A Cross-lingual Comparison of Human and Classification Model Entrainment Behavior in Code-switched Speech Settings

Debasmita Bhattacharya[1]*, Siying Ding[1], Alayna Nguyen[2]†, Julia Hirschberg[1]
[1]Department of Computer Science, Columbia University [2]Instagram

Abstract

Conversational entrainment is well-studied in monolingual and written contexts, but remains underexplored in spoken code-switching (CSW). We present a cross-lingual analysis of entrainment in Mandarin-English, Hindi-English, and Spanish-English dialogue, replicating and extending an established Spanish-English framework (Bhattacharya et al., 2024) to two new language pairs, and show that, while lexical entrainment generalizes across language pairs, entrainment over acoustic-prosodic and CSW style aspects exhibits context-specific variation. We build on these findings by asking whether classification models capture these human behavioral patterns. Using feature importance and ablation analyses, we find that classical and Transformer-based classifiers detect entrainment reasonably well but consistently prioritize features other than those most salient to human entraining behavior. Our approach introduces a human-grounded framework for evaluating model decision-making in multilingual stylistic contexts, and suggests challenges for developing conversational agents capable of producing naturalistic code-switched speech.

1 Introduction

When speakers *entrain* to one another, each subconsciously adapts aspects of the interlocutor's communication style into their own language production, coordinating features of diction, syntax, speaking rate, voice quality, pauses, and even facial expression and gesture, among others (Danescu-Niculescu-Mizil et al., 2011; Giles et al., 1991; Levitan and Hirschberg, 2011; Maurer and Tindall, 1983; Burgoon et al., 2006; Chartrand and Bargh, 1999). This interactive, multimodal phenomenon shapes conversational dynamics, enabling successful communication by ensuring mutual comprehension, fostering rapport, and making a positive impression on the interlocutor (Reitter and Moore, 2007; Nenkova et al., 2008; Levitan et al., 2012).

While most existing work on entrainment has focused on monolingual, and especially English, communication, the last decade has increasingly seen entrainment studied in additional languages (Xia et al., 2014; Levitan et al., 2015; Kejriwal and Štefan Beňuš, 2023; Weise et al., 2021) and in multilingual settings (Kootstra et al., 2010; Bawa et al., 2018; Parekh et al., 2020). Among studies of *code-switched entrainment*, where speakers adapt while alternating between producing at least two different language varieties, most analyses have focused on only a few language pairs, most often studying entraining behavior in Spanish-English (Soto et al., 2018; Ahn et al., 2020; Bhattacharya et al., 2024). Though some studies have incorporated additional linguistic contexts and performed cross-lingual comparisons of lexical entrainment in written code-switching, few have done so in the spoken domain, giving rise to the present work.

We explore the conversational dynamics of spoken code-switched dialogue in Mandarin-English, Hindi-English, and Spanish-English – three CSW pairs that each involve English – performing cross-lingual comparisons that validate and extend prior work, e.g., Bhattacharya et al. (2024), by grounding it in broader linguistic context. We first ask whether code-switched entrainment patterns in speech are consistent across language pairs involving different language families, levels of typological distance, and cultural settings. We then build upon this by asking whether entrainment detection models attend to the same features of entrainment as people do during spontaneous code-switched communication. We find **1)** generally consistent multimodal patterns of entrainment across diverse language pairs, with slight variation reflecting linguistic and cultural idiosyncrasies of the languages involved; and **2)** while classifier models detect entrainment reasonably well across levels of conversational

*Corresponding author: debasmita.b@cs.columbia.edu
†Work done as an undergraduate at Columbia University.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 16906–16936
October 24-29, 2026 ©2026 Association for Computational Linguistics

granularity and language pairs, they seem not to use the most relevant entraining features to do so, relative to humans. Our approach and findings suggest that empirically grounded human behavioral patterns provide a valuable lens for interrogating how and why classification models detect stylistic phenomena, which has implications for understanding stylistic tasks beyond entrainment.

Our work is the first to cross-lingually validate entrainment patterns of spoken code-switching (CSW), and to use those patterns as a human-grounded framework for evaluating classifier behavior. In sum, our contributions are as follows:

- We replicate the Spanish-English entrainment approach applied in Bhattacharya et al. (2024) across new Mandarin-English and Hindi-English settings, and extend it in a cross-lingual comparison of statistical entrainment patterns across three corpora.
- We analyze models' feature importance attribution and ablation behavior to go beyond evaluating *whether* models can detect entrainment, and also highlight *why* they do so.
- We release a manually annotated version of the MaSaC corpus with new CSW style labels to support further research on spoken CSW.[1]

2 Related Work and Research Questions

Entrainment in code-switched contexts. Our work builds directly on Bhattacharya et al. (2024), who studied lexical, acoustic-prosodic, and multilingual stylistic entrainment in spontaneous Spanish-English conversations from the Bangor Miami (BM) corpus (Deuchar, 2011). The authors applied the analytical framework of Levitan and Hirschberg (2011), which distinguishes entrainment at the turn-level from that at the conversation-level, and defines *proximity*, *convergence*, and *synchrony* as unique manifestations of spoken entraining behavior (see Appendix A.1 for definitions). Bhattacharya et al. adapted this framework to study entrainment across CSW aspects and acoustic-prosodic features, while applying methods that draw on salient word classes and perplexity-based measures (Nenkova et al., 2008; Gravano et al., 2014) to examine lexical entrainment in BM speech. They found that entrainment trends from monolingual domains appear in code-switched language production, with new patterns specific to spoken CSW, but did not examine whether these findings generalize beyond Spanish-English. Other work, such as Bawa et al. (2018) and Parekh et al. (2020), examined entrainment in Hindi-English CSW settings, focusing on lexical representations of speech transcripts and human-machine generated written exchanges, respectively. While both studies offered comparisons to entrainment patterns in Spanish-English CSW, these remained restricted to the written modality, giving rise to our first research question **RQ1:** *To what extent do entrainment patterns of spoken CSW generalize across distinct language pairs?* To answer **RQ1**, we build on existing work on Spanish-English in BM, and additionally focus on Mandarin-English and Hindi-English; these four languages together represent the most spoken world languages (Ethnologue, 2026).

Comparing human and model multilingual behavior. Recent work has proposed frameworks for explaining multilingual language models' ability to classify (and produce) human-like linguistic style and interaction dynamics (Namazifard and Galke Poech, 2025; Blevins et al., 2026; Lin et al., 2025). We highlight Havaldar et al. (2023), who created multilingual politeness lexica and compared feature importance values derived from model encodings to human productions. Such work informs our second research question **RQ2:** *To what extent do classifier models attend to those entraining features used by humans who perform entrainment in code-switched speech?* To answer **RQ2**, we examine a suite of classical and Transformer-based machine learning classifiers, using binary entrainment detection across language pairs to analyze feature importance in each setting.

3 Corpora

We examine three corpora of informal multilingual speech. Section 5.1 focuses on the **SEAME** corpus of spontaneous Mandarin-English dialogues (Lyu et al., 2010) and the **MaSaC** corpus of Hindi-English situational comedy television speech (Bedi et al., 2023), comparing both to the **Bangor Miami** corpus of Spanish-English spontaneous conversations (Deuchar, 2011).[2] Section 5.2 analyzes all three corpora, which comprise recorded and transcribed monolingual and code-switched utterances from conversations between two or more speakers.

The corpora cover shared topics, e.g., daily routines, interpersonal relationships, and aca-

[1] Data available at https://tinyurl.com/2245wxe4

[2] See Appendix A.2 for data licensing details.

Corpus	Speech hours	# of dialogues	# of utterances	# of unique speakers	% code-switched
Bangor Miami	35	56	46,901	84	5
SEAME	192	256	110,146	156	52
MaSaC	13	1,190	6,476	5	54

Table 1: Summary of relevant corpus statistics for each data set.

demic/professional concerns. Unlike SEAME and BM, MaSaC does not consist of naturally-occurring speech. However, scripted television conversation is designed to sound natural and is therefore a suitable proxy for studying naturalistic phenomena like entrainment (Bawa et al., 2018; Danescu-Niculescu-Mizil and Lee, 2011), especially given the lack of other genre-matched, openly accessible speech data on the Hindi-English language pair.

Both BM and SEAME contain utterance-level annotations (Bhattacharya et al., 2024, 2025) that identify code-switched and monolingual utterances, and distinguish between simple *insertional* code-switches of single words or short phrases (*I*), complex *alternational* code-switches at grammatical clause boundaries (*A*), and "other" (*O*) strategies of CSW resembling tag-switching (Muysken, 2000; Poplack, 1980). We add the same sets of labels to MaSaC. First, we use Microsoft's open-source language identification (LID) tool[3] to obtain token-level LID tags for the Romanized Hindi and English in MaSaC's speech transcripts, and accordingly infer utterance-level language labels: code-switched vs. monolingual English or Hindi. We then add manual annotations of CSW strategies on the multilingual utterances. All strategy labels are created by the first author and verified by a volunteer MS student, with no disagreements. Both have first-language proficiency in English; the former has intermediate competency in Hindi comprehension, while the latter is a native speaker of Hindi. We find that 92% of code-switched MaSaC utterances use insertional CSW, 14% use alternational, and less than 1% use "other" CSW. This strategy distribution is comparable to those in BM (72% *I*, 13% *A*, 18% *O*) and SEAME (89% *I*, 12% *A*, 7% *O*) (Bhattacharya et al., 2024, 2025). We summarize additional corpus statistics in Table 1.

To enable direct comparison to known *dyadic* entrainment patterns in 39 of the BM conversations (Bhattacharya et al., 2024), we filter SEAME and MaSaC for conversations between just two speakers where speech audio and corresponding transcripts are available for both speakers. We retain only those containing 4 or more turns to ensure sufficiently long dialogue for conversation-level entrainment calculations (Section 4.1). We accordingly use 40% of SEAME (42,893 utterances over 33 dialogues) and 25% of MaSaC (1,623 utterances over 334 dialogues) in our subsequent analyses.

[3] https://github.com/microsoft/LID-tool

4 Methods

4.1 Computing code-switched entrainment.

For statistical analyses of entrainment in Mandarin-English and Hindi-English CSW, we replicate Bhattacharya et al. (2024)'s method, which we summarize here and detail fully in Appendix A.1.

The **lexical** feature set comprises four word classes known to be often used and entrained upon in spontaneous dialogue (Nenkova et al., 2008; Bhattacharya et al., 2024): *corpus-level most frequent words* (top-100 and top-25), *conversation-level most frequent words* (top-25), *affirmative cues*, and *fillers* (Appendix A.3 lists affirmative cues and fillers in each corpus). It also operationalizes broad similarity between speakers' *overall language use* via perplexity from Kneser-Essen-Ney smoothed trigram language models trained on each speaker's transcripts with the KenLM toolkit. The **acoustic-prosodic** feature set covers utterance-level *pitch*, *intensity*, *jitter*, *shimmer*, *harmonics-to-noise ratio* (HNR), and *speaking rate* in syllables per second. All acoustic-prosodic features are extracted via Parselmouth (Jadoul et al., 2018) with default parameter values and are z-score normalized by speaker. Utterance-level **CSW style** metrics include a binary flag for *CSW presence*, a continuous *CSW ratio* normalizing switching quantity by utterance length (Soto et al., 2018), and a categorical label of the *CSW strategy* in use (*I*, *A*, or *O*), if any.

We study entrainment over **acoustic-prosodic** and **CSW style** features via turn- and conversation-level proximity, convergence, and synchrony – respectively, absolute similarity, change in similarity over time, and interlocutor coordination – following the operationalizations of Levitan and Hirschberg (2011) as adapted by Bhattacharya et al. (2024). We measure entrainment over **lexical** word

classes using proximity and the stated perplexity approach for overall language use.

4.2 Evaluating model entrainment behavior.

We next examine how a suite of binary classifiers attends to turn- and conversation-level entrainment in spoken CSW with the goal of verifying whether, during classification, models assign the most importance to the conversational features over which humans significantly entrain. We use both classical machine learning methods and larger Transformer-based architectures. The former group comprises eight classical models including *SVC*, *Decision Tree*, *Random Forest*, and *MLP* classifiers, all implemented in `scikit-learn 1.6.1` (full list in Table 4). The latter group includes a baseline *XGBoost* model, and adapted *TabTransformer* (Huang et al., 2020), *FT-Transformer* (Dai et al., 2025), *XLM-R* (Conneau et al., 2020), *mBERT* (Devlin et al., 2019), *RemBERT* (Chung et al., 2020), *WavLM* (Chen et al., 2021), and *mHuBERT* (Zanon Boito et al., 2024) pre-trained contextual embedding models; we fuse each with an MLP for classification[4] and list their licenses in Appendix A.4.

To evaluate these models, we create a 75/15/10 train/validation/test stratified data split per corpus. For each classifier, we first perform grid search with 5-fold cross-validation, selecting the best hyperparameters by accuracy.[5] Applying those in training, we calculate the model's test-time feature importance values using built-in attributes when available, and Shapley values (Lundberg and Lee, 2017) otherwise. We then perform feature-level ablations with the same (best) model and 5-fold cross-validation to indirectly evaluate the importance of each feature of entrainment from Section 4.1.

At the **conversation-level**, the positive (entraining) class comprises conversations exhibiting significant proximity on an empirically determined feature set of interest:[6] all CSW style metrics (BM, SEAME) or lexical word classes (MaSaC); the negative class is its complement. Classes per corpus are roughly balanced. At the **turn-level**, the positive class is similarly empirically determined (Section 5.1) and comprises consecutive turn pairs with significant proximity on a majority of CSW style metrics (BM, SEAME) or acoustic-prosodic features (MaSaC), with class imbalance handled via oversampling. At both levels of conversational granularity, a model input is a feature vector covering all feature sets from Section 4.1 for each speaker, with continuous features averaged and non-continuous features first converted to binary flags; full construction details are in Appendix A.6. This results in a *tabular/numerical* representation, which we concatenate with model-specific embeddings of conversation transcripts (text embeddings) or audio recordings (speech embeddings).

5 Results

5.1 Cross-lingual human entraining behavior.

We first examine whether human entrainment patterns over lexical, acoustic-prosodic, and CSW style features in code-switched speech are consistent across language pairs differing in typological distance, language family, and cultural context. Table 2 shows representative examples in each case. We report effect sizes and Bonferroni-corrected significance results in Appendices A.7 and A.8.

Entrainment on lexical aspects of code-switched speech. Within SEAME, we find strong lexical entrainment on all word classes. Over the most frequent corpus-level words, 100% of SEAME conversations exhibit entrainment ($p < 0.05$ in each case). We find similar entrainment on affirmative cues (97% of conversations, $p < 0.05$) and fillers (91% of conversations, $p < 0.05$). In terms of overall language use, when including out-of-vocabulary (OOV) words, 88% of conversations exhibit within-conversation entrainment ($p < 0.05$); 77% involve both interlocutors entraining to their partner's overall language use, while in the remaining 11% only one speaker entrains to the other. Results excluding OOV words are nearly identical (Appendix A.9).

Within MaSaC, we find significant lexical entrainment on the word classes involving frequent words. Over the corpus-level most frequent words (top-100 and top-25), 84% and 92% of MaSaC conversations respectively exhibit entrainment ($p < 0.05$ for both). While entrainment over the top-25 conversation-level words is significant ($p < 0.05$), it appears only in 10% of conversations. In terms of overall language use, 70% of conversations show

[4] We train only the MLP within each pre-trained fusion model, freezing the parameters of the pre-trained foundation.

[5] We list all hyperparameter details in Appendix A.5.

[6] Based on Section 5.1, only the feature set of interest should discriminate entraining from non-entraining speech: other feature sets show entrainment in either nearly all or nearly no instances, providing little discriminative signal. This data-dependence is a property of the empirical entrainment patterns reported in Section 5.1 rather than a fixed methodological choice made in advance, and we adopt the same feature-set determination procedure uniformly across corpora.

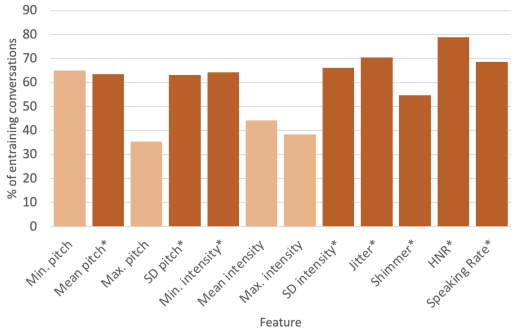

Figure 1: Turn-level acoustic-prosodic proximity in MaSaC. * and darker bars indicate significant features.

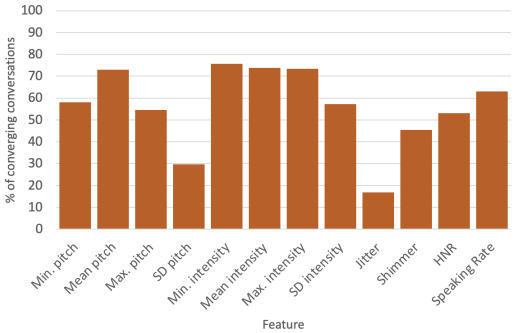

Figure 2: Turn-level acoustic-prosodic convergence in MaSaC. Most conversations converge *strongly*; non-converging ones diverge *weakly*. All features significant.

evidence of within-conversation entrainment when including OOV words ($p < 0.05$): 52% exhibit bidirectional entrainment while 18% involve only one speaker entraining to the other. OOV-excluded results are almost identical (Appendix A.9).

Entrainment on acoustic-prosodic aspects of code-switched speech. We find limited acoustic-prosodic entrainment within SEAME. Most conversations exhibit turn-level proximity and synchrony, but these achieve statistical significance on fewer than half of the acoustic-prosodic features (Figures 5 and 7 in Appendix A.9), with uniformly small effect sizes across tested turn-level features ($|d| < 0.04$; Appendix A.7). Similarly, significant turn-level convergence occurs in half of SEAME conversations over only a few pitch and intensity features (Figure 6 in Appendix A.9), suggesting inconsistent acoustic-prosodic entrainment of this kind. In MaSaC, in contrast, we find much stronger evidence of acoustic-prosodic entrainment. In terms of turn-level proximity, more than 60% of conversations entrain significantly over two-thirds of acoustic-prosodic features (Figure 1). The majority of conversations also consistently converges strongly across features at the turn-level (Figure 2).

Entrainment on CSW style aspects of code-switched speech. We find significant proximity at both the turn-level (67% of conversations, $p < 0.05$) and the conversation-level (97% of conversations, $p < 0.05$) for CSW presence in SEAME, as well as weak, but statistically insignificant, turn-level synchrony in 67% of conversations. In terms of CSW ratio, we again find significant turn- and conversation-level proximity in 70% and 97% of conversations, respectively ($p < 0.05$ in both). As with CSW presence, there is weak turn-level synchrony over CSW ratio in 91% of conversations, though this, too, fails to reach statistical significance. Finally, examining CSW strategies, we find significant turn-level synchrony over *insertional* CSW (67% of conversations, $p < 0.05$), and significant conversation-level proximity over *alternational* CSW (88% of conversations, $p < 0.05$). Effect sizes reinforce granular asymmetry: SEAME's turn-level CSW style proximity effects are consistently weak, while the corresponding conversation-level effects are medium-to-large (Appendix A.7).

Our findings on MaSaC are strikingly different from those on SEAME. We find only significant turn-level convergence across metrics of CSW style: 57% of conversations converge on CSW presence (54% *strongly*, 3% weakly or moderately), 91% converge on CSW ratio (82% *strongly*, 9% weakly or moderately), and 62% converge (*strongly*) on *insertional* CSW. Across metrics, this convergence is consistently statistically significant. Since MaSaC's scripted nature limits our ability to disentangle whether these patterns reflect Hindi-English-specific entrainment norms or corpus- and genre-specific effects (see Discussion below), we scope these findings to the MaSaC corpus only.

Discussion. Bhattacharya et al. (2024) previously found that monolingual entrainment trends extend to Spanish-English CSW in BM, alongside new patterns specific to CSW itself (e.g., entrainment on CSW presence, ratio, and strategy), but did not test whether these findings generalize beyond Spanish-English. Compared to this prior work, we generally find consistent *lexical* patterns across language pairs, robust to Bonferroni correction both across and within corpora (Appendix A.8). Both Mandarin-English and Spanish-English CSW exhibit lexical entrainment over all word classes. In Hindi-English, entrainment is absent over conversation-level frequent words, affirmative cues, and fillers. We attribute the former

Corpus	Type	Exchange	Feature	Value(s)
SEAME	Lexical	**S1:** 对 *then* 我 就 *like yah* 我 那 时候 [*Right, then I was like, yeah, at that point I...*] **/ S2:** *then* 你 就 过 不 了 [*...then you wouldn't be able to get across.*]	Frequent word	'*then*'
MaSaC	Lexical	**S1:** *Chalo, chalo, come on, Paris nahi jaana!* [Come on, come on, come on, I'm not going to Paris!] **/ S2:** *Nahi, nahi, nahi!* [*No, no, no!*]	Frequent word	'*nahi*'
SEAME	Acoustic-prosodic, non-entraining	**S1:** *what what what* 什么? [*What what what what?*] **/ S2:** *do you think our hall has pretty girls?*	SD pitch (Hz)	211.5 vs. 12.4
MaSaC	Acoustic-prosodic, entraining	**S1:** *That's totally ridiculous Indravardhan* **/ S2:** *Arey Maya...yakin karo* [*Oh Maya...believe me*]	Mean pitch (Hz)	231.3 vs. 231.4
SEAME	CSW style	**S1:** 对 那个 *area five* 多少 [*right, that Area Five, how much is it?*] **/ S2:** *okay kay* 让我拿出我的 *calculator* [*Okay, okay, let me take out my calculator*]	CSW presence	—
MaSaC	CSW style	**S1:** *Poore nau rupay waste!* [*All 9 rupees wasted!*] **/ S2:** *...Pachhatar paise vaala pen hai ye?* [*...Is this a 75-paise pen?*]	CSW strategy: *I*	—

Table 2: Turn-level examples of entrainment (and disentrainment) as discussed throughout Section 5.1.

absence to shorter conversations in MaSaC limiting its scope: fewer dialogue turns and shorter utterances (Table 1) leave less opportunity for a stable high-frequency word set to emerge and for speakers to entrain on them. We suspect the latter two null results are due to these lexical categories not appearing in scripted MaSaC dialogue, an entrainment effect noted in other scripted speech settings (Blevins et al., 2026); their effect sizes are near zero (Appendix A.7), consistent with a genuine absence of entrainment rather than low statistical power. This underscores a general caveat: MaSaC findings may reflect corpus properties as much as linguistic ones, and replication on naturally occurring Hindi-English speech is necessary before strong conclusions can be drawn. Entrainment on overall language use, in contrast, is robust across all three corpora (SEAME, MaSaC, and BM).

In terms of *acoustic-prosodic* entrainment, both Mandarin-English and Hindi-English mirror known Spanish-English patterns on the *absence* of conversation-level proximity and convergence, and the *absence* of significant turn-level synchrony. Where entrainment is *present*, MaSaC's Hindi-English patterns align more closely with BM's Spanish-English ones than do those from SEAME's Mandarin-English, which lacks acoustic-prosodic entrainment. One explanation is Mandarin's unique tonal character and distinct intonation norms relative to Spanish and Hindi, though corpus-level differences in speaker population and interaction setting may also contribute slightly.

With respect to *CSW style*, the Mandarin-English patterns we find align strongly with those known for Spanish-English: speech in both language pairs exhibits entrainment over most multilingual features of CSW conversation via proximity and synchrony. MaSaC's Hindi-English CSW differs markedly, exhibiting only turn-level convergence, and less prominent entrainment over CSW style metrics overall. We offer a preliminary hypothesis for this difference. Hindi-English speakers exhibit strikingly higher M- and I-index (Barnett et al., 2000; Guzman et al., 2016) values ($M = 0.49$; $I = 0.48$) than Mandarin-English ($M = 0.14$; $I = 0.13$) or Spanish-English ($M = 0.08$, $I = 0.01$) speakers, indicating more abundant and frequently interleaved CSW. One interpretation is that when speakers already code-switch prevalently and independently, there is less room for one speaker's style to influence the other's, leading to a ceiling on stylistic entrainment. We do not independently test this hypothesis and note that disentangling possible contributions of CSW linguistic norms and speech genre would require comparing naturally-occurring Hindi-English dyads across several CSW rates, which we leave for future work.

In response to **RQ1**, *lexical* entrainment in spoken CSW generalizes robustly across language pairs, while *acoustic-prosodic* and *CSW style* patterns exhibit variation that may reflect the unique linguistic character of the non-English language and its associated cultural norms, as well as potential corpus-specific interactional factors. We offer further qualitative comparison to prior monolingual and CSW entrainment findings in Appendix A.10.

5.2 Classifier models' entrainment behavior.

We next investigate how classifiers behave when detecting entrainment, relative to human entraining

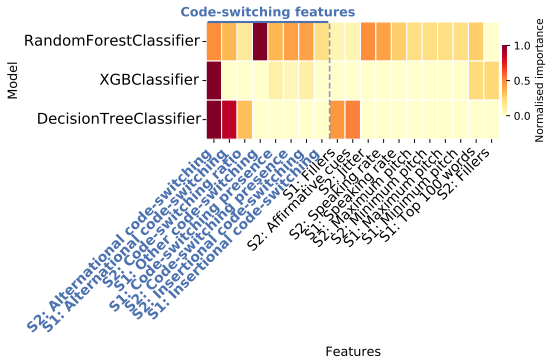

Figure 3: Highest feature importance values across best-performing BM models at the conversation-level. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable. For visualization across all features, including many near-zero values, see Figure 8 in Appendix A.11.

behavior from multiple CSW contexts and levels of conversational granularity. We apply both model-internal and model-external evaluation methods, first inspecting alignment between the most important features to the best-performing models[7] and those known to be statistically salient for entrainment per corpus, and then examining the impact of feature-level ablations across all models. Table 3 summarizes best-performing classifier accuracy by corpus and conversational granularity; additional per-model details are in Figures 8-12. We visualize conversation-level results for BM; SEAME and MaSaC figures are in Appendices A.11 and A.12.

Since each corpus empirically differs in which feature set discriminates entraining from non-entraining speech (Footnote 6), the specific classification task differs by corpus; cross-corpus comparisons of model behavior should be interpreted with this in mind. Note that statistical salience in human entrainment analyses need not reflect causal importance in such conversations, and classifiers' feature importance estimates may be influenced by the task operationalization. We examine alignment between model predictive behavior and our statistical characterization of human entrainment, rather than underlying cognitive mechanisms.

Models' feature importance values. At the **conversation level** of BM, 3 models achieve near perfect accuracy on distinguishing dialogue in which speakers entrain on CSW style features from that lacking such entrainment: *Decision Tree*,[8] *Random Forest*,[9] and *XGBoost*.[10] Each assigns the greatest feature importance to at least one relevant CSW style metric (Figure 3), yet two unexpected patterns emerge across these models. First, both *DecisionTree* and *XGBoost* concentrate importance on a *single* CSW feature rather than across the full set of human-relevant ones; this is surprising given that positive-class conversations exhibit significant entrainment on *each* CSW feature. We also note the assignment of relatively high importance to acoustic-prosodic features, particularly in the *RandomForest*; these should provide little signal for distinguishing conversations with and without entrainment over CSW aspects, given that nearly all BM conversations exhibit significant conversation-level acoustic-prosodic entrainment (Bhattacharya et al., 2024), and these do not correlate with CSW features (details in Table 13 in Appendix A.13).

These alternative predictive patterns replicate across corpora. All but one top-performing SEAME classifier assigns the greatest importance to features outside the relevant CSW style set (Figure 9); the best MaSaC models do the same for the relevant lexical set (Figure 10). Across corpora, we

Corpus	Level	Classifiers (Accuracy)
BM	Conv.	*Decision Tree** (0.99), *Random Forest** (0.99), *XGBoost*† (0.99)
	Turn	Excluded (see Footnote 11)
SEAME	Conv.	*RemBERT+WavLM fusion*† (0.99), *MLPClassifier** (0.99), *XLM-fusion*† (0.99), *mBERT-fusion*† (0.99), *RemBERT-fusion*† (0.99), *GradientBoostingClassifier** (0.99), *GaussianNB** (0.99), *TabTransformer-adapted*† (0.99), *FT-Transformer*† (0.75)
	Turn	*mBERT-fusion*† (0.70), *RemBERT-fusion*† (0.67), *TabTransformer-adapted*† (0.65)
MaSaC	Conv.	*LogisticRegression** (0.71), *mBERT-fusion*† (0.68), *SVC** (0.65)
	Turn	*mBERT+WavLM-fusion*† (0.77), *WavLM-fusion*† (0.74), *mBERT-fusion*† (0.70), *mHuBERT-fusion*† (0.70), *XLM-fusion*† (0.69)

Table 3: Best-performing *classical and †Transformer-based classifiers by corpus & conversational granularity.

[7]For conversation-level feature importance, we examine only models whose performance exceeds a test-time accuracy of 75% on BM and SEAME, and 65% on MaSaC; for the turn-level we use a 65% threshold across corpora. These maintain the reliability of feature importance observations. Slightly lower thresholds for MaSaC and turn-level analyses account for substantially smaller or noisier datasets; higher thresholds would overly restrict qualitative feature interpretation.

[8]criterion: gini, max. depth: 3, min. samples split: 2

[9]criterion: log loss, max. depth: 1, min. samples split: 5

[10]booster: gbtree, lr: 0.01, max. depth: 3, # estimators: 50

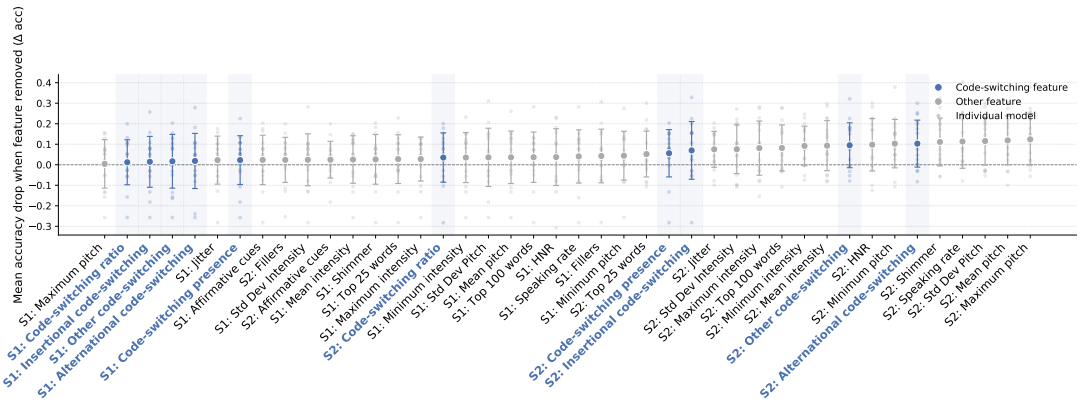

Figure 4: Aggregated feature ablation results for BM at the conversation-level. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful.

find little evidence that **conversation-level** classification decisions are guided by the features most expected from human entrainment behavior.

At the **turn level**, several models surpass the classification accuracy threshold with respect to CSW style features on SEAME and acoustic-prosodic features on MaSaC (Table 3 and Appendix A.11).[11] For SEAME, none assigns meaningful importance to any relevant CSW style features (Figure 11), instead concentrating on acoustic-prosodic cues, which are not among those that SEAME speakers use to significantly entrain at the turn level (Section 5.1), nor do these correlate with CSW features (Table 14; Appendix A.13). This mirrors the conversation-level for SEAME, where the best-performing models similarly bypass relevant CSW style features in favor of acoustic-prosodic signal. For MaSaC, we observe a divergence by model type. Text-embedding models show the same concentration effect noted for conversation-level BM, assigning the greatest importance to a narrow subset of pitch and intensity cues while making limited use of the broader acoustic-prosodic profile characterizing human entraining behavior (Figure 12). Speech-embedding models, in contrast, distribute importance across the full acoustic-prosodic feature set, but also assign notably high importance to CSW style features, which do not define the positive class for this task; turn pairs are labeled by acoustic-prosodic proximity, not CSW style entrainment. We propose an explanation for this cross-feature-set pattern in Appendix A.11.

These **turn-level** results largely reinforce those from the conversation level in their lack of sufficient evidence for model classification decisions being systematically guided by the features most salient to human entrainment. While speech-grounded representations offer partial improvement in feature coverage, these exhibit diffuse rather than targeted reliance on human-grounded turn-level entrainment. We also note that model performance is generally lower at the turn level than at the conversation level across corpora, a pattern consistent with the finer-grained and more variable nature of turn-level conversational dynamics, which may reduce the separability of entraining and non-entraining instances. The relative advantage of contextual embedding-based models at the turn level is consistent with this interpretation: contextual representations may provide useful inductive bias when local sequential context is more diagnostic than aggregate conversational statistics.

Feature-level ablations. Since feature importance estimates are classifier-specific and derived from only a subset of models, we expand our analysis to all models via feature-level ablations. At the **conversation level** of BM, ablating each feature produces a significant *drop* in accuracy relative to the baseline, per paired *t*-tests; all features thus provide some useful signal (Figure 4). However, most effect sizes are small, and no relevant CSW style features rank among the top contributors to accuracy drops. SEAME ablations are similar and generally produce negligible change in performance (Figure 13). While CSW style features account for 3 of the top-5 contributors to performance, more are linked to accuracy *increases* upon ablation, suggesting that they introduce noise rather than useful signal. This effect is more pronounced in MaSaC: only one lexical feature ablation results in an accuracy drop, and none rank among the top contribu-

[11] No BM model exceeds the 65% threshold, so we exclude BM from turn-level analysis; discussion in Appendix A.14.

tors to performance across models (Figure 14).

We find a larger divergence between model and human entraining behavior at the **turn level**. For SEAME and BM, ablations produce near-zero changes in accuracy (Figures 15 and 17), suggesting that models rely on diffuse (or redundant) cues rather than any identifiable set of uniquely important entrainment features to make decisions. MaSaC models are only slightly more informative: while some salient acoustic-prosodic features contribute to performance, effect sizes remain small and importance is narrowly concentrated, with only speaking rate as a partial exception (Figure 16). These results suggest that, as conversational granularity increases, entrainment becomes harder for models to capture in a way that reflects statistically observed human entrainment patterns.

Discussion. Across language pairs and evaluation methods, we observe two recurring patterns in model behavior. At the conversation-level, models achieve reasonable performance while relying on features different from those identified as statistically salient in human entrainment. At the turn-level, feature ablations produce near-zero changes, suggesting reliance on diffuse signals rather than a coherent, human-grounded feature set. In response to **RQ2**, our entrainment detection models generally prioritize features other than those most salient in naturalistic CSW, and, at finer-grained levels of interaction, either distribute importance too broadly across feature sets or place limited emphasis on any interpretable, human-relevant features. Similar to Blevins et al. (2026), we speculate that this may reflect fundamental differences between underlying human and model entrainment mechanisms, as well as the inherent difficulty of explicitly modeling a *subconscious* human process. Other explanations include a need for alternative representations of entraining code-switched speech, or the possibility of our data and models capturing latent aspects of entrainment not covered by the formal theories used to operationalize our approach. Further investigation of these factors, particularly on how models might be encouraged to attend to and robustly represent human-grounded entrainment features, is an important direction for future work.

As noted in prior work (Havaldar et al., 2023), multilingual models generally classify stylistic aspects of language well, but struggle to generate them appropriately (Briakou et al., 2021; Srinivasan and Choi, 2022). Our findings raise the possibility that generating code-switched speech with human-like entrainment patterns may pose a similar challenge, though results from classifier-based settings may not directly transfer to generative models. One potential concern is that classifiers, which show promise for steering and controlling generative models' stylistic outputs (Dathathri et al., 2020; Scalena et al., 2026; Dinan et al., 2020; Liang et al., 2024), could propagate feature-level misalignment when used to guide generation. So, strong performance in detecting entrainment need not imply that models capture the underlying phenomenon in a way that would support naturalistic generation. Recent findings in monolingual settings reinforce this concern: Blevins et al. (2026) identified mismatches in the linguistic form and motivations of entrainment between humans and language models in spontaneous text production, while Hua et al. (2025) showed that post-training fails to close the gap between LLMs and humans in task-oriented stylistic adaptation. Taken together, these results suggest that it may be valuable to directly investigate whether generative models face similar challenges in the more demanding context of code-switched speech, and whether interventions (e.g., prompting or training strategies) can encourage more human-like entrainment behavior.

6 Conclusion

We explore cross-lingual entrainment behavior in spoken CSW, studying both human speakers and classification models across language pairs spanning diverse typological families and cultural contexts. We find that many established patterns of code-switched entrainment extend to new language pairs, while observed cross-corpus variation may reflect relevant multilingual norms. We also find that the predictive cues used by our entrainment classifiers are not well aligned with statistically salient features observed in human speech under our operationalization – a finding we surface through feature importance attribution methods that are directly tied to model behavior and enable human-interpretable explanations. We hope that these new insights, together with the stylistically annotated data we release, will inform the development of multilingual language models capable of generating naturalistic code-switched speech that embodies the human-like stylistic patterns needed for successful communication and social cohesion.

Limitations

Our work is the first to conduct a cross-lingual analysis of entrainment in code-switched speech. While we study multiple language pairs covering distinct language families and cultural contexts, all involve English as one of the languages. This design choice is partly motivated by data availability constraints, but necessarily leaves much of the world's rich and varied CSW landscape unaddressed. Our claims about cross-lingually consistent entrainment patterns are therefore limited to English-involved CSW, and may not extend to bilingual settings excluding English.

We also acknowledge that our corpora each represent a specific community of CSW speakers (Miami-based Hispanic communities, Singaporean and Malaysian Mandarin-speakers, one particular Hindi-language television show), in addition to differing in size, genre, speaker demographics, and overall CSW rate (Table 1). Fully disentangling language-pair-specific entrainment patterns from the corpus-level differences we find would require new, genre- and size-matched corpora across all three language pairs, which do not currently exist; collecting or identifying such data, particularly a naturally-occurring Hindi-English counterpart to MaSaC, is a concrete direction we intend to pursue in future work. Since models are highly context-sensitive and entrainment norms may vary within a language pair across regions, social contexts, and speaker demographics, we eagerly anticipate future work that tests the robustness of our findings on new data. Given limited access to suitable open-source code-switched speech corpora, our work is a first step towards expanding cross-lingual understanding of conversational dynamics in CSW.

Separately, we acknowledge a potential data preprocessing limitation specific to MaSaC. The Romanized representation of Hindi in its speech transcripts may have posed a challenge for the LID tool used to obtain token-level language tags, as language-ambiguous words can be difficult to label automatically without contextual or pronunciation information. We manually reviewed and corrected LID labels prior to inferring utterance-level language assignments, reading and listening to *every* MaSaC transcript in full, but some residual errors could remain. If any, these would most likely result in underestimating the number of code-switched utterances, and their corresponding style annotations, by miscategorizing them as monolingual. This is a particularly acute concern for MaSaC given its limited size. This limitation reflects a broader challenge of computational work on Hindi-English CSW: the scarcity of openly accessible, genre-matched spoken corpora means that errors in automatic pre-processing carry greater consequence, as few alternative data sources exist.

Finally, we acknowledge three related limitations in our model evaluation methodology. First, our binary classification framing is necessarily a simplification of what is a continuous and multidimensional linguistic phenomenon. The thresholds used to construct entraining and non-entraining classes directly shape what models are able to learn, and different operationalizations of entrainment could yield different classification boundaries and, consequently, different feature importance profiles. Since our entrainment labels are derived from the same analytical framework used to characterize human behavior, the resulting classification task necessarily reflects that operationalization of entrainment rather than the full underlying phenomenon.

Second, we operationalize features salient to human entraining behavior as those exhibiting statistically significant entrainment patterns (Section 5.1). We note that statistical significance in corpus-level analysis is an imperfect proxy for behavioral or cognitive salience: a feature may show significant entrainment at the population level without being a primary driver of individual speakers' adaptation, and vice versa. Our comparison between human and model feature reliance should therefore be understood as a first approximation, i.e., a way of asking whether models are at least sensitive to the same dimensions of speech that correlate with entrainment, rather than a direct test of the underlying mechanism. Future work using behavioral data, e.g., listener judgments or physiological measures of rapport, could provide a more direct index of which features matter more to human speakers.

Third, Shapley feature importance estimates, which we rely on for a subset of models in RQ2, are known to be sensitive to feature correlations; since our acoustic-prosodic features are weakly to moderately correlated with one another (Tables 13 and 14), some instability in their attributed importances is possible. Both of these factors introduce uncertainty into the comparison between human and model entrainment behavior, and future work would benefit from exploring alternative operationalizations and importance attribution methods to assess the robustness of our conclusions.

Ethical Considerations

This study uses only secondary data and does not involve human experiments. All speakers in the Bangor Miami and SEAME corpora consented to data collection and sharing, and both datasets were released in de-identified form by their original authors. The MaSaC corpus consists of fictional dialogue performed by professional actors and was used with permission from its creators.

While our work is primarily exploratory, it has potential implications for multilingual conversational systems. Modeling conversational entrainment in code-switched settings may support applications such as adaptive dialogue systems or language learning tools, but could also be misused in contexts where stylistic adaptation is intended to increase trust or rapport, including persuasive or deceptive scenarios such as targeted scams or disinformation. Care should therefore be taken in the deployment of downstream applications.

Our findings highlight how entrainment patterns vary across language pairs and contexts, and systems trained on such data may reflect these underlying norms. This may lead to uneven performance or misalignment when applied to speakers whose communication style differs from those represented in these datasets, and raises broader questions about whether entrainment should be treated as descriptive behavior or a normative benchmark.

We also find that models detect entrainment without relying on the same features as humans, suggesting that decisions may be driven by indirect or spurious signals. This raises concerns about robustness and reliability, as systems that appear to capture human-like interaction patterns may behave unpredictably in new settings.

To mitigate these risks, we encourage future work that prioritizes evaluation across diverse multilingual contexts. We also caution against treating entrainment as an isolated normative measure of communicative competence, and recommend safeguards such as usage restrictions, misuse monitoring, and human oversight in sensitive applications.

Acknowledgments

We thank Nicholas Deas for several helpful discussions and feedback, and Chhavi Dixit for verifying Hindi-English LID and CSW annotations. DB is supported by the National Science Foundation under Grant IIS 2418307.

References

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What Code-Switching Strategies are Effective in Dialog Systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 254–264, New York, New York. Association for Computational Linguistics.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 2000. The LIDES Coding Manual: A document for preparing and analyzing language interaction data version 1.1—July, 1999. *International Journal of Bilingualism*, 4(2):131–270.

Anshul Bawa, Monojit Choudhury, and Kalika Bali. 2018. Accommodation of Conversational Code-Choice. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 82–91, Melbourne, Australia. Association for Computational Linguistics.

M. Bedi, S. Kumar, M. Akhtar, and T. Chakraborty. 2023. Multi-modal Sarcasm Detection and Humor Classification in Code-mixed Conversations. *IEEE Transactions on Affective Computing*.

Debasmita Bhattacharya, Siying Ding, Alayna Nguyen, and Julia Hirschberg. 2024. Measuring Entrainment in Spontaneous Code-switched Speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2865–2876, Mexico City, Mexico. Association for Computational Linguistics.

Debasmita Bhattacharya, Anxin Yi, Siying Ding, and Julia Hirschberg. 2025. Code-switching in Context: Investigating the Role of Discourse Topic in Bilingual Speech Production. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 64–80, Suzhou, China. Association for Computational Linguistics.

Terra Blevins, Susanne Schmalwieser, and Benjamin Roth. 2026. Do language models accommodate their users? A study of linguistic convergence. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–807, Rabat, Morocco. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A Benchmark for Multilingual Formality Style Transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Judee K. Burgoon, David B. Buller, Jerold L. Hale, and Mark A. de Turck. 2006. Relational Messages Associated with Nonverbal Behaviors. *Human Communication Research*, 10(3):351–378.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: the perception-behavior link and social interaction. *Journal of personality and social psychology*, 76 6:893–910.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei. 2021. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *Preprint*, arXiv:2010.12821.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Huangliang Dai, Shixun Wu, Jiajun Huang, Zizhe Jian, Yue Zhu, Haiyang Hu, and Zizhong Chen. 2025. FT-Transformer: Resilient and Reliable Transformer with End-to-End Fault Tolerant Attention. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '25, page 1085–1098, New York, NY, USA. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my Words! Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th International Conference on World Wide Web*. ACM.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.

Margaret Deuchar. 2011. Miami corpus: Preliminary documentation - bangortalk.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Preprint*, arXiv:1810.04805.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Ethnologue. 2026. What is the most spoken language? `https://www.ethnologue.com/insights/most-spoken-language/`. Accessed: 2026-04-06.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. *Accommodation theory: Communication, context, and consequence*, page 1–68. Studies in Emotion and Social Interaction. Cambridge University Press.

Agustín Gravano, Štefan Beňuš, Rivka Levitan, and Julia Hirschberg. 2014. Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 578–583.

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple Tools for Exploring Variation in Code-switching for Linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023. Comparing Styles across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.

Yilun Hua, Evan Wang, and Yoav Artzi. 2025. Post-training for Efficient Communication via Convention Formation. In *Proceedings of the Conference on Language Modeling*.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. 2020. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *Preprint*, arXiv:2012.06678.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Jay Kejriwal and Štefan Beňuš. 2023. Relationship between auditory and semantic entrainment using Deep Neural Networks (DNN). In *Interspeech 2023*, pages 2623–2627.

Gerrit Jan Kootstra, Janet G. van Hell, and Ton Dijkstra. 2010. Syntactic alignment and shared word order in code-switched sentence production: Evidence from

bilingual monologue and dialogue. *Journal of Memory and Language*, 63(2):210–231.

Rivka Levitan, Štefan Beňuš, Agustín Gravano, and Julia Hirschberg. 2015. Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–334, Prague, Czech Republic. Association for Computational Linguistics.

Rivka Levitan, Agustin Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-Prosodic Entrainment and Social Behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada. Association for Computational Linguistics.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. Interspeech 2011*, pages 3081–3084.

Xun Liang, Hanyu Wang, Shichao Song, Mengting Hu, Xunzhi Wang, Zhiyu Li, Feiyu Xiong, and Bo Tang. 2024. Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5797–5814, Bangkok, Thailand. Association for Computational Linguistics.

Qi Lin, Weikai Xu, Lisi Chen, and Bin Dai. 2025. V-VAE: A Variational Auto Encoding Framework Towards Fine-Grained Control over Human-Like Chat. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29693–29706, Suzhou, China. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Chng, and Haizhou Li. 2010. Mandarin–English code-switching speech corpus in South-East Asia: SEAME. In *Language Resources and Evaluation*, volume 49, pages 1986–1989.

Richard E. Maurer and Jeffrey H. Tindall. 1983. Effect of postural congruence on client's perception of counselor empathy. *Journal of Counseling Psychology*, 30(2):158–163.

Pieter Muysken. 2000. *Bilingual speech: a typology of code-mixing*. Cambridge University Press.

Danial Namazifard and Lukas Galke Poech. 2025. Isolating Culture Neurons in Multilingual Large Language Models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 768–785, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Ani Nenkova, Agustin Gravano, and Julia Hirschberg. 2008. High Frequency Word Entrainment in Spoken Dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172, Columbus, Ohio. Association for Computational Linguistics.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding Linguistic Accommodation in Code-Switched Human-Machine Dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577, Online. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics*, 18:581–618.

David Reitter and Johanna D. Moore. 2007. Predicting Success in Dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Scalena, Gabriele Sarti, Arianna Bisazza, Elisabetta Fersini, and Malvina Nissim. 2026. Steering Large Language Models for Machine Translation Personalization. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4681–4701, Rabat, Morocco. Association for Computational Linguistics.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The Role of Cognate Words, POS Tags and Entrainment in Code-Switching. In *Interspeech 2018*, pages 1938–1942.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andreas Weise, Vered Silber-Varod, Anat Lerner, Julia Hirschberg, and Rivka Levitan. 2021. "Talk to me with left, right, and angles": Lexical entrainment in spoken Hebrew dialogue. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 292–299, Online. Association for Computational Linguistics.

Zhihua Xia, Rivka Levitan, and Julia Hirschberg. 2014. Prosodic Entrainment in Mandarin Chinese and English: A Cross-Linguistic Comparison. In *Speech Prosody 2014*, pages 65–69.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*, pages 3939–3943.

A Appendix

A.1 Computing code-switched entrainment: definitions & full methodological details.

We reproduce the method of Bhattacharya et al. (2024) here for completeness. Note that *proximity* refers to absolute similarity in speech, *convergence* refers to how much interlocutors' speech becomes more similar over time, and *synchrony* refers to relative coordination between interlocutors' speech.

We compute **lexical entrainment** over word classes using Eq. 1 (Nenkova et al., 2008), which calculates entrainment between speakers S_A and S_B on a given word class W as the negated sum of absolute differences in each word $w \in W$'s usage rate between S_A and S_B.

$$entr(S_A, S_B) = -\sum_{w \in W} |\frac{count_{S_A}(w)}{ALL_{S_A}} - \frac{count_{S_B}(w)}{ALL_{S_B}}| \tag{1}$$

We also compute **lexical entrainment** on overall language use in each corpus using Kneser-Essen-Ney smoothed trigram language models trained on each speaker's transcripts via the KenLM toolkit. For each conversation, we evaluate the model trained on S_A's transcripts through its perplexity on S_B's speech transcripts. We do the opposite to evaluate the S_B-trained model. Resulting negated perplexity values are unidirectional entrainment scores; lower perplexity reflects greater entrainment to the interlocutor.

We study entrainment over **acoustic-prosodic** and **CSW style** features via turn- and conversation-level proximity, convergence, and synchrony.

At the **turn-level**, we measure **proximity** by computing a *partner difference* (Eq. 2) and *other difference* (Eq. 3) for each target speaker turn, such that $turn_p$ is adjacent to the target turn and uttered by the target turn speaker's conversation partner, and $turn_i$ is uttered by the target turn speaker's conversation partner but is not adjacent to the target turn, over ten random turns. We compare *partner differences* and *other differences* for a given feature with a paired *t*-test, and infer proximity when partner differences from the prior speaker turn are smaller than differences from other speaker turns.

$$difference_{partner} = |turn_t - turn_p| \tag{2}$$

$$difference_{other} = \frac{\sum_{i=1}^{10} |turn_t - turn_i|}{10} \tag{3}$$

For turn-level **convergence**, we compute the Pearson correlation coefficient between the abso-

lute feature difference between adjacent turns and the turn number. We infer strong, moderate, and weak convergence from $r \geq 0.7$, $0.5 \leq r < 0.7$, and $0 < r < 0.5$, respectively. We similarly infer strong, moderate, and weak divergence from $r \leq -0.7$, $-0.7 < r \leq -0.5$, and $-0.5 < r < 0$. We similarly compute turn-level **synchrony**, using Pearson correlation between feature values of adjacent turns from different speakers, testing for significance with a two-sided t-test. We infer the degree of synchrony or asynchrony using the same ranges of r-values as for turn-level convergence.

At the **conversation-level**, we calculate **proximity** using paired t-tests to compare two sets of differences for each speaker: *partner differences*, i.e., the difference between the speaker's value for a given feature and that of their partner, versus *other differences*, i.e., the mean of differences between the speaker's value and the values of each speaker who was not their interlocutor. We infer proximity when partner differences are smaller than other differences. We compute **convergence** by halving each conversation and comparing the difference in the two speakers' mean values for a given feature in the first half to that in the second half using a paired t-test. We infer convergence when differences in the second half are significantly smaller.

A.2 Datasets' license information.

The SEAME corpus may be used for research per the LDC User Agreement for Non-Members. The MaSaC corpus is released by the authors in a public Github repository at https://github.com/LCS2-IIITD/MSH-COMICS. The BM corpus is made available under the GNU General Public License version 3 or later. Our use of these corpora is consistent with their intended use cases in research.

A.3 Affirmative cues and fillers.

The members of the *affirmative cues* word class for SEAME are: 'yeah', 'right', 'ok', 'mmhmm', 'ya', 'yes', 'huh', 'alright', 'yup', 'hah', 'yep', 'yah', '对', '嗯', '啊哈', '啊哈', '哦', '哈', '是', '当然', '可以', '行'. We derive this list through a combination of SEAME data exploration and reference to the following articles:

- https://direct.mit.edu/coli/article/38/1/1/2144/Affirmative-Cue-Words-in-Task-Oriented-Dialogue
- https://lcchineseschool.com/say-yes-in-chinese-quick-guide-to-affirmation/

The members of the *fillers* word class for SEAME are: 'um', 'ah', 'uh', 'eh', 'er', 'hmm', 'mm', 'mmm', 'umm', 'ar', 'hm', '啊', '呃', '呃', '啊', '呃', '嗯', '嗯', '哼'. This list is derived similarly to the one for affirmative cues in SEAME, with additional reference to the article at https://www.chineseclass101.com/blog/2021/09/09/chinese-filler-words/.

The members of the *affirmative cues* word class for MaSaC are: 'haan', 'arey', 'aray', 'arre', 'oh', 'ooh', 'ohh', 'achchha', 'achcha', 'ya', 'yah', 'vah', 'wah', 'waah', 'ok', 'okay', 'ha', 'bol', 'right', 'oho', 'oo', 'huh', 'yes', 'alright', 'hein', 'really', 'jee', 'waa'.

The members of the *fillers* word class for MaSaC are: 'voh', 'woh', 'wooh', 'aa', 'aah', 'ahh' 'aaah', 'na', 'umm', 'hmm', 'ae', 'achchha', 'achcha', 'pch', 'aaa', 'mmm', 'o', 'im', 'oye', 'haww', 'uhhh', 'uhh', 'm', 'ahem', 'mmmmm', 'uh'.

Both word classes for MaSaC are derived through data exploration.

A.4 Pre-trained models' license information.

Our use of these models is consistent with their intended use cases in research.

- *TabTransformer*: Creative Commons CC0 1.0 Universal.
- *FT-Transformer*: Creative Commons Attribution 4.0 International.
- *XLM-R*: MIT license.
- *mBERT*: Apache License 2.0
- *RemBERT*: Apache License 2.0.
- *WaVLM*: Creative Commons Attribution-ShareAlike 3.0 Unported
- *mHuBERT*: Creative Commons Attribution Non Commercial Share Alike 4.0

A.5 Hyperparameters for grid search for each model.

Within Table 4, all models from *TabTransformer* onwards encode input data with the stated architecture, which is then fused with an MLP and softmax function for predictions. We encode transcript text alongside tabular data as input in the *XLM-R*, *mBERT*, and *RemBERT* models, splitting transcripts into 512-token chunks using the corresponding tokenizer for each model. The parameters

of each of these pre-trained encoders are frozen so that only the numerical *MLP*, pooling, fusion, and classifier train, and feature importance attribution is computed only over tabular features. We encode audio recordings alongside tabular data as input in the *WavLM* and *mHuBERT* models, chunking audio in 5-second segments and using averaged pooled acoustic-prosodic embeddings for each. We similarly freeze the parameters of these encoders during training. For each corpus, the final multimodal model combining text and speech embeddings with tabular inputs is determined based on the best-performing text and speech embedding methods (*mBERT* + *mHuBERT* for BM, *RemBERT* + *WavLM* for SEAME, *mBERT* + *WavLM* for MaSaC). All models use an Adam optimizer and binary cross-entropy loss over 10 epochs and batch size of 32. Note that we adapt the *TabTransformer* architecture since our data lacks categorical features after pre-processing for evaluation.

Conversation-level experiments collectively take about 130 hours to run on a Google Colab A100 GPU. Turn-level experiments collectively take about 460 hours to run on an A5500 GPU.

A.6 Evaluating model entrainment behavior: data construction details.

At the **conversation-level**, we identify conversations that exhibit significant proximity on all CSW style metrics (for BM and SEAME) or lexical word classes (for MaSaC); positive (entraining) and negative (non-entraining) classes per corpus are roughly balanced and directly informed by the statistical results in Section 5.1. We structure model inputs as a feature vector for each conversation covering speaker-separated lexical, acoustic-prosodic, and CSW style attributes, as defined in Section 4.1; these hand-selected features effectively approximate an entrainment lexicon. We use the conversation-level mean of continuous features and convert utterance-level non-continuous features to binary flags (similar to one-hot encodings) before computing their conversation-level mean within each feature vector. This results in a *tabular/numerical* representation, which we concatenate with model-specific embeddings of conversation transcripts (text embeddings) or audio recordings (speech embeddings).

At the **turn-level**, we apply Section 5.1's statistical results to identify pairs of consecutive turns that exhibit significant proximity on a majority of CSW style metrics (for BM and SEAME) or acoustic-prosodic features (for MaSaC), handling class imbalance with oversampling. Turn-level feature vectors have the same speaker-separated structure and treatment of continuous vs. non-continuous features for associated turn pairs as do conversation-level ones; since a single speaker's turn may comprise multiple consecutive utterances, we use the mean of utterance-level values within that turn in the corresponding feature vector.

We acknowledge one apparent tension in our design: at both levels of conversational granularity, the positive class is defined by significant entrainment on a particular feature set (e.g., significant proximity on CSW style metrics for conversation-level BM), and we evaluate whether models assign high importance to that same feature set. In principle, a model that perfectly learns to perform the classification task as defined should rely on those features. Our findings in Section 5.2 on imperfect alignment, where classifiers perform reasonably while bypassing the defining feature set, is therefore more striking than it may first appear; it suggests that classifiers may find alternative shortcuts rather than solving the problem as defined. We note two important caveats to this interpretation, however. First, classifiers could, in principle, achieve high accuracy via unmeasured causal features outside our theoretically-motivated feature sets (e.g., latent signal captured by text or speech embeddings), or via correlated proxy features within our feature space; on the latter, we find near-zero correlations between our lexical, acoustic-prosodic, and CSW style feature sets across all corpora and granularities (Appendix A.13, Tables 13-14), suggesting that this alignment gap is not simply an artifact of models exploiting a correlated proxy elsewhere in the same feature space. Second, this design means that we cannot use high feature-importance alignment as straightforward validation, as such alignment could reflect class-definition circularity rather than genuine behavioral correspondence. We therefore treat our findings primarily as evidence of imperfect alignment (the negative result) rather than a strong positive claim about what more obviously aligned model behavior would mean.

A.7 Cross-lingual entrainment behaviors: effect sizes.

Tables 5-8 report paired Cohen's d for turn- and conversation-level proximity and conversation-level convergence comparisons and lexical entrainment (i.e., proximity) comparisons. Tables 9 and

10 report aggregate mean and median Pearson r for turn-level convergence and synchrony, following standard practice for correlation-based measures. These results generally support our main claims in Section 5.1 and offer additional evidence beyond what significance testing alone can show. We highlight a few examples below.

First, significant turn-level CSW style proximity effects in SEAME are consistently weak in magnitude ($|d| \leq 0.05$), while the corresponding conversation-level proximity effects are medium-to-large ($0.55 \leq |d| \leq 1.58$), aligning with greater conversation-level CSW style entrainment within this corpus. Second, MaSaC's null lexical results for affirmative cues and fillers correspond to effect sizes close to zero ($|d| = 0.01$ and $|d| = 0.02$, respectively), supporting the interpretation that these reflect real absences of entrainment. Third, acoustic-prosodic entrainment effect sizes are uniformly small across every tested feature in SEAME (turn-level $|d| < 0.04$ throughout), providing quantitative support for the characterization of Mandarin-English acoustic-prosodic entrainment as limited, in addition to which individual features crossed the significance threshold.

A.8 Cross-lingual entrainment behaviors: Bonferroni-corrected significance results.

Table 11 summarizes cross-corpus Bonferroni-corrected pairwise Fisher's exact comparisons; Table 12 summarizes within-corpus Bonferroni-corrected significance for the same measures. See Section 5.1 for relevant interpretation.

A.9 Cross-lingual entrainment behaviors: additional results.

Entrainment over lexical aspects of code-switched speech. In the SEAME corpus, when excluding OOV words, 88% of conversations exhibit within-conversation entrainment ($p < 0.05$); 82% involve both partners entraining to each other, while in the remaining 6% only one speaker entrains to their partner.

When excluding OOV words in the MaSaC corpus, 72% of conversations exhibit entrainment ($p < 0.05$); 58% involve both interlocutors entraining to each other, while the remaining 14% involve only unidirectional entrainment.

See Figures 5 through 7 for visualizations of turn-level proximity, convergence, and synchrony in SEAME.

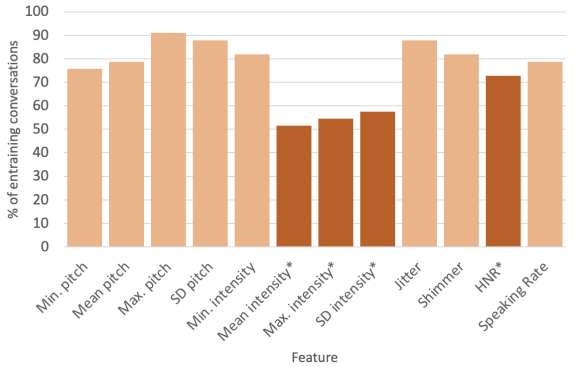

Figure 5: Proximity at the turn-level in SEAME. * and darker bars indicate significant features.

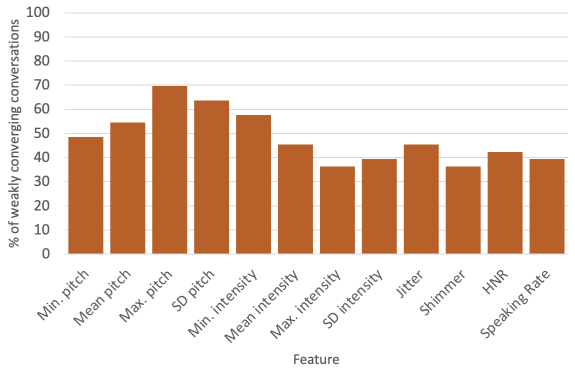

Figure 6: Convergence at the turn-level in SEAME. All conversations converge *weakly*. All non-converging conversations diverge *weakly*. All acoustic-prosodic features are significant.

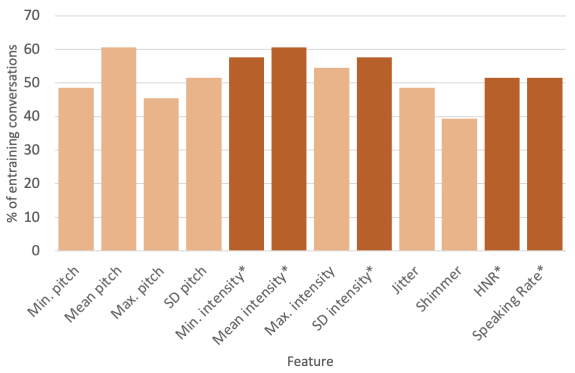

Figure 7: Synchrony at the turn-level in SEAME. All conversations are *weakly* synchronous. All acoustic-prosodic features are significant. All non-synchronous conversations are *weakly* asynchronous.

A.10 Cross-lingual entrainment behaviors: comparison to prior monolingual and CSW entrainment findings.

While exact quantitative comparisons are limited by differences in corpora, modality, and entrain-

ment operationalization across studies, we offer several qualitative points of comparison to prior work below.

Bawa et al. (2018) study Hindi-English CSW accommodation in a Hindi movie corpus using a distinct entrainment metric from our own. Their finding that a high base rate of CSW corresponds to reduced accommodation is consistent with our hypothesis (Section 5.1) that MaSaC's higher M- and I-index values may create a ceiling on stylistic entrainment, since there is less room for one speaker's style to influence the other's when both already code-switch prevalently.

Levitan et al. (2015) study monolingual acoustic-prosodic entrainment in Spanish, English, and Mandarin. Their finding of significant monolingual English turn-level proximity on mean and maximum intensity aligns directionally with our finding that SEAME's limited turn-level proximity is nonetheless significant specifically on intensity features (Figure 5), suggesting that the English side of the language pair may drive this signal. Their slight yet largely negative Mandarin proximity and strongly negative Mandarin pitch synchrony are consistent with our characterization of Mandarin-English as our weakest acoustic-prosodic pair, lending external support to our explanation involving Mandarin's distinct tonal character.

Xia et al. (2014) report that monolingual Mandarin and English each individually show strong conversation-level proximity on intensity, pitch maxima, and speaking rate, with English additionally showing conversation-level convergence. Our SEAME findings show the opposite pattern: an absence of conversation-level acoustic-prosodic proximity and convergence, mirroring Spanish-English. This divergence suggests that the CSW context itself, rather than either language alone, may suppress conversation-level entrainment that is otherwise present monolingually in both languages.

Parekh et al. (2020) study lexical and CSW entrainment in written human-machine Hindi-English dialogue. Their finding of strong lexical entrainment in both English and Hindi, using the same method as Nenkova et al. (2008) that we also adopt, is broadly consistent with our MaSaC lexical findings. We also note a partial parallel between their reported entrainment ranking (CSW presence, then insertional CSW, then language choice, then alternational CSW) and our finding that MaSaC's turn-level convergence is strongest for CSW presence and insertional strategy; however, their study concerns accommodation to a machine interlocutor in text rather than human-human speech, likely reflecting a different underlying social dynamic.

A.11 Classifier models' entrainment behavior: models' feature importance values.

Figure 8 shows the full set of feature importance values across the best-performing BM models at the conversation-level, including many near-zero values. Figures 9 and 11 visualize conversation-level and turn-level features importance values, respectively, for the SEAME corpus. Figures 10 and 12 visualize the same in MaSaC.

With respect to the degree of importance assigned to CSW style features in the best-performing turn-level speech-embedding models for MaSaC (Figure 12), this could partly reflect very weak turn-level cross-feature-set correlations (Table 14; Appendix A.13). Although the mean correlation between CSW metrics and acoustic-prosodic features is near zero ($r = 0.015$), 35.4% of feature pairs reach statistical significance; this is likely an artifact of the large number of turn pairs inflating test sensitivity rather than evidence of a meaningful relationship. However, even such weak statistical regularities can be sufficient for models to exploit CSW features as incidental predictors of acoustic-prosodic entrainment labels. We highlight that this is importantly different from our interest in models learning a theoretically-motivated entrainment signal.

A.12 Classifier models' entrainment behavior: feature-level ablations.

Figures 13 and 15 visualize feature ablations on the SEAME corpus at the conversation-level and the turn-level, respectively. Figures 14 and 16 visualize the same on the MaSaC corpus.

A.13 Within- and between-feature-set correlations for each corpus.

At the **conversation-level**, cross-feature-set correlations are near zero across all modality pairs and corpora, with mean Fisher-z-transformed correlations close to zero and very low proportions of significant relationships where $p < 0.05$ (Table 13). This suggests that CSW, lexical, and acoustic-prosodic feature spaces capture largely non-overlapping dimensions of variation, supporting their complementarity rather than redundancy. In contrast, within-set correlations are substantially higher, particularly for CSW and acoustic features

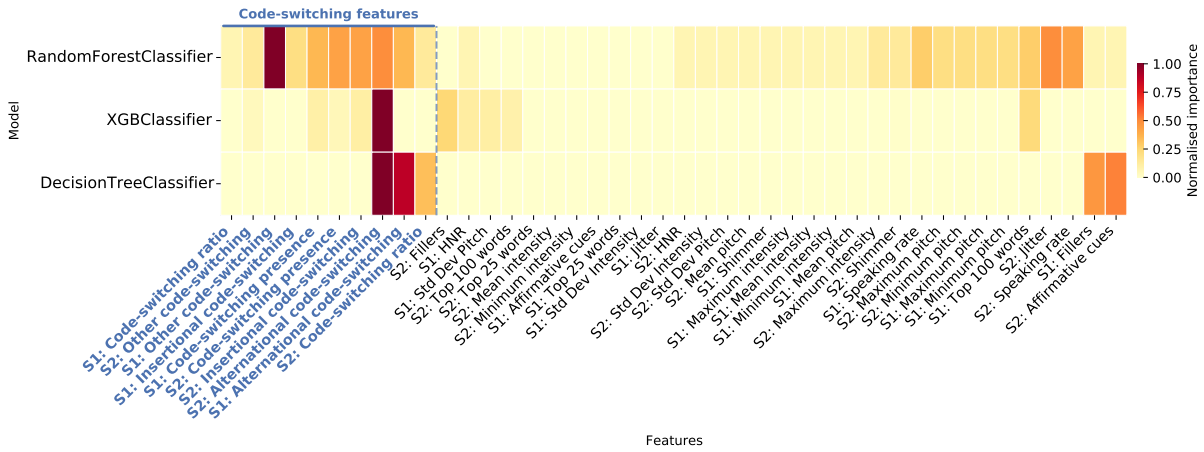

Figure 8: Feature importance values across best-performing BM models at the conversation-level. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable.

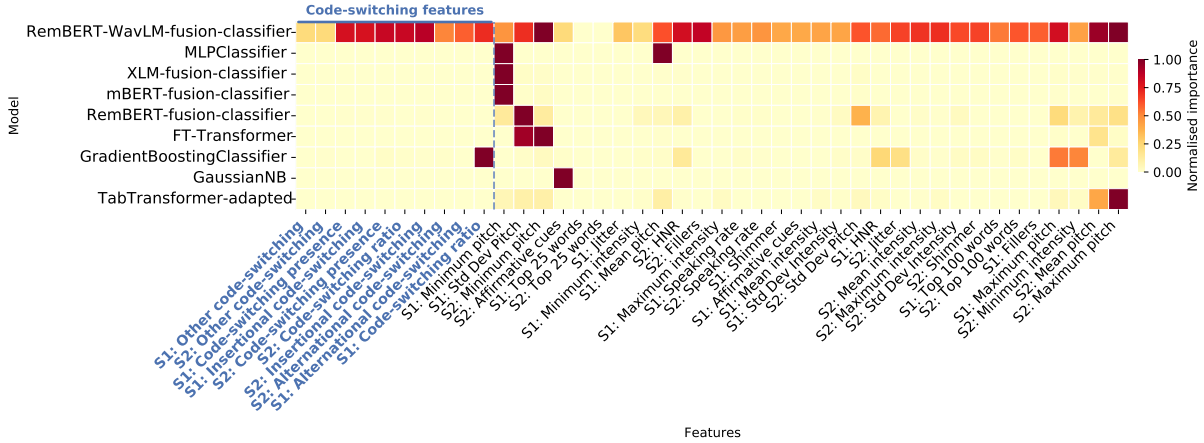

Figure 9: Feature importance values across best-performing SEAME models at the conversation-level. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable. All models except *FT-Transformer* achieve near-perfect test-time accuracy; *FT-Transformer achieves 75%* test-time accuracy. Individual model hyperparameter settings are as follows: *MLPClassifier* has activation: tanh, alpha: 1e-05, hidden layer sizes: 250, solver: lbfgs; *GradientBoostingClassifier* has criterion: friedman mse, lr: 0.01, loss: exponential, n estimators: 50; *TabTransformer-adapted* has lr: 0.001; *XLM-fusion* has lr: 0.001; *mBERT-fusion* has lr: 0.001; *RemBERT-fusion* has lr: 0.001; *RemBERT + WavLM fusion* has lr: 0.01; *FT-Transformer* has lr: 0.001.

in BM and SEAME, indicating weak to moderate internal structure but cross-modally distinct feature spaces in those corpora. Conversation-level patterns of cross-feature-set correlations are consistent at the **turn-level** (Table 14), where even within-set correlations are weak.

A.14 Classifier models' entrainment behavior: excluding BM from turn-level analysis of models' feature importance values.

None of the BM models exceed the 65% test-time accuracy threshold at the turn level, so we exclude the corpus from turn-level analysis. This poor performance may be due to class imbalance artifacts and/or the relative size of the corpus, especially given BM's lower overall CSW rate (5%, Table 1), which could reduce the variance in CSW style features at the turn level and may make the classification boundary harder to learn. We acknowledge that the exclusion of the corpus is an unfortunate limitation, as the absence of turn-level BM results leaves the cross-lingual picture at finer conversational granularity incomplete.

A.15 Classifier models' entrainment behavior: conversation-level analysis using CSW style features in MaSaC.

Our primary classification tasks for each corpus use the empirically-determined feature set that best discriminates entraining from non-entraining con-

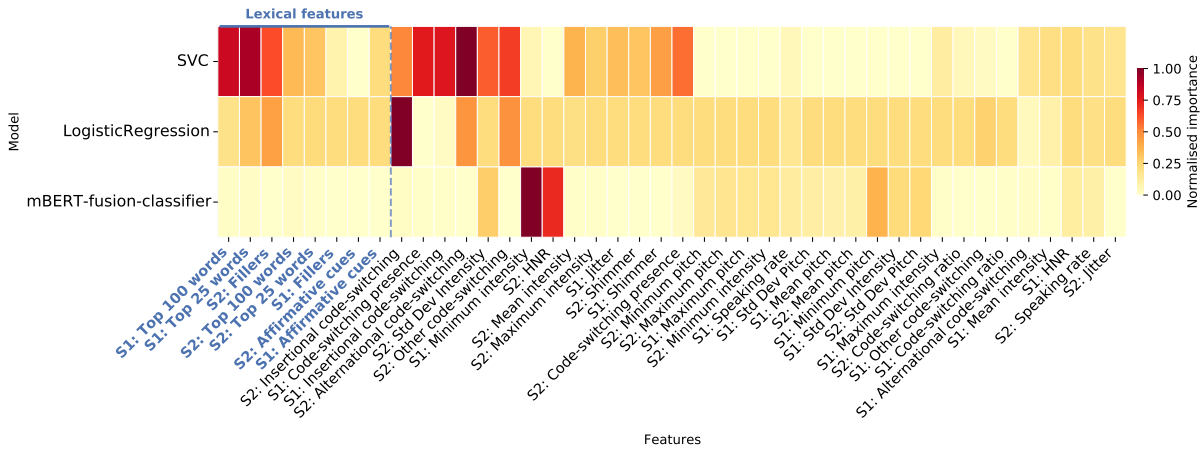

Figure 10: Feature importance values across best-performing MaSaC models at the conversation-level. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable. *LogisticRegression* with hyperparameter C: 1 achieves 71% test-time accuracy, *mBERT-fusion* with hyperparameter lr: 0.001 achieves 68%, and *SVC* with hyperparameter C: 1.75 achieves 65%.

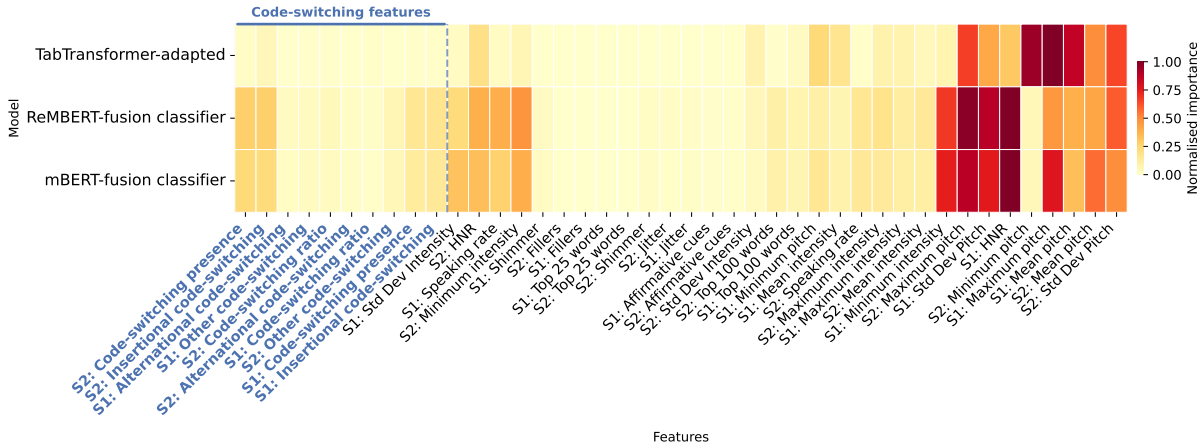

Figure 11: Feature importance values across best-performing SEAME models at the turn-level. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable. *mBERT-fusion* with hyperparameter lr: 1e-4 achieves 70% test-time accuracy, *RemBERT-fusion* with hyperparameter lr: 1e-4 achieves 67%, and *TabTransformer-adapted* with hyperparameter lr: 1e-3 achieves 65%.

versations (Footnote 6); we additionally explore classifier behavior when the feature set in use is not naturally suited to this purpose in a given corpus. While CSW style metrics are appropriate for evaluating proximity in BM and SEAME at the conversation-level, these do not yield a naturally balanced classification task for MaSaC, as the corpus lacks significant conversation-level proximity over this feature set (Sections 4.2 and 5.1). Therefore, here, we construct an *artificial* entraining/non-entraining split for MaSaC by taking the median of an aggregate score across CSW presence, ratio, and strategy, and assigning conversations above the median to the positive class. We otherwise follow the same feature construction, speaker-averaging, and grid search procedure described in Section 4.2 and Appendix A.6.

We caution that this analysis differs in an important respect from our primary classification tasks. Since the entraining/non-entraining label used here is constructed directly as a threshold on the average of the same CSW style features that are also included in the classifier's input feature vector, the resulting feature importance and ablation results (Figures 18 and 19) are expected to concentrate heavily on these features by construction, independent of any genuine model sensitivity to human-relevant entraining behavior. This is distinct from the circularity concern we raise for our primary analyses in Appendix A.6: there, the positive class

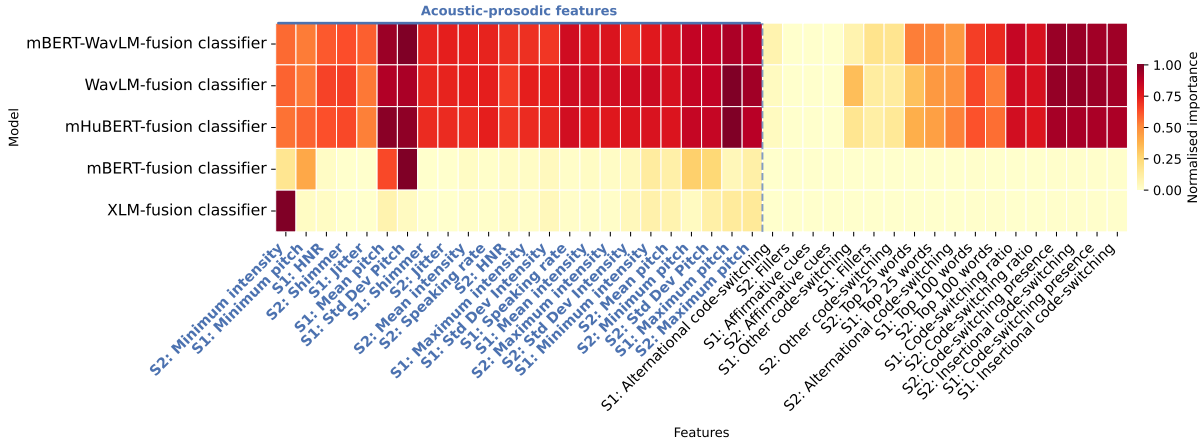

Figure 12: Feature importance values across best-performing MaSaC models at the turn-level. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable. *mBERT-fusion* with hyperparameter lr: 2e-5 achieves 70% test-time accuracy, *XLM-fusion* with hyperparameter lr: 5e-5 achieves 69%, *WavLM-fusion* with hyperparameter lr: 1e-3 achieves 74%, *mHuBERT-fusion* with hyperparameter lr: 1e-3 achieves 70%, and *mBERT+WavLM-fusion* with hyperparameter lr: 1e-3 achieves 77%.

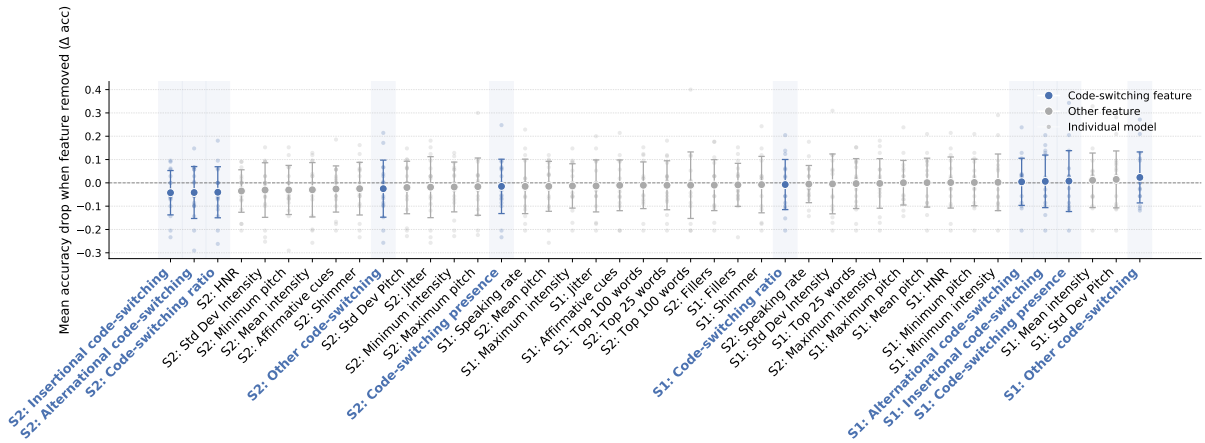

Figure 13: Aggregated feature ablation results for SEAME at the conversation-level. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful.

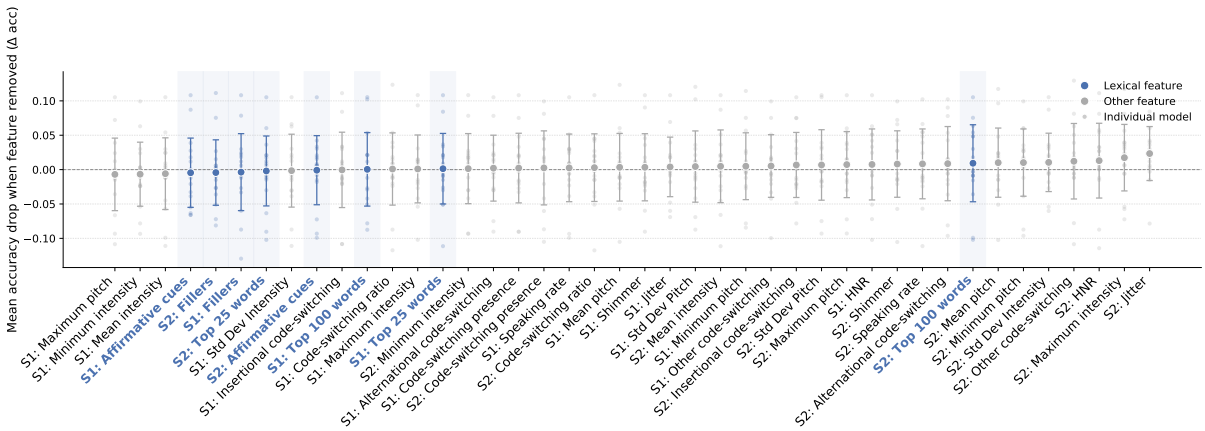

Figure 14: Aggregated feature ablation results for MaSaC at the conversation-level. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful.

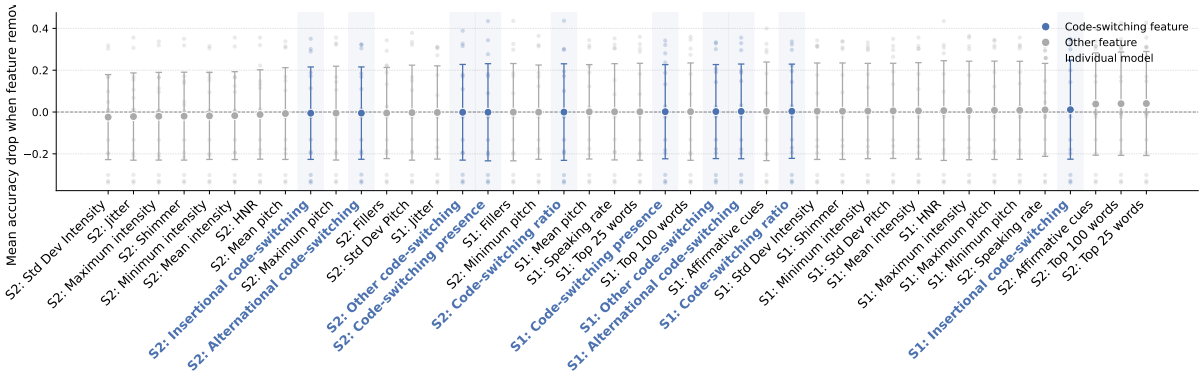

Figure 15: Aggregated feature ablation results for SEAME at the turn-level. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful.

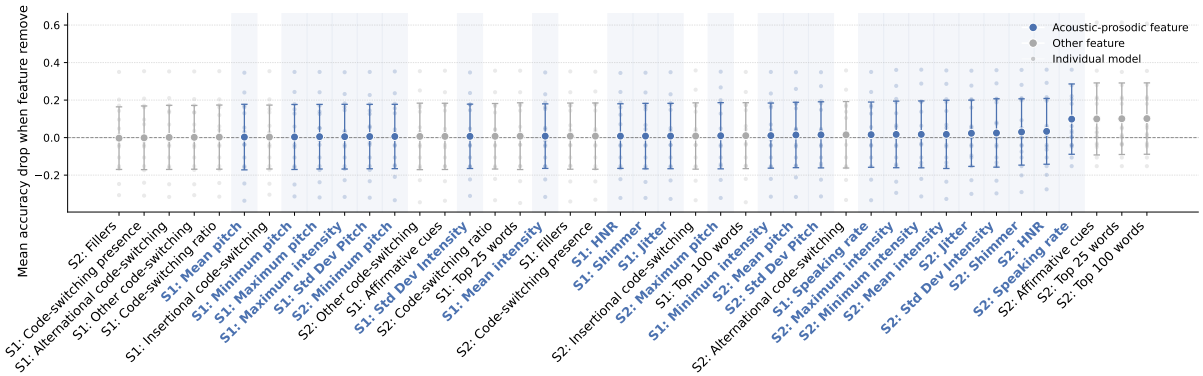

Figure 16: Aggregated feature ablation results for MaSaC at the turn-level. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful.

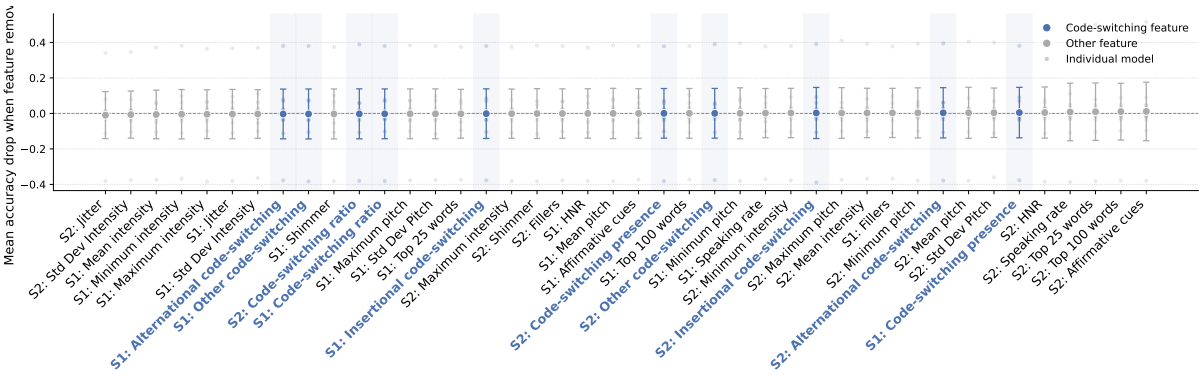

Figure 17: Aggregated feature ablation results for BM at the turn-level. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful.

is defined by a statistical significance test over observed human entrainment, a non-trivial target that a classifier is not guaranteed to recover from the raw feature values alone. Here, in contrast, the label is a direct arithmetic function of a subset of the input features themselves, so strong classifier reliance on those features reflects the construction of the task rather than a positive, human-grounded finding.

Consistent with this, Figure 18 shows near-maximal importance concentrated almost entirely on the CSW style features across models, and Figure 19 shows a similar pattern: three of the five features with the largest ablation impact are CSW style features. Together, these results reflect the label's own construction rather than demonstrating that classifiers attend to human-relevant CSW style features when detecting entrainment – a direct consequence of MaSaC's lack of significant conversation-level CSW style proximity, noted above. We report

this analysis for completeness and to illustrate the effect of feature-label circularity when it is present in an unambiguous form, but we do not interpret these results as evidence of human-grounded model behavior, and we caution against comparing them directly to the findings reported in Section 5.2.

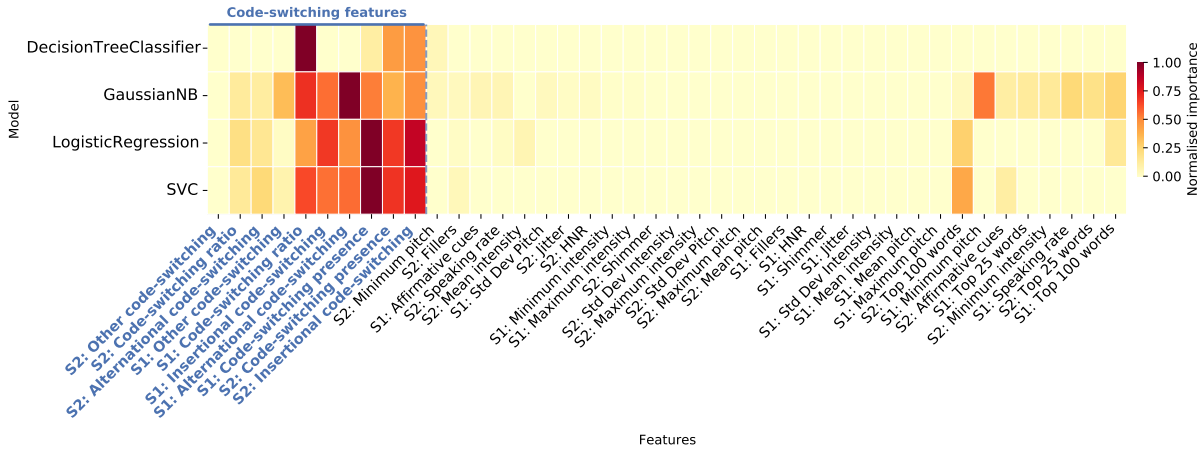

Figure 18: Feature importance values across the best-performing models from the secondary conversation-level analysis for MaSaC described in Appendix A.15. We normalize values per model using row-wise min.-max. so that models with different importance scales are comparable. Unlike the corresponding figures for our primary analyses, the entraining/non-entraining label underlying this task is constructed directly from the CSW style features shown here, so the concentration of importance on these features reflects the label's construction rather than a positive, human-grounded finding; see Appendix A.15 for discussion.

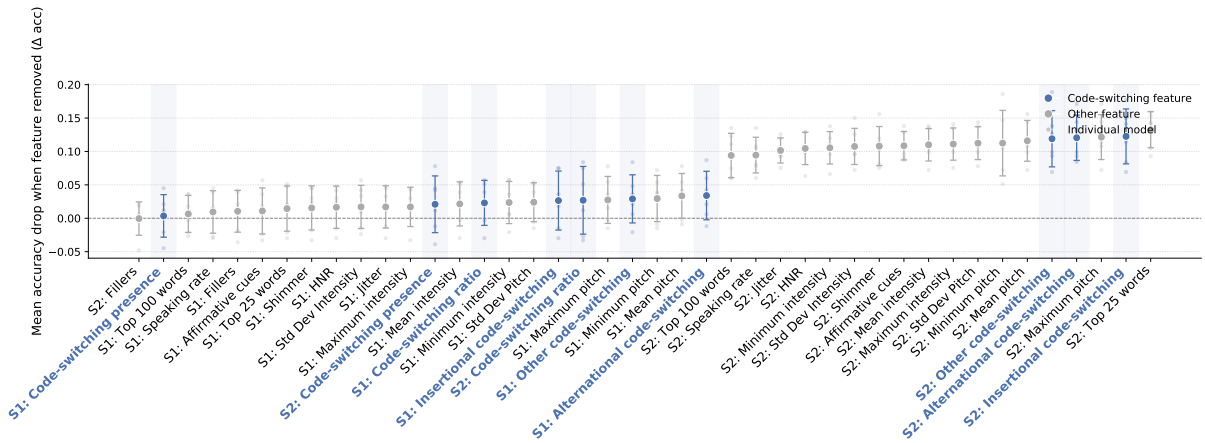

Figure 19: Aggregated feature ablation results for the secondary conversation-level analysis for MaSaC described in Appendix A.15. Features are ordered left to right by increasing impact; $\Delta_{acc.} > 0$ means accuracy *drops* when the feature is removed, i.e., the feature is useful. As with the corresponding feature importance results (Figure 18), the entraining/non-entraining label for this task is constructed directly from the CSW style features shown here; the resulting ablation pattern should not be interpreted as evidence that classifiers attend to human-relevant CSW style features when detecting entrainment.

Model	#Parameters	Hyperparameters tried
LogisticRegression(class_-weight='balanced', max_iter=1000)	43	• C = [0.001, 0.001, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 10]
SVC(class_weight='balanced', kernel='linear', max_iter=-1)	43	• C = [0.001, 0.001, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 10]
DecisionTreeClassifier(class_-weight='balanced')	22 / 3K / 98K[a]	• criterion = ['gini', 'entropy', 'log_loss'] • max_depth = [1, 3, 5, 10, 12, 15, 20, 25, 30] • min_samples_split = [2, 5, 10]
RandomForestClassifier(class_-weight='balanced')	400 / 9.4K[a]	• criterion = ['gini', 'entropy', 'log_loss'] • max_depth = [1, 3, 5, 10, 12, 15, 20, 25, 30] • min_samples_split = [2, 5, 10]
GaussianNB()	170	NA
KNeighborsClassifier()	0	• n_neighbors = [1, 3, 5, . . . , 49] • weights = ['distance', 'uniform'] • algorithm = ['ball_tree', 'kd_tree', 'brute'] • leaf_size = [5, 10, . . . , 45]
MLPClassifier(max_iter=1000)	4.5K	• hidden_layer_sizes = [10, 50, 100, . . . , 500] • activation = ['identity', 'logistic', 'tanh', 'relu'] • solver = ['lbfgs', 'adam'] • alpha = [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1]
GradientBoostingClassifier()	1.1K / 3.3K / 7.7K[b]	• loss = ['log_loss', 'exponential'] • learning_rate = [1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1] • n_estimators = [50, 100, . . . , 500] • criterion = ['friedman_mse', 'squared_error']
XGBClassifier(scale_pos_-weight=pos_class_weight, max_delta_step=1)	1.1K / 1.2K / 11K[c]	• learning_rate = [1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1] • n_estimators = [50, 100, . . . , 500] • max_depth = [1, 3, 5, 10] • booster = ['gbtree']
(adapted) TabTransformer(emb_dim=16, transformer_dim=32, depth=2, heads=4)	141K	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
FT-Transformer(emb_dim=32, depth=3, heads=4, dropout=0.1)	43K	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
XLM-R(hidden_dim=128)	133K (+278M frozen)	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
mBERT(hidden_dim=128)	133K (+178M frozen)	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
RemBERT(hidden_dim=128)	182K (+575M frozen)	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
WavLM-Base-Plus(sample_rate=16000, emb_dim=768, hidden_dim=256, dropout=0.2)	986K	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
mHuBERT-147(sample_rate=16000, emb_dim=768, hidden_dim=256, dropout=0.2)	986K	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]
mBERT/RembERT + WavLM/mHuBERT(hidden_dim=256, dropout=0.2)	724K (+178M frozen) / 822K (+575M frozen)[d]	• learning_rate = [5e-5, 2e-5, 1e-5, 1e-4, 1e-3, 0.01, 0.05, 0.1, 0.25, 0.5, 1]

Table 4: Models with number of learned parameters and hyperparameters tried in grid search. **Note:** [a] depends on `max_depth`; [b] depends on `n_estimators`; [c] depends on `max_depth` and `n_estimators`; [d] depends on the model combination selected.

Feature	**Turn-level Proximity** (d)	**Conversation-level Proximity** (d)	**Conversation-level Convergence** (d)
CSW style			
CSW presence	-0.018	-1.211	—
CSW ratio	-0.050	-1.579	—
CSW strategy (*I*)	-0.016	-0.673	—
CSW strategy (*A*)	-0.0003	-0.552	—
CSW strategy (*O*)	-0.0017	-0.689	—
Acoustic-prosodic			
Min. pitch	—	—	—
Mean pitch	—	0.054	—
Max. pitch	—	—	—
SD pitch	—	0.019	—
Min. intensity	—	-0.049	—
Mean intensity	-0.037	—	—
Max. intensity	-0.030	—	—
SD intensity	-0.0099	—	—
Jitter	—	-0.048	—
Shimmer	—	0.028	—
HNR	-0.021	—	—
Speaking rate	—	—	—

Table 5: Cohen's d for turn- and conversation-level proximity and conversation-level convergence comparisons in SEAME. "—" indicates a feature not identified as significant in Section 5.1's within-corpus analysis, for which an effect size was therefore not computed.

Feature	**Lexical Entrainment** (d)
Top-100 words	1.843
Top-25 words	1.795
Top-25 (conv.-level)	1.460
Affirmative cues	1.679
Fillers	0.510

Table 6: Cohen's d for lexical entrainment comparisons in SEAME.

Feature	Turn-level Proximity (d)	Conversation-level Proximity (d)	Conversation-level Convergence (d)
CSW style			
CSW presence	—	—	—
CSW ratio	—	—	—
CSW strategy (*I*)	—	—	—
CSW strategy (*A*)	—	—	—
CSW strategy (*O*)	—	—	—
Acoustic-prosodic			
Min. pitch	—	—	—
Mean pitch	-0.060	0.026	—
Max. pitch	—	—	—
SD pitch	-0.027	0.214	—
Min. intensity	0.027	—	—
Mean intensity	—	—	-0.155
Max. intensity	—	—	-0.165
SD intensity	0.030	—	—
Jitter	-0.011	—	—
Shimmer	0.023	0.228	—
HNR	0.029	0.157	—
Speaking rate	0.011	0.148	0.132

Table 7: Cohen's d for turn- and conversation-level proximity and conversation-level convergence comparisons in MaSaC. "—" indicates a feature not identified as significant in Section 5.1's within-corpus analysis, for which an effect size was therefore not computed.

Feature	Lexical Entrainment (d)
Top-100 words	0.374
Top-25 words	0.291
Top-25 (conv.-level)	-1.461
Affirmative cues	-0.009
Fillers	0.015

Table 8: Cohen's d for lexical entrainment comparisons in MaSaC.

Feature family	Turn-level Convergence		Turn-level Synchrony	
	Mean r	Median r	Mean r	Median r
CSW style	-0.015	-0.017	0.031	0.027
Acoustic-prosodic	-0.009	-0.017	0.012	0.009

Table 9: Mean and median Pearson's r for turn-level convergence and synchrony in SEAME, aggregated by feature family.

Feature family	Turn-level Convergence		Turn-level Synchrony	
	Mean r	Median r	Mean r	Median r
CSW style	0.070	0.080	-0.259	-0.059
Acoustic-prosodic	0.017	0.023	0.006	-0.001

Table 10: Mean and median Pearson's r for turn-level convergence and synchrony in MaSaC, aggregated by feature family.

Feature family	Surviving	Total	Notable corrected results
Lexical	1	18	Only one contrast (top-100 corpus-level words, SEAME vs. MaSaC) survives correction, supporting the claim that lexical entrainment generalizes robustly across language pairs.
Acoustic-prosodic	11	178	The majority of surviving contrasts involve SEAME diverging from both BM and MaSaC, consistent with the characterization of Mandarin-English acoustic-prosodic patterns as the most distinct of the three pairs.
CSW style	12	60	Surviving contrasts are concentrated in the features we highlight in Section 5.1: MaSaC vs. BM and MaSaC vs. SEAME differ significantly on conversation-level alternational CSW proximity ($p_{corrected} < 0.001$, both) and turn-level insertional CSW synchrony ($p_{corrected} < 0.001$, both), while BM and SEAME remain statistically indistinguishable from each other on the same features. This is consistent with the characterization of MaSaC's CSW style entrainment as markedly different from the other two corpora.

Table 11: Cross-corpus Bonferroni-corrected pairwise comparisons by feature family.

Corpus	Feature family	Surviving	Total	Notable corrected results
SEAME	Lexical	6	6	All five word classes and overall language use remain significant ($p_{corrected} < 0.001$, each).
MaSaC	Lexical	4	6	Four of six features remain significant: corpus-level top-100 and top-25 words, conversation-level top-25 words, and overall language use ($p_{corrected} < 0.01$, each); affirmative cues and fillers do not survive correction ($p_{corrected} = 1.0$, both). This indicates that the Hindi-English lexical entrainment asymmetry reported in Section 5.1 is robust to multiple-comparison correction rather than an artifact of uncorrected significance testing.
SEAME	Acoustic-prosodic	17	60	Few turn-level proximity contrasts survive correction (3 of 12 features: mean and maximum intensity and HNR), consistent with the reported pattern of inconsistent proximity in Mandarin-English. Conversation-level proximity and convergence largely fail to survive correction, supporting the claim that Mandarin-English mirrors the absence of conversation-level acoustic-prosodic entrainment previously reported for Spanish-English.
MaSaC	Acoustic-prosodic	26	60	Turn-level proximity contrasts remain overwhelmingly significant (10 of 12 features: all except minimum pitch and maximum intensity; $p_{corrected} < 0.05$), supporting the claim of substantially stronger acoustic-prosodic entrainment in Hindi-English than in Mandarin-English. All turn-level convergence contrasts survive correction (12 of 12 features; $p_{corrected} < 0.05$, each). Conversation-level proximity and convergence largely fail to survive correction, supporting the claim that Hindi-English mirrors the absence of conversation-level acoustic-prosodic entrainment previously reported for Spanish-English.
SEAME	CSW style	9	20	Significant contrasts that survive correction are concentrated in turn- and conversation-level proximity, supporting the findings reported in Section 5.1.
MaSaC	CSW style	4	20	Significant contrasts that survive correction are concentrated exclusively in turn-level convergence across CSW presence, ratio, and insertional strategy ($p_{corrected} < 0.001$, each); no other measure or entrainment level survives correction. This supports the claim that Hindi-English CSW style entrainment in MaSaC is distinctively convergence-driven, in contrast to the proximity- and synchrony-dominant patterns observed observe for Mandarin-English (and those previously reported for Spanish-English).

Table 12: Within-corpus Bonferroni-corrected significance by feature family.

Corpus	FS 1	FS 2	Mean r	SD r	% sig.
BM	CSW	CSW	0.547	0.259	53.3
BM	L	L	-0.117	0.329	46.7
BM	AP	AP	0.369	0.310	51.8
BM	CSW	L	0.003	0.138	1.67
BM	CSW	AP	-0.083	0.155	10.4
BM	L	AP	-0.023	0.202	9.72
SEAME	CSW	CSW	0.573	0.434	46.7
SEAME	L	L	-0.063	0.310	20.0
SEAME	AP	AP	0.209	0.326	23.9
SEAME	CSW	L	0.072	0.350	31.7
SEAME	CSW	AP	-0.021	0.231	13.8
SEAME	L	AP	-0.052	0.205	16.0
MaSaC	CSW	CSW	0.197	0.297	24.4
MaSaC	L	L	-0.024	0.165	40.0
MaSaC	AP	AP	0.057	0.294	63.8
MaSaC	CSW	L	0.050	0.133	22.0
MaSaC	CSW	AP	0.016	0.073	11.3
MaSaC	L	AP	0.002	0.086	15.8

Table 13: Summary of conversation-level feature set correlations. FS = feature set; CSW = code-switching style aspects; L = lexical features; AP = acoustic-prosodic features.

Corpus	FS 1	FS 2	Mean r	SD r	% sig.
BM	CSW	CSW	0.191	0.251	42.2
BM	L	L	0.185	0.251	57.1
BM	AP	AP	0.249	0.236	86.6
BM	CSW	L	0.026	0.049	46.3
BM	CSW	AP	0.001	0.029	18.8
BM	L	AP	0.019	0.047	32.8
SEAME	CSW	CSW	0.159	0.269	44.4
SEAME	L	L	-0.030	0.058	40.0
SEAME	AP	AP	0.153	0.231	84.8
SEAME	CSW	L	0.024	0.105	53.3
SEAME	CSW	AP	0.017	0.072	68.3
SEAME	L	AP	0.012	0.084	49.3
MaSaC	CSW	CSW	0.182	0.295	46.7
MaSaC	L	L	0.162	0.242	33.3
MaSaC	AP	AP	0.056	0.263	70.7
MaSaC	CSW	L	-0.019	0.052	24.3
MaSaC	CSW	AP	0.015	0.083	35.4
MaSaC	L	AP	0.004	0.052	20.2

Table 14: Summary of turn-level feature set correlations. FS = feature set; CSW = code-switching style aspects; L = lexical features; AP = acoustic-prosodic features.