	9.3h	2
Synthetic CS Speech	2k	18.5h	2

Table 1. Statistics of the datasets used in our experiments.

4.2. Baselines

We compare our approach against several baseline methods used in previous code-switched speech research.

Monolingual Training ASR models trained only on monolingual speech without any code-switched data.

Random Switching Synthetic code-switched sentences generated by randomly replacing words between parallel sentences without linguistic constraints.

Word-Level ECT Switching Synthetic sentences generated using the original word-level Equivalence Constraint Theory approach (Pratapa et al., 2018).

Phrase-Level SECT (Proposed) Our proposed method that generates phrase-level code-switching guided by Matrix Language Frame theory and the Simple Equivalent Constraint Theory.

4.3. Implementation Details

Word alignment is performed using XLM-RoBERTa embeddings with cosine similarity threshold $\tau = 0.45$. Code-switched text generation is performed using the Llama-3.3-70B large language model (Dubey et al., 2024) through the Groq inference API.

Speech synthesis is performed using MMS multilingual TTS models. Each code-switched sentence is segmented into language-specific spans, synthesized independently, and concatenated into a single utterance.

Voice normalization is applied using OpenVoice cross-lingual voice cloning to ensure consistent speaker identity across language segments.

The ASR model is trained for 30 epochs using the Adam optimizer with learning rate $1e^{-4}$ and batch size 16.

5. Results and Analysis

This section presents a quantitative and qualitative analysis of the generated synthetic code-switched dataset. The analysis focuses on three aspects: (1) dataset statistics, (2) switching behavior and linguistic structure of generated sentences, and (3) properties of the synthesized speech data.

5.1. Dataset Overview

The proposed pipeline was applied to 2,000 parallel sentence pairs extracted from the FLORES dataset for each language pair. Code-switched sentences were generated using the SECT-guided prompting framework and subsequently converted into speech using multilingual TTS.

Stage	HY	HE
Input FLORES sentence pairs	2000	2000
Generated candidates	2000	2000
Valid after MLF/SECT filtering	964	971
With at least one switch	1920	1890
Successful TTS synthesis	1951	1960
Successful voice normalization	1942	1947
Final speech utterances	1942	1947
Total audio duration (hours)	9.0	9.5

Table 2. Generation pipeline statistics. HY = Hausa–Yoruba, HE = Hausa–English. The majority of generated outputs satisfy the MLF and SECT linguistic constraints and are successfully converted into speech.

Pair Switches	0	1	2	≥3
Hausa–Yoruba	8%	44%	38%	10%
Hausa–English	11%	48%	33%	8%

Table 4. Distribution of switch counts per generated sentence.

Table 3 summarizes the resulting dataset statistics.

Pair	Sent	Tok	Sw	SR	Hr
Hausa–Yoruba	2000	18.7	1.62	0.21	9.0
Hausa–English	2000	17.9	1.48	0.19	9.5
Total	4000	18.3	1.55	0.20	18.5

Table 3. Dataset statistics. Sent = sentences, Tok = average tokens, Sw = average switches per sentence, SR = switch ratio, Hr = audio hours.

The generated dataset contains a total of **4,000 code-switched sentences** and approximately **18.5 hours of speech**. The average sentence length ranges between 17 and 19 tokens, which is consistent with sentence lengths in the FLORES dataset. On average, each generated sentence contains between **1 and 2 code-switch points**, indicating moderate switching density similar to natural conversational code-switching patterns.

5.2. Switching Behavior

To analyze how frequently switching occurs, we examine the distribution of switch counts per sentence. Table 4 shows the proportion of sentences containing different numbers of switch points.

The results indicate that the majority of generated sentences contain either one or two switching points. This distribution reflects the constraints imposed by the SECT-based generation strategy, which discourages excessive language alternation. Only a small proportion of sentences exhibit three or more switches.

5.3. Phrase-Level Switching Analysis

Because SECT operates at the phrase level, we analyze the types of syntactic phrases inserted during switching. Table 5 reports the distribution of inserted phrase types.

Pair	NP	VP	PP	Other
Hausa–Yoruba	41.2	27.5	21.3	10.0
Hausa–English	38.7	29.6	20.4	11.3

Table 5. Phrase-type distribution of switching segments (%).

The results show that noun phrases account for the largest

proportion of switching segments in both language pairs. This observation aligns with linguistic studies of natural code-switching, which report that content-bearing phrases such as noun phrases are commonly inserted into the matrix language structure.

5.4. Generation Quality

We further evaluate the effectiveness of the generation pipeline by measuring several quality indicators, including valid sentence generation, switching occurrence, and speech synthesis success.

Metric	Hausa–Yoruba	Hausa–English
Valid generated sentences (%)	96.4	97.1
Sentences with at least one switch (%)	92.0	89.0
Average embedded span length	3.8	3.5
Speech synthesis success (%)	98.6	98.9
Voice normalization success (%)	97.8	97.5

Table 6. Generation quality metrics for the proposed pipeline.

As shown in Table 6, more than **96%** of generated outputs satisfy the linguistic constraints and validation checks defined in the generation pipeline. Nearly all generated sentences are successfully converted into speech using the MMS text-to-speech system.

The examples illustrate that switching occurs primarily at phrase boundaries while preserving the Hausa grammatical frame of the sentence. Yoruba and English segments appear as embedded phrases, consistent with the Matrix Language Frame assumption used in the generation process.

5.5. Audio Duration Analysis

Finally, we analyze the duration of the synthesized speech recordings. Table 7 reports statistics for the generated audio.

Pair	Avg Duration (s)	Min (s)	Max (s)
Hausa–Yoruba	8.4	2.1	15.3
Hausa–English	9.1	2.4	16.8

Table 7. Duration statistics of synthesized code-switched speech.

The synthesized utterances range between approximately 5 and 17 seconds depending on sentence length and speech rate. The average utterance duration is 9 seconds, resulting in approximately **18.5 hours of generated code-switched speech**. This dataset therefore represents one of the largest automatically generated code-switched speech resources for African languages.

The examples illustrate that switching occurs at phrase boundaries rather than isolated word replacements. This behavior aligns with predictions from both the Matrix Language Frame model and Equivalence Constraint Theory. In particular, the Hausa matrix language maintains the grammatical structure of the sentence while Yoruba phrases are inserted as embedded constituents.

Compared with word-level switching methods, phrase-level switching produces more natural bilingual utterances because syntactic constituents remain intact during language alternation. This observation supports previous findings that linguistically constrained switching improves grammaticality in generated code-switched text.

5.6. Effect of Synthetic Data on ASR Performance

The ASR experiments demonstrate that incorporating synthetic code-switched speech significantly improves recognition performance compared to monolingual training. In particular, models trained with SECT-generated speech show consistent reductions in both Word Error Rate (WER) and Mixed Error Rate (MER).

Method	WER (%) ↓	MER (%) ↓
Monolingual ASR	34.8	37.5
Random Switching	31.6	34.2
Word-Level ECT	29.4	31.7
Phrase-Level SECT (Proposed)	**25.8**	**28.1**

Table 8. ASR performance comparison on synthetic code-switched speech. Lower is better.

Table 8 shows that linguistically constrained phrase-level switching significantly improves ASR performance compared to monolingual and unconstrained switching approaches. The proposed SECT-based method achieves the lowest WER and MER, demonstrating that syntactically valid code-switching patterns provide more effective training signals for multilingual ASR systems.

5.7. Impact of Voice Cloning

We also analyze the effect of voice normalization through OpenVoice. Without voice cloning, synthesized segments produced by multilingual TTS models exhibit different speaker identities across languages. This inconsistency introduces acoustic discontinuities that negatively affect ASR training.

Voice cloning ensures consistent speaker identity across language segments, producing smoother acoustic transitions and improving recognition performance. As shown in the ablation study, removing the voice cloning stage results in a measurable degradation in ASR accuracy.

These results highlight the importance of maintaining acoustic consistency when generating synthetic code-switched speech.

6. Conclusion

This paper presented a framework for generating synthetic intra-sentential code-switched speech using large language models and multilingual text-to-speech synthesis. The proposed approach combines Matrix Language Frame theory with a phrase-level extension of the Equivalence Constraint Theory to generate linguistically valid code-switched text. The generated text is converted into speech using MMS TTS and normalized with OpenVoice voice cloning to ensure consistent speaker identity.

Experiments on Hausa–Yoruba and Hausa–English demonstrate that the resulting synthetic dataset improves ASR performance and enables effective switch-aware ASR training through explicit switch-point annotations.

Our findings show that linguistically grounded synthetic data generation can serve as an effective strategy for developing speech technologies for underrepresented multilingual communities.

Limitations

This study has several limitations. First, the proposed dataset remains relatively small compared with large-scale multilingual speech corpora, which may limit generalization across speakers, dialects, and domains. Second, the work focuses on Hausa–English and Hausa–Yorùbá, so the findings may not transfer directly to other African code-switched language pairs. Third, switch-region annotation may still involve ambiguity, especially in cases of lexical borrowing, overlap, or gradual transitions between languages. Finally, the synthetic augmentation pipeline, although linguistically constrained, may not fully capture the prosody, spontaneity, and pronunciation variability of naturally occurring code-switched speech.

References

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

Costa-jussà, M. R., Cross, J., Ørschler, D., Webster, K., Heafield, K., Pino, J., Subramanian, S., Wu, Z., Elbayad, M., Maillard, J., et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Dubey, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hussein, A., Biswas, A., Yılmaz, E., and van den Heuvel, H. Speech collage: Code-Switching data augmentation without TTS. *arXiv preprint arXiv:2309.15674*, 2023.

Kuwanto, G., Agarwal, C., Winata, G. I., and Wijaya, D. T. EZSWITCH: Linguistically guided Code-Switching generation with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Liang, Y., Chen, Z., and Li, H. Speech editing for Code-Switching speech data augmentation. *arXiv preprint arXiv:2306.08588*, 2023.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. SEAME: A mandarin–English Code-Switching speech corpus in south-east asia. In *Proceedings of Interspeech 2010*, pp. 1986–1989, 2010.

Myers-Scotton, C. *Duelling Languages: Grammatical Structure in Code-Switching*. Oxford University Press, Oxford, 1997.

Nakayama, Y., Tanaka, T., and Toda, T. Machine speech chain with Code-Switching for semi-supervised ASR and TTS. *IEICE Transactions on Information and Systems*, E104-D(10):1690–1700, 2021. doi: 10.1587/transinf.2021EDP7005.

Nguyen, T., Li, X., and Fung, P. Synthetic Code-Switched text generation for semi-supervised bilingual ASR. *arXiv preprint arXiv:2210.12214*, 2022.

Poplack, S. Sometimes i'll start a sentence in spanish y termino en español: Toward a typology of Code-Switching. *Linguistics*, 18(7–8):581–618, 1980.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Pratapa, A., Choudhury, M., and Sitaram, S. 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 (ACL)*, pp. 1543–1553, 2018. doi: 10.18653/v1/P18-1143.

Qin, Z., Zhao, W., Yu, X., and Sun, X. OpenVoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2024.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 28492–28518, 2023.

Sharma, P., Yılmaz, E., van den Heuvel, H., and van Leeuwen, D. Improving low-resource Code-Switched ASR using synthetic data augmentation. In *Proceedings of Interspeech 2020*, 2020.

Sitaram, S., Choudhury, M., Bali, K., He, Y., Rao, S., and Black, A. W. A survey of Code-Switched speech and language processing. *Computer Speech and Language*, 54:28–44, 2019.

Solorio, T. and Liu, Y. Learning to predict Code-Switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 973–981, 2008.

Winata, G. I., Madotto, A., Lin, Z., and Fung, P. Multilingual speech recognition for Code-Switched speech. In *Proceedings of Interspeech 2021*, pp. 3730–3734, 2021.

Winata, G. I., Madotto, A., Lin, Z., and Fung, P. Code-Switching in speech and language processing: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:146–166, 2023.

Yu, J., Li, B., Zhang, S., and Meng, H. Synthetic Code-Switching data augmentation for end-to-end speech recognition. *arXiv preprint arXiv:2303.10949*, 2023.

Zhang, Z., Winata, G. I., and Fung, P. Code-Switching language modeling with neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3220–3232, 2023.

A. Full Pipeline Algorithm

Algorithm 1 Full pipeline for linguistically constrained code-switched speech generation and voice normalization

Input: Source sentence S_h, target sentence S_t, similarity threshold τ, reference speaker embedding v_r
Output: Voice-normalized code-switched speech waveform X'

$E_h \leftarrow \text{Encoder}(S_h)$ $E_t \leftarrow \text{Encoder}(S_t)$
alignments $\leftarrow \emptyset$
for $i \leftarrow 1$ **to** $|S_h|$ **do**
  for $j \leftarrow 1$ **to** $|S_t|$ **do**
    sim $\leftarrow \text{Cosine}(E_h(i), E_t(j))$
    if sim $> \tau$ **then**
      Add (i, j) to alignments
phrases $\leftarrow \text{ExtractPhrases}(S_h)$
validPhrases $\leftarrow \emptyset$
foreach $P_k \in$ phrases **do**
  if *all tokens in* P_k *are aligned* **then**
    Add P_k to validPhrases
$S_{cs} \leftarrow S_h$
foreach $P_k \in$ validPhrases **do**
  Replace P_k in S_{cs} with the aligned phrase from S_t
segments $\leftarrow \text{SplitByLanguage}(S_{cs})$
outputSegments $\leftarrow \emptyset$
foreach $L_i \in$ segments **do**
  $x_i \leftarrow \text{MMS\_TTS}(L_i)$
  $x'_i \leftarrow \text{VoiceClone}(x_i, v_r)$
  Add x'_i to outputSegments
$X' \leftarrow \text{Concat}(\text{outputSegments})$
return X'

B. Qualitative Examples

Table 9 presents representative examples of generated intra-sentential code-switched sentences produced by the proposed SECT-guided generation framework.

Hausa	Translation	Generated Code-Switched Sentence
Rahoton ya yi matuar sukar tsarin gwamnati a kasar.	Ìròyìn náà ṣofintoto eto ìjba náà gidigidi.	Rahoton ya yi matuar sukar *eto ìjba náà gidigidi*.
Likitan ya bayyana cewa binciken yana kan matakin farko.	The doctor explained that the research is still in its early stage.	Likitan ya bayyana cewa *the research is still in its early stage*.
Yanzu muna nazarin sabbin hanyoyin magani.	A n ṣe ìtúpal àwn na ìtjú tuntun.	Yanzu muna nazarin *àwn na ìtjú tuntun*.

Table 9. Examples of generated intra-sentential code-switched sentences. Embedded-language segments are shown in italics.