Does Context Matter? A Prosodic Comparison of English and Spanish in Monolingual and Multilingual Discourse Settings

Debasmita Bhattacharya*[1], David Sasu*[2], Michela Marchini[3], Natalie Schluter[2], Julia Hirschberg[1]

[1]Columbia University, [2]IT University of Copenhagen, [3]University of Michigan

Correspondence: debasmita.b@cs.columbia.edu; dasa@itu.dk

Abstract

Different languages are known to have typical and distinctive prosodic profiles. However, the majority of work on prosody across languages has been restricted to monolingual discourse contexts. We build on prior studies by asking: how does the nature of the discourse context influence variations in the prosody of monolingual speech? To answer this question, we compare the prosody of spontaneous, conversational monolingual English and Spanish both in monolingual *and* in multilingual speech settings. For both languages, we find that monolingual speech produced in a *monolingual* context is prosodically different from that produced in a *multilingual* context, with more marked differences having increased proximity to multilingual discourse. Our work is the first to incorporate multilingual discourse contexts into the study of native-level monolingual prosody, and has potential downstream applications for the recognition and synthesis of multilingual speech.

1 Introduction

Prosodic analysis has been a cornerstone of speech processing research for decades (Cutler et al., 1997; Wagner and Watson, 2010). Much of the early work in this area has traditionally focused on monolingual productions of highly-resourced languages, partly due to practical concerns of data availability and modeling capabilities (Xu, 2011). With the field's growing interest in lower-resourced languages and multilingual domains (Hasija et al., 2022; Nikolaev et al., 2015), there is considerable scope for exploring prosody in diverse language settings. In particular, the impact on prosody of multilingual discourse contexts is one area in which our understanding remains incomplete, but would be essential for improving both the comprehension and synthesis of human-like monolingual and multilingual speech, among other applications.

In this work, we take the first step toward filling this gap by exploring the prosodic features of monolingual speech across different types of multilingual discourse context. Such contexts exist on a spectrum, at one end of which are purely monolingual settings, followed by domains consisting primarily of monolingual speech in one language alternating infrequently with other languages between sentences, paragraphs, or even entire documents. At the opposite end of the spectrum are code-switched domains where various languages are interspersed more granularly and typically at the intra-sentential level. This work compares monolingual speech from two extremes of the multilingual spectrum: one involving little to no contact between distinct languages, and the other involving code-switching. We define the former as a *monolingual* discourse context in which all speech consists of a single language; in contrast, the latter is a *multilingual* discourse context consisting of code-switched speech alternating between two languages within and between utterances.

We specifically examine the prosody of spontaneous, informal productions of monolingual English and Spanish, situated in both monolingual and multilingual, i.e. code-switched, contexts. To do so, we investigate U.S. English and Spanish as spoken in the monolingual CallHome corpora (Canavan et al., 1997; Canavan and Zipperlen, 1996) and compare this to the U.S. English and Spanish spoken in the Bangor Miami corpus of multilingual and code-switched speech (Deuchar, 2011).

We find evidence of monolingual speech produced in a monolingual context having significant prosodic differences from that produced in a multilingual context, which holds true across both English and Spanish. These differences are marked enough to be learned by end-to-end predictive models, and become even more pronounced with increased proximity of monolingual utterances to multilingual discourse. Overall, we find that multi-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23320–23334

lingual speech settings have a meaningful influence on the prosody of monolingual speech. Our work is the first to study monolingual prosody across different types of discourse context, and our main contribution is a novel insight on nuanced prosodic production in varied contexts. We hope that our work will inform innovation in downstream speech applications, including predicting, recognizing, and synthesizing multilingual and code-switched speech.

2 Related work

There exists an extensive literature on the prosodic patterns of various languages. Computational work on intonation has focused on analyzing native-level monolingual speech (Schack, 2000; Torres, 2024), for which the ToBI framework (Silverman et al., 1992), along with its extensions, has been particularly important. Related studies such as Rosenberg et al. (2012) have shown that different languages have distinct, typical prosodic profiles, which can be leveraged to perform language identification (LID) on both read and spontaneous speech (Rouas et al., 2003) across many languages including English, German, Japanese, Mandarin, Spanish, and Hindi, among others (Cummins et al., 1999; Rao et al., 2015; Bhattacharjee and Sarmah, 2013). Such prosodically-driven LID systems have used a variety of front- and back-ends and configurations such as hierarchical and fusion classifiers (Ambikairajah et al., 2011), but almost all have implicitly assumed a monolingual discourse context. Prior work that has explicitly considered contextual influence on prosody, such as Cole (2015), has largely defined context in terms of syntactic, lexical, and discourse levels which vary according to the spontaneity of the speech situation. While prosody has been interpreted relative to features of neighboring words, syntactic boundaries, and discourse units, it is yet to be considered relative to the broader multilingual status of the conversation as a whole.

Studies most closely related to our work have examined prosodic interference, e.g. Nikolaev et al. (2015), and focused on the interaction of prosodic elements in second language acquisition settings (Bowen, 1956; Graham, 1978; Kainada and Lengeris, 2015; Ding and Hoffmann, 2015). Some of these studies performed very fine-grained analyses – for instance, only comparing pitch and intonational contours between questions and declarative statements in English and Spanish (Delattre et al., 1962). Such work, however, has been restricted to the effect of native prosody on foreign language speech, rather than the prosodic interplay between two first languages. The few studies that have considered a speech setting involving interactions between multiple native languages have generally conducted small-scale studies focused on language development and prosodic mixes among young children (Schmidt and Post, 2015; Cruz-Ferreira, 1999), or on listeners' ability to anticipate upcoming monolingual and code-switched speech (Piccinini and Garellek, 2014).

Overall, prior work indicates that close contact between languages influences prosody particularly as produced by children and language learners, but it remains unclear what the impact of more distant language contact in multilingual discourse settings is on the prosody of adult, native-level speakers. To address this gap, we ask **RQ:** How does the prosodic character of monolingual speech in a given language vary depending on its monolingual versus multilingual discourse context? In other words, how does a multilingual discourse context influence the prosody of monolingual speech productions?

3 Data

We examine two monolingual corpora, CallHome English (CH-E) and CallHome Spanish (CH-S) (Canavan et al., 1997; Canavan and Zipperlen, 1996), and one multilingual corpus, Bangor Miami (BM) (Deuchar, 2011). Both CallHome English and CallHome Spanish are made available under the LDC User Agreement for Non-Members. The Bangor Miami corpus is made available under the GNU General Public License version 3 or later.

As specified in individual corpus documentation, all three data sets were recorded in similar time periods and consist of spontaneous, informal conversations among friends or relatives on common topics of discussion (e.g. family life, relationships, hobbies, etc.), limiting systematic differences between corpora in discourse theme or interlocutor rapport. Each corpus is reasonably gender-balanced over the represented speakers. The monolingual corpora include recorded telephone speech, while BM comprises recordings of in-person conversations. From BM, we use the monolingual English and Spanish subsets of the corpus, which respectively make up about 63% and 26% of the data set. The remaining 11% of BM comprises code-switched utterances

Corpus	Total hours of speech	Number of dialogues	Number of speakers	Mean [SD] utterance length (s)
CH-E	33	71	142	3.7 [6.8]
CH-S	26	84	168	3.4 [6.0]
BM	35	56	84	3.1 [87.5]

Table 1: Summary of corpus statistics for each data set. Further detail on the idiosyncrasies of BM is in Appendix A.

and non-speech segments, e.g. laughter, which we exclude from our analysis. We summarize additional corpus statistics in Table 1 and share further detail about the BM corpus in Appendix A.

4 Method

4.1 Feature extraction and statistical testing.

For each corpus, we extract the Disvoice set[1] of 103 utterance-level prosodic features derived from fundamental frequency, energy, and duration across voiced and unvoiced segments and six statistical functionals: mean, standard deviation, maximum, minimum, skew, and kurtosis. These features are grouped into three prosodic categories: the first 30 features are related to pitch, the next 48 are related to energy, and the last 25 are related to duration. We transform the extracted features into vector representations again using DisVoice (Dehak et al., 2007; Vasquez et al., 2018). We compare these features of English and Spanish across the two discourse contexts using feature-level independent t-tests over feature distributions of a) English in CallHome versus Bangor Miami, and b) Spanish in CallHome versus Bangor Miami. Later, to examine the specific, positional influence of the multilingual discourse context on prosody, we use the same statistical testing method to compare monolingual BM utterances that are "closer" to the multilingual context, i.e. a code-switched utterance, to those that are further away.

4.2 Model building.

Inspired by existing literature on the use of prosody for LID (see Section 2), we attempt a novel approach that leverages LID models to understand variations in prosody. Note that LID as a task is *not* the primary focus of this work; we only use LID as a *tool* for understanding prosody. We hypothesize that a prosodic LID model trained on monolingual English and Spanish from a monolingual discourse context will perform better when tested on speech from a monolingual context than when tested on speech from a multilingual context, due to inherent prosodic differences that occur from the broader discourse setting. We expect the opposite to be true for the performance of a prosodic LID model trained on monolingual speech from a multilingual discourse context.

To test this hypothesis, we build two custom prosodic LID models for binary classification, both using a Transformer base model architecture (Vaswani et al., 2017). We select these models by performing an extensive hyperparameter grid search over 30 epochs, testing the efficacy of 50 different combinations of viable hyperparameters across 6 different model architectures, and choosing the ones that maximize accuracy (val. accuracy: 0.51 and 0.61). We do so using the Optuna hyperparameter optimization framework.[2] Each model takes utterance-level prosody vector representations as input, and outputs a language label for that utterance: either English or Spanish. One model (Transformer-CH) is trained on the prosodically-encoded combined train sets of CH-E and CH-S with a learning rate of 8.84 x 10^{-5}, while the other (Transformer-BM) is trained on the prosodically-encoded monolingual train set of BM with a learning rate of 1.89 x 10^{-4}. Both are trained over 200 epochs using the Adam optimizer, chosen for its adaptive learning rate and efficient convergence properties. Additionally, both models employ a batch size of 32, ensuring stable and efficient gradient updates. Architecturally, the encoder in Transformer-CH comprises a single layer, in contrast to the three-layer encoder in Transformer-BM. In both cases, the encoder's feed-forward network is configured with a hidden size of 128 to facilitate the learning of complex non-linear representations. To enrich the embedding space, Transformer-CH uses a projection dimension of 320, while Transformer-BM utilizes a projection dimension of 256. In terms of overall model complexity, Transformer-CH comprises 848,450 param-

[1] https://disvoice.readthedocs.io/en/latest/Prosody.html

[2] https://optuna.org

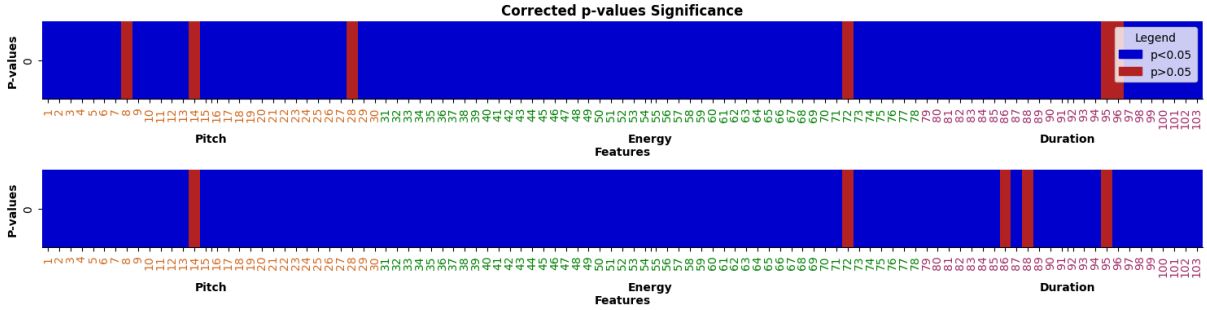


Figure 1: p-values for prosodic feature distribution comparisons between CallHome and Bangor Miami: English (top) and Spanish (bottom). Blue indicates feature-level statistical significance; red indicates insignificance. Orange ticks indicate pitch features, green ticks indicate energy features, and pink ticks indicate duration features.

eters, while Transformer-BM contains 1,273,474 parameters. The data used follows a 70%-20%-10% train-test-val split. To prevent the models from learning biases toward the majority class, we ensure that the training, testing, and validation sets are balanced in terms of class distribution. We run experiments on three v100 GPUs, on which hyperparameter search took about 8 hours and training took about 2 hours.

To confirm the reliability of our custom models, we also perform inference with a large (1B parameters) off-the-shelf LID model, Facebook-MMS-LID (Pratap et al., 2024), pre-trained on monolingual speech in 256 languages, including English and Spanish, from monolingual contexts. We use this model to perform the same binary classification task, i.e. prosodic LID, as above, testing it on data from both the monolingual and multilingual discourse contexts. Note that this large pre-trained model makes use of acoustic features in addition to the prosodic ones that we focus on. We would have preferred to use a prosody-only LID model for controlled comparison, but such pre-trained models are not openly accessible.

5 Results

5.1 Prosodic profiles differ significantly in monolingual versus multilingual discourse contexts.

We begin with a statistical comparison of the prosodic features of monolingual English and Spanish between the monolingual discourse contexts of CH-E and CH-S and the multilingual discourse context of BM. We find that the vast majority of prosodic features in both English and Spanish are significantly different ($p < 0.05$) in a monolingual context than in a multilingual one (Figure 1). A few prosodic features diverge from this general trend, e.g. MSE and tilt of voiced fundamental frequency (F0) linear estimation[3] (features 8 and 14) are not significantly different for monolingual English in CH-E compared to BM, while pause duration (feature 95) is not significantly different for monolingual Spanish in CH-S compared to BM. However, the overall prosodic profile of a language does indeed differ meaningfully depending on the nature of the discourse context, which is especially striking given that we compare monolingual instances of the *same* language in each case.

The above result holds true even when we account for 1) the differing noise conditions of the CallHome and Bangor Miami corpora – only the latter corpus contains babble noise as a byproduct of its public recording setup; 2) the variation in utterance length distribution beyond 7 seconds in each language (Figure 2); and 3) the differing recording conditions of the CallHome and Bangor Miami corpora, whereby the former data were collected over the telephone and the latter using in-room microphones. We replicate our results using 1) denoised versions of the BM data sourced from prior work (Bhattacharya et al., 2024)[4] (Figure 3), 2) the subset of monolingual utterances in each corpus of length less than 7 seconds (Figure 4), and 3) downsampled BM audio files of 8 kHz, which match the audio sampling rate of both CallHome corpora (Figure 8 in Appendix A.4).

Over the corpus as a whole, when considering prosodic features by type (pitch, energy, or duration), group-level statistically significant differences are similar in proportion for both English and Spanish; each prosodic feature group con-

[3]We provide a glossary of these terms in Appendix A.

[4]Please see Bhattacharya et al. (2024) for complete methodological details of how denoising was carried out.

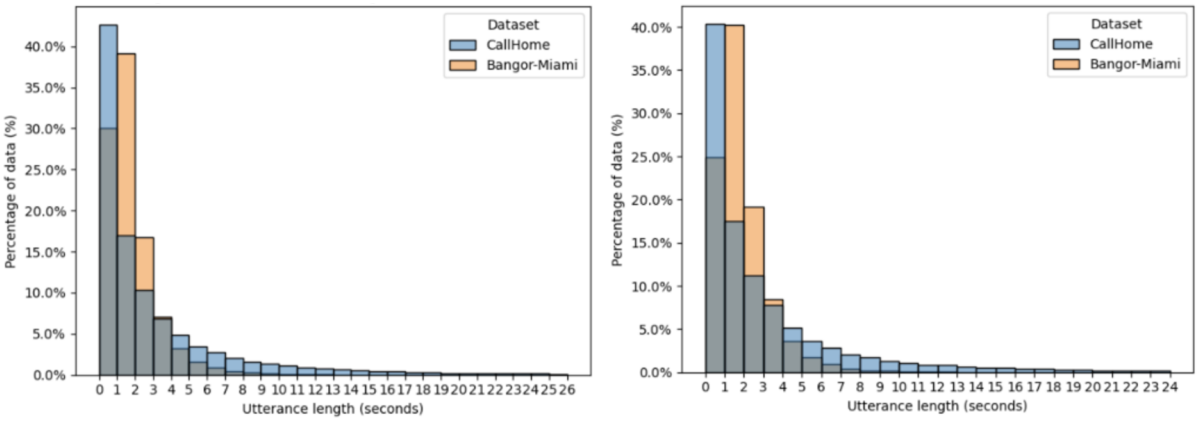


Figure 2: Comparing utterance length distributions of English (left) and Spanish (right) in CallHome to Bangor Miami.

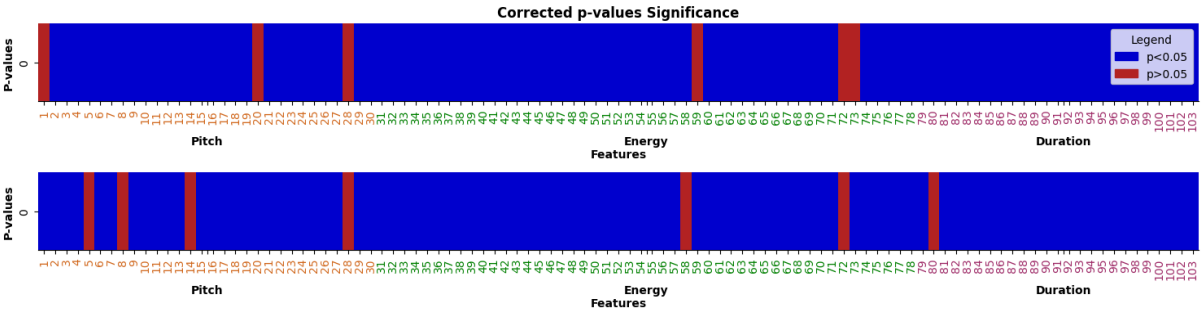


Figure 3: p-values for prosodic feature distribution comparisons between CallHome and denoised Bangor Miami: English (top) and Spanish (bottom). Blue indicates feature-level statistical significance; red indicates insignificance. Orange ticks indicate pitch features, green ticks indicate energy features, and pink ticks indicate duration features.

tains at least one feature that shows statistically insignificant differences in both languages. For neither language is there a prosodic feature group that consistently represents a greater proportion of statistically significant differences than the other groups across the various experimental control settings (Figures 1, 3 , 4, and 8). This suggests that even granular prosodic differences in monolingual speech between monolingual and multilingual discourse contexts may be similar for both English and Spanish.

We further inspect the nature of these differences in prosodic profiles, highlighting a few observations from our qualitative comparison of monolingual English and Spanish in each discourse context. For both English and Spanish, F0 contours generally have lower values in the multilingual setting of BM compared to the monolingual settings of CH-E and CH-S, with less variation in F0 contour in the multilingual discourse context. Initial voiced segments in both languages have lower average F0 in a multilingual setting. This is also true for final voiced segments in Spanish. In terms of energy, initial and final voiced segments in both languages have greater mean values in a multilingual setting than in a monolingual one. This difference is even more marked for the average energy of unvoiced segments. Finally, voicing rate and the ratio of duration of pauses to combined voiced and unvoiced segments are lower in a multilingual context than in a monolingual one. We visualize some of these trends in Figure 5. Cumulatively, we find that English and Spanish produced in a multilingual context tend to sound lower-pitched, louder, less rhythmic, and less disjointed than English and Spanish produced in a monolingual context. The consistency of these qualitative patterns across both languages is striking, and suggests that a multilingual discourse context has a uniform impact on the prosody of monolingual speech. In sum, these results provide the motivation for the remainder of the work.

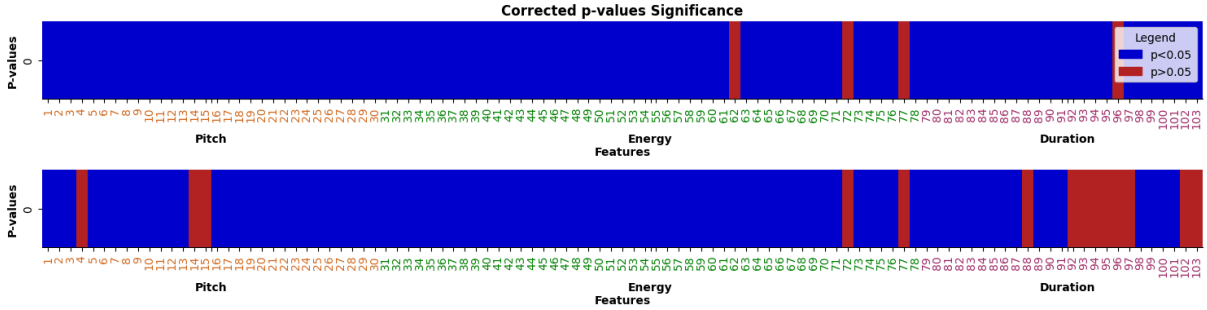
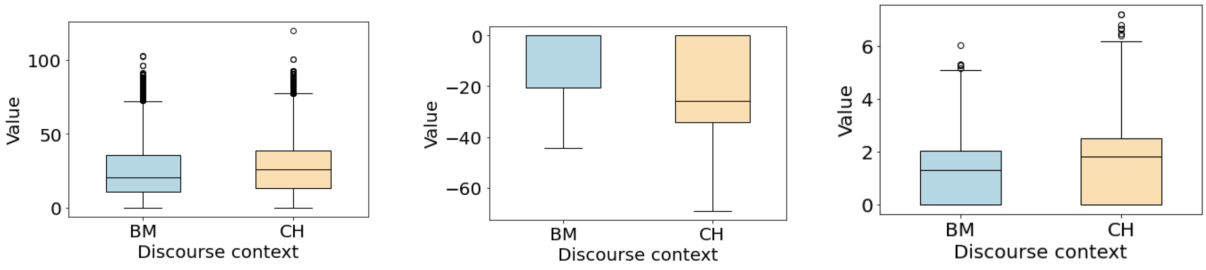


Figure 4: p-values for prosodic feature distribution comparisons between CallHome and denoised Bangor Miami over length-controlled utterances (<=7 seconds): English (top) and Spanish (bottom). Blue indicates feature-level statistical significance; red indicates insignificance. Orange ticks indicate pitch features, green ticks indicate energy features, and pink ticks indicate duration features.



(a) (b) (c)

Figure 5: Visualizing trends in (a) mean F0 in monolingual Spanish, (b) mean energy of initial unvoiced segments in monolingual English, and (c) voicing rate in monolingual Spanish across monolingual (CH; CallHome) and multilingual (BM; Bangor Miami) discourse contexts.

5.2 Models can learn prosodic profile differences between monolingual and multilingual discourse contexts.

Next, we examine whether end-to-end predictive models can learn the underlying differences in monolingual English and Spanish prosody that seem to arise from the discourse context.

5.2.1 Evaluating our custom models.

We find that our custom model trained on English and Spanish in *monolingual* contexts (Transformer-CH) performs better at the LID task when tested on monolingual data from a *monolingual* context than when tested on monolingual data from a *multilingual* context (Table 2). This difference in performance is statistically significant with $p < 0.001$ according to a z-test of proportions comparing model accuracy on the test samples. This is also consistent with the results from our custom model trained on a *multilingual* context (Transformer-BM), which performs better at LID when tested on monolingual English and Spanish from a *multilingual* context than from *monolingual* contexts. Again, the difference in performance, according to a z-test of proportions, is statistically significant with $p < 0.001$. While the overall performance of our custom models in each testing configuration is modest, the key aspect of these results is the *relative* performance in distinct discourse contexts. Since the languages and conversational genre under consideration in each of these settings are identical, and we have already shown that the impact of noise, utterance length, and recording channel discrepancies across settings is minimal, these differences in model performance provide further support for prosodic variation in monolingual speech depending on the multilingual nature of the broader discourse context.

5.2.2 Ablating our custom models.

To develop further insight into which *types* of prosodic features matter the most to prosodic variation that is driven by the nature of the discourse context, we evaluate the contribution of each group of features to overall model performance through feature-group-level ablations. In other words, we re-train each model after removing each of the pitch, energy, and duration feature groups and assess the corresponding change in model perfor-

Model	Train set	Test set	Accuracy	F1 Score
Transformer-CH	CH-E+CH-S	CH-E+CH-S	0.592	0.591
Transformer-CH	CH-E+CH-S	BM	0.482	0.479
Transformer-BM	BM	CH-E+CH-S	0.478	0.478
Transformer-BM	BM	BM	0.551	0.540
Facebook-MMS	–	CH-E+CH-S	0.817	0.816
Facebook-MMS	–	BM	0.803	0.806

Table 2: Comparing model accuracy and F1 score on monolingual versus multilingual discourse context train/test configurations. Baseline accuracy associated with random/blind guessing in each case is 0.5. For associated confusion matrices, see Appendix A.3. For the equivalent custom model results on sampling-rate-matched audio data, see Table 10 in Appendix A.4.

mance at inference time. In each case, changes in model performance relative to the baseline show that pitch features play the most important role, followed by energy, and finally duration features (Table 3). This suggests that prosodic variation in monolingual speech between monolingual and multilingual contexts may primarily stem from pitch variations. However, since the absolute differences between the feature groups' contributions to overall model performance are relatively small, further work is required to definitively quantify any meaningful impact of variation in each of pitch, energy, and duration features on shaping overall prosodic differences between monolingual and multilingual discourse settings.

5.2.3 Replication with a large pre-trained model.

To rule out the performance differentials between monolingual and multilingual contexts found in Section 5.2.1 as being unique to our custom models, we replicate these results using Facebook-MMS-LID (Table 2). As with Transformer-CH, the test performance of Facebook-MMS-LID on data from a *monolingual* setting exceeds that on data from a *multilingual* setting, with a performance gap is that statistically significant according to a z-test of proportions ($p = 0.045$).[5] Since Facebook-MMS-LID was pre-trained on data from *monolingual* contexts only (Pratap et al., 2024), which importantly do not include any of the corpora we use, and both the CallHome and Bangor Miami corpora are out-of-distribution for this model, its relative performance across contexts lends validity to our custom model results. We note the moderate performance of this large pre-trained model is slightly surprising given its position as a state-of-the-art model, but the relevant takeaway from its inference results is the *difference* in performance between discourse contexts, rather than the absolute performance metric values.

Note that we do not perform ablations over Facebook-MMS-LID, as we lack the resources to re-train such a large, pre-trained model in multiple ablation settings.

Overall, our model performance evaluation results reinforce the statistically significant differences in prosodic profile of monolingual English and Spanish between monolingual and multilingual settings. Not only do such differences exist, these can also be learned by prosodic models whose downstream predictions are, in turn, influenced by such variation. In sum, this provides further evidence in support of the influence of discourse context on prosodic variation in monolingual English and Spanish.

5.3 Prosodic profile differences are enhanced with increased proximity to multilingual discourse.

To complete our investigation, we examine the specific impact of proximity to multilingualism on the prosody of monolingual speech in BM. We define monolingual BM utterances that are preceded or followed by code-switched utterances as being *close* to the multilingual discourse context of the conversational data; all other monolingual BM utterances are defined as being *far* from the multilingual context. Statistical testing again reveals significant differences over the vast majority of prosodic features between monolingual BM utterances spoken *close* to the multilingual context, compared to monolingual BM utterances spoken *far* from the multilingual context (Figure 6). This

[5]We speculate that this relatively small performance gap between discourse contexts could be due to acoustic, rather than exclusively prosodic, feature contributions to distinguishing between languages, but further work is required to confirm or refute this.

Model	Train set	Test set	Excluded feature group	Accuracy	F1 Score
Transformer-CH	CH-E+CH-S	CH-E+CH-S	–	0.592	0.591
Transformer-CH	CH-E+CH-S	CH-E+CH-S	Pitch	0.493	0.380
Transformer-CH	CH-E+CH-S	CH-E+CH-S	Energy	0.585	0.585
Transformer-CH	CH-E+CH-S	CH-E+CH-S	Duration	0.598	0.598
Transformer-CH	CH-E+CH-S	BM	–	0.482	0.479
Transformer-CH	CH-E+CH-S	BM	Pitch	0.476	0.476
Transformer-CH	CH-E+CH-S	BM	Energy	0.477	0.476
Transformer-CH	CH-E+CH-S	BM	Duration	0.494	0.395
Transformer-BM	BM	BM	–	0.551	0.540
Transformer-BM	BM	BM	Pitch	0.477	0.460
Transformer-BM	BM	BM	Energy	0.539	0.522
Transformer-BM	BM	BM	Duration	0.553	0.548
Transformer-BM	BM	CH-E+CH-S	–	0.478	0.478
Transformer-BM	BM	CH-E+CH-S	Pitch	0.456	0.454
Transformer-BM	BM	CH-E+CH-S	Energy	0.493	0.478
Transformer-BM	BM	CH-E+CH-S	Duration	0.516	0.484

Table 3: Comparing our custom models' accuracy and F1 score across subsets of the entire prosodic feature set. The first row for each model train/test configuration, where no features are excluded, denotes the performance of that model on the entire feature set, as originally shown in Table 2, and serves as the baseline for that configuration.

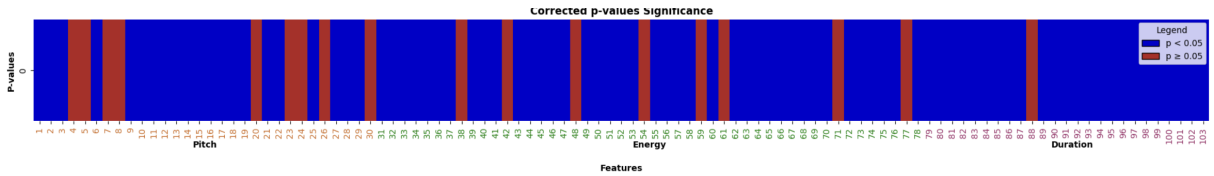
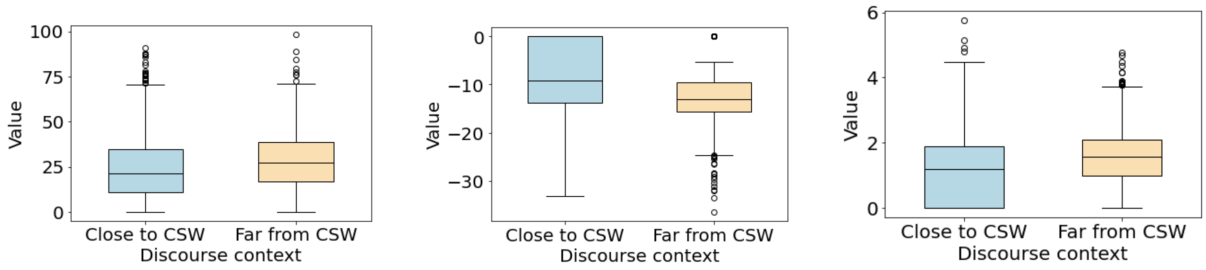


Figure 6: p-values for prosodic feature distribution comparisons between monolingual BM utterances close to and far from multilingual discourse context. Blue indicates feature-level statistical significance; red indicates insignificance. Orange ticks indicate pitch features, green ticks indicate energy features, and pink ticks indicate duration features.

is supported by a qualitative inspection whose resulting patterns mirror those found in Section 5.1; monolingual utterances that are close to the multilingual context generally have lower fundamental frequency feature values, higher energy feature values, and lower voicing rate and pause duration ratio than monolingual utterances that are further away from the multilingual context (see Figure 7 for selected visualizations). So, increased proximity to the multilingual discourse context results in further variation in the prosodic profile of monolingual speech uttered in a broadly multilingual setting, for both English and Spanish. In sum, these results strengthen our finding on the variation in prosodic profile of monolingual speech by discourse context, while adding another dimension of insight as to the specific contribution of a multilingual discourse context to such prosodic differences.

6 Discussion

We examine how the prosody of monolingual English and Spanish varies between monolingual and multilingual discourse contexts. Most studies of prosody have focused on monolingual contexts; our work explores how multilingual environments influence even monolingual speech, a research direction that is original and under-explored, in a way that bridges aspects of sociolinguistics and computational modeling. We find and interpret consistent statistically significant differences in the prosodic profile of both languages in distinct language settings, which become more pronounced with increased proximity of monolingual utterances to multilingual discourse, i.e. code-switching. Since statistical relationships in data inform model behavior, we apply the contextual differences we find toward interpreting the performance of end-to-end predictive LID models; our models seem to be able



(a) (b) (c)

Figure 7: Visualizing trends in (a) mean F0, (b) mean energy of initial unvoiced segments, and (c) voicing rate in utterances close to and far from a multilingual, i.e. code-switched (CSW), discourse context.

to indirectly learn these significant prosodic differences, though further work is required to improve their absolute task performance. Our use of LID models for diagnostic rather than predictive purposes is novel and under-utilized in prosody research, and enables quantitative interpretation of latent variation, allowing us to show that our models can internalize context-driven prosodic cues.

We conclude that the multilingual nature of the discourse context has a meaningful influence on the prosodic profile of both monolingual English and monolingual Spanish. It would be valuable to examine how our findings on prosodic differences across discourse contexts affect real-world speech systems, including their impact on the performance of speech recognition, synthesis, or speaker identification, thereby informing and guiding practical improvements in multilingual downstream applications; we plan to do so in future work that builds directly on the present study.

The main novelty of our work is based on our exploration of the broader discourse context in relation to prosody, and our largely language-agnostic[6] approach that leverages LID models to do so. In addition, we focus on speech produced by native speakers and, differing from prior work, on prosodic characteristics *across* types of utterances as opposed to *specific* speech acts, at a higher level of abstraction than any individual prosodic feature. These conditions in our research question and methodology enable us to work toward the goal of extending our research to additional languages, including lower-resourced ones, in the future, an objective that would be difficult to achieve if only considering stand-alone, highly language-specific measures of prosody, e.g. any of intonational contour, pitch accent, or lexical tone *alone*. Our work contributes novel insight and a nuanced understanding of the subtleties of prosodic production in distinct discourse contexts, which can be further developed in future studies investigating other languages from different typological families, accented and dialectal monolingual speech, as well as code-switched speech.

Limitations

Our work focuses on only two languages spoken within the United States. Fruitful extensions of our work would examine additional languages and cultural contexts to verify the robustness of our findings. While the corpora that we use were recorded in a single country, we note that this does not imply a lack of linguistic and cultural diversity therein. Across the three data sets, represented speakers are from over two-thirds of the 50 U.S. states, among which there is notable state-level and regional diversity in linguistic patterns and culture, in both spoken English and spoken Spanish. Each corpus also consists of speakers of a variety of ages. These factors combined indicate the extent of linguistic and cultural diversity of the speech we analyze.

To prevent our analyses from becoming overly complex, we assume that the *distribution* of dialects and accents of both English and Spanish across the CallHome and Bangor Miami corpora are comparable, since all three data sets are collected in the same country and have broad overlap in the Latin American origin countries of represented speakers. We acknowledge that our assumption may not necessarily hold true in the face of more fine-grained regional differences or other

[6]In using this term, we refer specifically to the nature of the input features to our models. These prosodic input features are suprasegmental and are not tied to any single language's phonemic or lexical structure. This is in contrast to much existing work on prosody that is inherently language dependent, through the study of highly language-specific intonational contours, pitch accents, or lexical tone, e.g. Delattre et al. (1962); Torres (2024).

hidden variables that could lead to undiagnosed corpus mismatches. However, we note that it is required to enable the best use of currently available resources, given the scarcity of large-scale, openly-available monolingual *and* multilingual data in two languages, with both spoken in the same accent/dialect. We believe this was a reasonable design choice for this work, which represents a first step toward studying prosody across discourse contexts, and we highlight that the effects of dialectal variation are an important future direction of research on prosody in such contexts.

Separately, there may be interesting associations between the prosodic differences we find and phonological and/or pragmatic concepts from speech literature. We omit any such analyses in the present study due to lack of relevant annotations rendering this out of scope, but believe this would be another fruitful direction for future work.

Our study intentionally focuses on discourse-level context as a driver of prosodic variation, rather than on the linguistic mechanisms (e.g. syntactic structure, lexical content, or phonological phrasing) that often require fine-grained annotations. Enriching the data sets under investigation with such annotations would allow for a more detailed analysis of how prosodic cues relate to underlying linguistic form and could potentially provide insight into the structural mechanisms underlying or interacting with prosodic differences in varied discourse contexts. However, large-scale, multilingual corpora with aligned syntactic or phonological labels are rare, and our aim in the present study was to assess whether prosodic traces of multilingual context are detectable without relying on such detailed supervision. We view our work as complementary to linguistically annotated studies and hope it will motivate future research that integrates discourse context with fine-grained linguistic structure in more richly labeled corpora.

Ethics Statement

This study was conducted exclusively on secondary data, and did not require human experiments. We did not access any information that could uniquely identify individuals within each corpus, as the original authors de-identified all speakers as outlined in the documentation of each data set. We did not collect any of the data used in this work, but note that all participants in the corpora had explicitly consented to sharing the data analyzed in our study.

We believe this work is important and highly relevant in a globalized world where multilingualism is growing in prevalence. We hope our work will lead to further study of multilingual prosody, possibly in a communication oriented framework.

Acknowledgements

*DB and DS are equal contributors to this work and share first authorship. This work was carried out while DS and MM were visiting researchers at Columbia University.

We thank Jie Chi for several helpful discussions and feedback. This work was partly supported by the National Science Foundation under Grant IIS 2418307.

References

Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu. 2011. Language identification: A tutorial. *IEEE Circuits and Systems Magazine*, 11(2):82–108.

Utpal Bhattacharjee and Kshirod Sarmah. 2013. Language identification system using MFCC and prosodic features. In *2013 International Conference on Intelligent Systems and Signal Processing (ISSP)*, pages 194–197.

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.

J. Donald Bowen. 1956. A Comparison of the Intonation Patterns of English and Spanish. *Hispania*, 39(1):30–35.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech LDC97S42.

Alexandra Canavan and George Zipperlen. 1996. CALLHOME Spanish Speech LDC96S35.

Jennifer Cole. 2015. Prosody in context: a review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

Madalena Cruz-Ferreira. 1999. Prosodic mixes: Strategies in multilingual language acquisition. *International Journal of Bilingualism*, 3(1):1–21.

Fred Cummins, Felix Gers, and Jürgen Schmidhuber. 1999. Language identification from prosody without explicit features. In *6th European Conference on Speech Communication and Technology*, pages 371–374.

Anne Cutler, Delphine Dahan, and Wilma Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40 (Pt 2):141–201.

Najim Dehak, Pierre Dumouchel, and Patrick Kenny. 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103.

Pierre Delattre, Carroll Olsen, and Elmer Poenack. 1962. A Comparative Study of Declarative Intonation in American English and Spanish. *Hispania*, 45(2):233–241.

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

Hongwei Ding and Rüdiger Hoffmann. 2015. *An Investigation of Prosodic Features in the German Speech of Chinese Speakers*, pages 221–241. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rosemary Graham. 1978. Intonation and Emphasis in Spanish and English. *Hispania*, 61(1):95–101.

Taniya Hasija, Virender Kadyan, Kalpna Guleria, Abdullah Alharbi, Hashem Alyami, and Nitin Goyal. 2022. Prosodic feature-based discriminatively trained low resource speech recognition system. *Sustainability*, 14(2).

Evia Kainada and Angelos Lengeris. 2015. Native language influences on the production of second-language prosody. *Journal of the International Phonetic Association*, 45(3):269–287.

Anatoliy Nikolaev, Evgeniy Parfenov, and Ivan Artemiev. 2015. Foreign language accent and prosody in the context of cross-cultural multilingualism. *Mediterranean Journal of Social Sciences*, 6.

Page Piccinini and Marc Garellek. 2014. Prosodic Cues to Monolingual versus Code-switching Sentences in English and Spanish.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research (JMLR)*.

K.S. Rao, V.R. Reddy, and S. Maity. 2015. *Language Identification Using Spectral and Prosodic Features*. SpringerBriefs in Speech Technology. Springer International Publishing.

Andrew Rosenberg, Erica Cooper, Rivka Levitan, and Julia Hirschberg. 2012. Cross-language prominence detection. In *Speech Prosody 2012*, pages 278–281.

J.-L. Rouas, J. Farinas, F. Pelligrino, and R. Andre-Obrecht. 2003. Modeling prosody for language identification on read and spontaneous speech. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 1, pages I–753.

Katrina Schack. 2000. Comparison of intonation patterns in Mandarin and English for a particular speaker.

Elaine Schmidt and Brechtje Post. 2015. *Language Interaction in the Development of Speech Rhythm in Simultaneous Bilinguals*, pages 271–291.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody.

Catalina Torres. 2024. Portuguese and German Intonation Contours in a Two-Way Immersion School. *Languages*, 9(2).

Juan Vasquez, Juan Rafael Orozco, Tobias Bocklet, and Elmar Noeth. 2018. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of Communication Disorders*, 76.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michael Wagner and Duane G. Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945. PMID: 22096264.

Yi Xu. 2011. Speech prosody: a methodological review. *Journal of Speech Sciences*, 1(1):85–115.

A Appendix

A.1 Notes on the Bangor Miami corpus.

As shown in Table 1, the BM corpus consists of 84 unique speakers across 56 dialogues. One important idiosyncrasy of the data set is that 15 of the 56 dialogues involve the same speaker ("Maria") in conversation with different interlocutors. The justification for this data collection design choice can be found in the original corpus documentation (Deuchar, 2011). Relatedly, "Maria" tends to speak very long utterances that represent outliers in the corpus in terms of utterance length. Excluding this speaker from utterance length statistics, the corpus-level mean and standard deviation are 1.89s and 1.45s respectively.

Below, we share examples of monolingual English and monolingual Spanish utterances, on which we conduct our analyses in this work, and contrast these with examples of Spanish-English code-switched utterances, which we do not use in the present study.

- **Monolingual English Example 1:** "So we went to the Heat game."
- **Monolingual English Example 2:** "You know, I'm just saying, I'm nice to people in general."
- **Monolingual Spanish Example 1:** "*Ay, qué estúpida.*"
- **Monolingual Spanish Example 2:** "*Vamos a ver qué dice pues.*"
- **Code-switched Spanish-English Example 1 [not used in any analysis in this work]:** "*Pero mi* printer doesn't work."
- **Code-switched Spanish-English Example 2 [not used in any analysis in this work]:** "*Pero no la puedes hacer* because you can't start checking it."

A.2 Glossary of selected prosodic features.

- **Voiced F0.** This refers to the fundamental frequency, i.e. perceptual pitch, of voiced segments of speech, which primarily involve vowels and certain groups of consonants. It is measured in terms of the rate of vibration of vocal folds, and can be estimated using signal processing techniques that leverage linear approximation.

- **Tilt.** When vocal folds vibrate during speech production, this produces a complex sound wave with a fundamental frequency as well as a series of higher frequencies known as harmonics. Tilt, or spectral tilt, describes how the energy of these higher frequency harmonics is distributed across the frequency range. Greater tilt indicates a greater difference in amplitude between lower and higher harmonics, and is perceived as a breathier voice.

- **Voicing rate.** This refers to the rate at which a speaker's vocal cords vibrate during the production of voiced sounds in speech.

A.3 Confusion matrices associated with models in Section 5.2.

	Predicted: P	Predicted: N
True: P	2481	2132
True: N	1633	2980

Table 4: Confusion matrix for Transformer-CH model tested on CH-E and CH-S. P and N refer to positive and negative classes, respectively.

	Predicted: P	Predicted: N
True: P	975	1397
True: N	1063	1309

Table 5: Confusion matrix for Transformer-CH model tested on BM.

	Predicted: P	Predicted: N
True: P	2189	2424
True: N	2393	2220

Table 6: Confusion matrix for Transformer-BM model tested on CH-E and CH-S.

	Predicted: P	Predicted: N
True: P	1663	709
True: N	1422	950

Table 7: Confusion matrix for Transformer-BM model tested on BM.

	Predicted: P	Predicted: N
True: P	4090	523
True: N	1162	3451

Table 8: Confusion matrix for Facebook-MMS model tested on CH-E and CH-S.

	Predicted: P	Predicted: N
True: P	1711	661
True: N	248	2124

Table 9: Confusion matrix for Facebook-MMS model tested on BM.

A.4 Replications using sampling-rate-matched utterances.

Below, we present additional results for the sampling-rate-matched data.

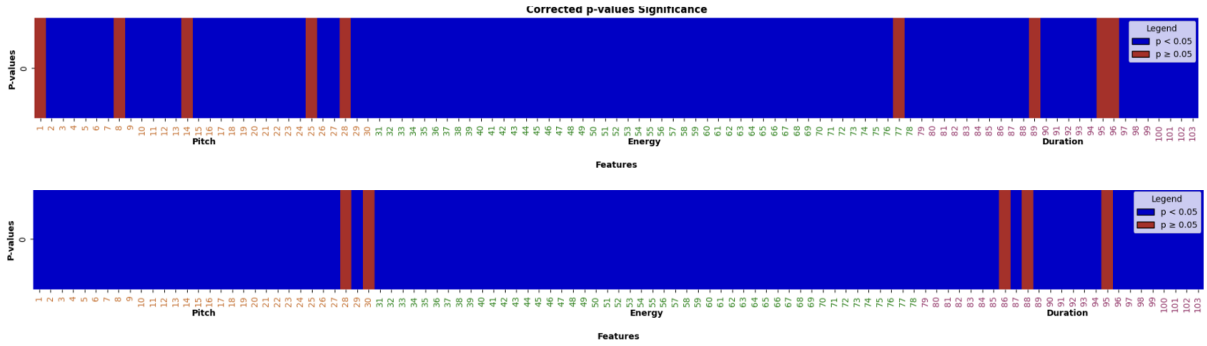


Figure 8: p-values for prosodic feature distribution comparisons between CallHome and downsampled Bangor Miami over length-controlled utterances (<=7 seconds): English (top) and Spanish (bottom). Blue indicates feature-level statistical significance; red indicates insignificance. Orange ticks indicate pitch features, green ticks indicate energy features, and pink ticks indicate duration features.

Model	Train set	Test set	Accuracy	F1 Score
Transformer-CH	CH-E+CH-S	CH-E+CH-S	0.581	0.576
Transformer-CH	CH-E+CH-S	BM	0.472	0.466
Transformer-BM	BM	CH-E+CH-S	0.467	0.467
Transformer-BM	BM	BM	0.557	0.546

Table 10: Comparing model accuracy and F1 score on monolingual versus multilingual discourse context train/test configurations, using sampling-rate-matched audio data. Baseline accuracy associated with random/blind guessing in each case is 0.5. These results are consistent with those presented in Table 2 and provide further evidence of the lack of channel-driven effects in a modeling setting.