Too few interruptions? Using data augmentation to improve offline automatic turn-taking annotation

Autorías: Gravano, Agustín; Gallo, Tomás; Molina, Nadia Guadalupe; Oppenheim, Abi; Sneider, Jazmín

Fecha: 26-29/05/2026

Publicado originalmente en: Speech Prosody 2026, Decimotercera conferencia internacional sobre prosodia del habla. Philadelphia, Pennsylvania, USA

¿Cómo citar este documento?
Gravano, A., Gallo, T., Molina, N.G., Oppenheim, A. Sneider, J. "Too few interruptions? Using data augmentation to improve offline automatic turn-taking annotation", accepted for presentation in Speech Prosody 2026, Philadelphia, PA. https://doi.org/10.21437/SpeechProsody.2026-48

Speech Prosody 2026
26-29 May 2026, Philadelphia, Pennsylvania, USA

Too few interruptions? Using data augmentation to improve offline automatic turn-taking annotation

Agustín Gravano[1,2,3], *Tomás Gallo*[2], *Nadia Guadalupe Molina*[2], *Abi Oppenheim*[4], *Jazmín Sneider* [2]

[1]Laboratorio de Inteligencia Artificial, Universidad Torcuato Di Tella, Buenos Aires, Argentina
[2]Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires, Argentina
[3]Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina
[4]Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina

agravano@utdt.edu, {tomas.gallo,nmolina,jsneider}@mail.utdt.edu, aoppenheim@dc.uba.ar

Abstract

Offline turn-taking annotation consists in classifying all transitions in a spoken conversation into several categories, such as smooth switches, backchannels, and interruptions. Previous research has reported low accuracy in identifying interruptions, possibly due to their infrequent occurrence in spontaneous spoken dialogue, resulting in a scarcity of data to effectively train machine-learning models. In this study, we explore three strategies to increase the number of interruptions available in existing corpora: 1) create copies of actual interruptions and subtly alter their acoustic-prosodic characteristics; 2) generate artificial interruptions at hold transitions, which are known to be prosodically similar to the speech preceding interruptions; and 3) combine the first two strategies. We report promising improvements in classification performance when using these data augmentation techniques.

1. Introduction and previous work

Spontaneous conversation is more than just a sequence of orderly sentences; it thrives on rich interactions including smooth and overlapping speaker transitions, backchannels, interruptions, and failed interruptions, all contributing to complex dialogue dynamics. Observing these elements can provide insights into the conversation's development and outcome; for example, a teacher might interpret a lack of backchannels as a student struggling, while a customer representative might see repeated interruptions as signs of customer frustration. Identifying these turn exchanges can offer valuable paralinguistic information that deepens our understanding of conversation flow. Thus, the task of automatically annotating turn-taking transitions could facilitate the analysis of large conversation volumes, for example for addressing sociolinguistic questions and for analyzing call-center interactions, among other practical applications.

This task can occur in two distinct settings: online and offline. Research has mostly concentrated on the online setting, where transitions are classified in real-time using information available up to what Sacks, Schegloff y Jefferson (1974) identified as a transition relevance place — a moment when a speaker change might happen [1]. Such an online task is typical for spoken dialogue systems, which must continuously decide whether to respond or keep listening. There is a vast amount of literature on this setting; e.g. see Skantze (2021) for a review [2]. In contrast, an offline setting gives access to the complete conversation, enabling the use of broader contextual cues around each transition, including adjacent transitions, speaker activity patterns, and remaining conversation time. Surprisingly, this setting has been explored only sparsely.

We build upon the work by Brusco and Gravano (2023), who experimented with different machine-learning techniques for classifying all turn-taking transitions in a conversation into several categories, including smooth switches, backchannels, and interruptions [3]. Their best-performing models consistently showed poor performance for interruptions. A good candidate to explain these results is the low number of instances of interruptions present in the three datasets used in that work, in American English, Slovak, and Argentine Spanish.

In general, interruptions are known to be relatively infrequent compared to other types of turn transitions. As noted by [1], participants tend to avoid interruptions and prefer to wait for natural transitions. Consistently, [4] found in their cross-cultural study that interruptions are less common, with smooth transitions being more characteristic of natural dialogue.

The few interruptions available in the dialogues may provide insufficient training material for the machine-learning models. In particular, **prosodic variability** is likely underrepresented in the data, making it difficult for the models to learn the prosodic patterns that characterize different turn-transition types, including turn-yielding, turn-holding, and backchannel-inviting cues. These prosodic signals have been extensively studied in the literature and are typically described as utterance-final differences in pitch, intensity, speech rate, voice quality, and pitch contour [5, 6, 7, 8], and they appear to be relatively stable across European languages [7, 9, 10]. This limitation is reflected in the feature-importance rankings reported in [3], where acoustic features appear much lower than expected. In short, the models do not have enough data to adequately learn the prosodic properties associated with the different transitions.

To mitigate this issue, [11] experimented with different oversampling techniques, but without much success. Thus, in the present paper we seek to explore the question of what would happen if we had a greater quantity and variety of interruptions. We simulate having more interruptions by means of three data augmentation techniques. In this way, we generate new instances of interruptions, retrain the machine-learning models, and compare the results, with the objective of improving the classification rates of the models from [3] and [11].

In what follows, Section 2 describes the corpus of task-oriented spontaneous dialogues in Argentine Spanish that we used in our experiments. In Section 3, we describe the machine-learning framework: algorithms, feature sets and other relevant details. Section 4 presents the three data augmentation strategies — the main contribution of this paper. Section 5 shows the improvements obtained by using each strategy, and Section 6 discusses the main findings.

10.21437/SpeechProsody.2026-48

2. Corpus

The UBA Games Corpus is a collection of dyadic conversations in Argentine Spanish that were recorded following the Objects Games paradigm [12, 8, 13]: each pair of subjects participated in a series of collaborative computer games that require verbal interaction. Each subject had a laptop displaying a board with 5-7 objects. One of the subjects (the Describer) described the position of a target object to their partner (the Follower), whose job was to place that same object in the exact location on their own screen. These task-oriented dialogues closely resemble real-world settings such as call centers or teaching lessons that also have two distinct speaker roles — roughly a leader (a customer service representative, an instructor) and a follower (a customer, a student).

Each Objects Game session consists of a series of 14 tasks, with subjects alternating roles as Describer and Follower. The subjects' speech is not restricted in any way, and they were awarded points on a 1-100 scale according to how well the Follower managed to locate the target object. Sessions were recorded in a soundproof booth, each participant using a laptop, and separated by an opaque curtain so that communication was entirely verbal. At the end of a recording session, subjects were rewarded with a modest monetary compensation for their time and an extra bonus proportional to their awarded points. We used only the first batch of sessions of this corpus, which was collected in 2012 at the University of Buenos Aires (UBA). 14 subjects participated (7 female, 7 male), aged between 19 and 59 years (M = 28.6, SD = 12.7). All subjects agreed to participate by signing a consent form, were native speakers of Argentine Spanish, and lived in the Buenos Aires area at the time of the study. Each session had exactly 14 Objects Games tasks, totaling 6.4 hours of dialogue in this first batch.

Following [8], an **inter-pausal unit** (IPU) is defined as a maximal sequence of words surrounded by silence longer than 50 ms, and a **turn** is a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. A **turn-taking transition** is defined as an event in which an IPU from the current speaker is immediately followed by either a new IPU from the same speaker or an IPU from the interlocutor (with or without overlap).

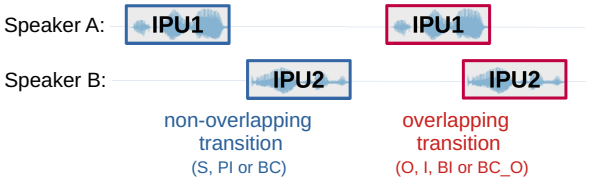

Figure 1: *Schematic illustration of a non-overlapping (left) and an overlapping (right) turn-taking transition from IPU1 to IPU2, produced by speakers A and B, respectively.*

All turn-taking transitions in the corpus were manually labeled by two trained annotators following the guidelines described in [8] and briefly summarized next. These annotation guidelines distinguish between two disjoint transition categories (see Figure 1) — with no overlap (when the interlocutor starts speaking after a short pause) and with overlap (when the interlocutor starts talking before the current speaker's IPU has finished). Transitions with no overlap include:

- *Smooth switch* (S): the current speaker completes their utterance; the interlocutor speaks after a pause;
- *Pause interruption* (PI): the current speaker does not complete their utterance; the interlocutor speaks after a pause;
- *Backchannel* (BC): after a pause, the interlocutor produces an utterance such as "yeah" or "uh-huh" to show attention and invite the current speaker to continue.

Transitions with overlap include:

- *Overlap* (O): equivalent to S, but with overlapping speech;
- *Backchannel with overlap* (BC_O): equivalent to BC, but with overlapping speech;
- *Interruption with overlap* (I), equivalent to PI, but with overlapping speech;
- *Butting-in* (BI): similar to I, but speaker B does not manage to take the floor (i.e., this is an unsuccessful interruption).

Note that the difference between S and PI lies in whether the current speaker has completed their utterance. This is typically reflected prosodically through the presence of turn-yielding cues in S and turn-holding cues in PI. In this sense, S and PI are opposites, and they are particularly challenging for a model to distinguish. The same applies to O and I.

The corpus has 2043 instances of S, 664 O, 845 BC, 86 BC_O, 257 PI, 192 I, and 83 BI. In addition to these seven main categories, there are other special transitions, including Hold (H) and X2, that occur when the current speaker continues speaking after a short pause or after a backchannel, respectively. These special labels are described in detail in [13] and are omitted in the present study, since they are almost always trivial to annotate automatically.

3. Machine-learning framework

For our experiments, we took as our starting point the automatic turn-taking transition classification models introduced in [3] and [11]. There, the authors explored two **algorithms**: random forest (RF) and recursive neural networks (RNN). In all cases, RNN models outperformed RF, but at the expense of a higher training cost. In the present study we choose to use the best-performing configurations of RF models, since they allow for greater exploration given their lower computational costs. It is reasonable to assume that the results presented here could also be improved using RNN, but we are interested in focusing on finding out what *relative* improvements may be achieved by means of data augmentation techniques.

Regarding the instance **features**, we use the set that was reported to lead to the best results. These include the following, computed for IPU1 and IPU2 (see Figure 1): IPU and turn durations in seconds, speech rate (in words/sec and phonemes/sec), speaker gender and role, and number of IPUs in the turn. The feature set also includes standard acoustic features, aimed at capturing the prosodic variation found around turn transitions: mean and stdev of pitch, intensity, logHNR, jitter and shimmer — all computed over the final 1000 ms of IPU1 and the initial 1000 ms of IPU2; as well as pitch and intensity slopes — computed over the final 250, 500, 750 and 1000 ms of IPU1.

Acoustic features are extracted from the audio signal using the open-source toolkit openSMILE (version 2.2) [14]. This tool extracts low-level audio descriptors by using a 50 ms sliding window with a 10 ms step, thus producing a time series for each acoustic feature. These five time series are then speaker-standardized using z-scores and finally aggregated into means, stdevs and slopes as indicated in the previous paragraph.

Two tasks per session and three complete sessions are separated as a **held-out set**. The remaining data (the **development set**) is used for running the cross-validation experiments described in the following sections. With this setup, the held-out set contains speakers not present in the development set, as well

as additional tasks for speakers who are. To address the issue of class imbalance in the data set, an **oversampling** technique is used which consists in replicating instances in each training fold until all classes have the same counts.

Our **initial RF model (INIT)** reaches a macro F1 score of .597 on the development set. F1 scores for individual labels are shown in Table 1. The performance is relatively low for interruptions (I, PI, BI), which may reasonably be attributed to the small number of instances available in the dataset for training. In the rest of the present paper, we describe three data augmentation techniques for addressing this issue. We then train identical models on the augmented data, and compare their performance against the initial model.

4. Data augmentation strategies

In this section, we present three methods for generating new interruptions. The first perturbs the acoustic characteristics of existing instances, the second transforms hold transitions into interruptions, and the third combines both approaches.

4.1. AUG1: Perturbation of acoustic features

Our first data augmentation strategy (AUG1) consists in creating new copies of each instance of I, PI and BI, and randomly altering their acoustic characteristics. As explained in Section 3, we consider five acoustic features: pitch, intensity, logHNR, jitter and shimmer. For each feature, we extract two time series (from the final 1000ms of IPU1; and from the initial 1000ms of IPU2) by sliding a 50ms window with a 10ms step, and speaker-standardize each time series.

At this point, we multiply each entire time series by a constant $\alpha \approx 1$ chosen at random, which allows the overall shape of the curve to be preserved while modifying its scale in a controlled manner. We choose values of α close to 1 in order to generate mild, plausible variations in pitch, intensity and voice quality. We emphasize that each feature is multiplied by a different α, thus producing variations among features within the same instance. The values for α are sampled uniformly from $[0.7, 1.3]$, which yielded the best results after experimenting also with $[0.6, 1.4]$, $[0.8, 1.2]$, and $[0.9, 1.1]$. After scaling each time series, we aggregate it into the same features described in above: means, standard deviations, and slopes.

Using the AUG1 strategy, we generate as many instances as necessary of the target classes (I, PI, BI) to match the number of instances of the most frequent class in the training set.

4.2. AUG2: Removal of turn-final IPUs

Our second data augmentation strategy (AUG2) is inspired by one of the conclusions of [15]: "in task-oriented dialogue, interruptions usually take place during or after IPUs that are more similar to turn-medial IPUs than turn-final ones" (p.858). In other words, turn-medial pauses are good candidates for interruption attempts to occur. Following this statement, our second technique consists in generating new interruptions by trimming conversation excerpts, so as to simulate an interruption where originally there was a hold transition (H). This procedure may be used for generating either new instances of PI (pause interruptions, with no overlapping speech) or I (simple interruptions, with overlapping speech).

Figure 2 illustrates the procedure for creating a new instance of PI. First we find IPUs 1 and 2 from speaker A, followed by a smooth switch (S) to IPU 3 from speaker B. We proceed to delete IPU 2, and slide IPU 3 to the left. If we assume

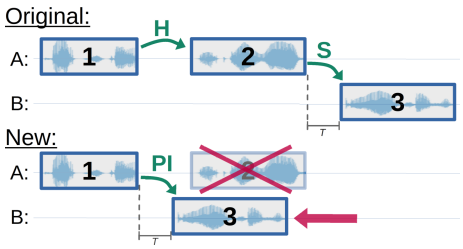

Figure 2: ***Generation of a new pause interruption (PI).*** *Originally, there are two IPUs from speaker A followed by a smooth switch transition to speaker B. We create a new instance of PI by removing A's rightmost IPU and moving B's IPU to the left.*

that A's turn is now incomplete (because we have removed its final IPU), then B's new turn is an interruption. Thus, this methods creates a new instance of PI that replaces the original S, and all of its features are computed as explained in the section 3.

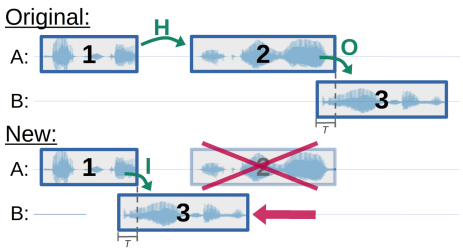

Figure 3: ***Generation of a new simple interruption (I).*** *Originally, there are two IPUs from speaker A followed by an overlap transition to speaker B. We create a new instance of I by removing A's rightmost IPU and moving B's IPU to the left.*

Analogously, Figure 3 illustrates how we generate a new instance of I. We find IPUs 1 and 2 from speaker A, followed by an overlap transition (O) to IPU 3 from speaker B. Again, we delete IPU 2 and slide IPU 3 to the left. If we assume that A's turn is now incomplete, then B's new turn is also an interruption. Thus, the new instance of I replaces the original O.

Of course, assuming that A's turn becomes incomplete after removing IPU 2 is not necessarily correct, as this segment may be an optional constituent. Still, an informal inspection of the data suggests that this approximation is sufficiently reliable.

We create with the AUG2 method 347 new instances of PI and 78 of I, increasing their total counts in the development data set from 162 to 509 and from 150 to 228, respectively. Note that we choose to remove the original instances of S and O that were used for this, since those instances would be somewhat correlated to the new ones. (As we will see below, this decision may not be optimal, as it considerably reduces the training data available for those two classes, with a corresponding impact on the detection performance.)

4.3. AUG1+2: Combination of strategies

Next, we explore how to combine the two strategies described above. We have considered several possibilities along dimensions such as the proportion of new instances generated with each strategy, and whether to apply the acoustic perturbations from AUG1 also to the instances created with AUG2. All combinations yielded comparable results, with little variation.

For brevity's sake, in the next section we present the results of one of the optimal combinations: 1) We create the same number of instances of I and PI as we did with AUG2, and we do

not apply acoustic perturbations to them. 2) We further generate new instances with AUG1 until the dataset is balanced.

5. Results

We report first the results of our experiments conducted using k-fold cross-validation on the development data set, with $k = 5$. We compare the performance of models trained on the original data (INIT) with that of models trained on data augmented using each of our three strategies (AUG1, AUG2, AUG1+2). Table 1 presents the number of instances available for cross-validation, the F1 score for each label, and the overall macro F1 score. The two labels of interest are highlighted in blue: pause interruption (PI) and simple interruption (I).

	S	O	BC	BC_O	PI	I	BI	Macro
INIT: Initial model								
F1:	.83	.71	.85	.58	.36	.41	.44	.597
N:	1393	488	527	54	162	150	61	
AUG1: Perturbation of acoustic features								
F1:	.83	.69	.85	.53	.39	.48	.46	.604
N:	1393	488	527	54	1393	488	488	
AUG2: Removal of turn-final IPUs								
F1:	.79	.70	.88	.57	.44	.44	.46	.611
N:	1046	410	527	54	509	228	61	
AUG1+2: Combination of strategies								
F1:	.79	.67	.87	.62	.45	.48	.46	.620
N:	1046	410	527	54	1046	410	410	

Table 1: *Model performance on the development set. F1 scores averaged across all CV folds, and instance counts used for CV.*

Regarding the number of instances, we first observe that, given that it is easy scalable, the AUG1 technique allows us to balance classes PI and I against their competitors S and O, yielding the same number of instances (1,393 and 488, respectively). In contrast, AUG2 enables the creation of only a limited number of new interruptions, at the cost of reducing the number of S and O instances.

In terms of performance, we find consistent improvements in the F1 scores of PI and I across all three techniques. Compared to INIT, AUG1 yields a larger gain for I (.41 to .48) than for PI (.36 to .39), whereas AUG2 shows the opposite pattern, improving PI more markedly (.36 to .44) than I (.41 to .44). Interestingly, the combined AUG1+2 technique achieves the largest improvement for both labels: PI (.36 to .45) and I (.41 to .48). In all cases, the F1 scores of the competing labels S and O decrease slightly.

Next, we assess how well these results generalize to new tasks, both for speakers present in the training data and for previously unseen speakers. We repeat the analysis of the AUG1+2 strategy on the held-out data. Table 2 summarizes the resulting per-label F1 scores as well as the overall macro F1 score.

		S	O	BC	BC_O	PI	I	BI	Macro
INIT:	F1:	.84	.76	.88	.36	.10	.25	.14	.476
AUG1+2:	F1:	.63	.61	.88	.36	.20	.32	.32	.474
	N:	553	169	262	29	51	41	22	

Table 2: *Model performance on the held-out set, and instance counts used for validation (the same in all cases).*

We again observe substantial performance gains for PI (.10 to .20) and for I (.25 to .32), but in this case the performance of the competing labels S and O shows a larger drop than in the development data. This suggests that removing the original S and O instances that were transformed into PI and I may not have been optimal, and that retaining them could be a better approach. Unlike the results on the development set, where the overall macro F1 improves with all three strategies, here the overall macro F1 remains unchanged.

In the Introduction, we hypothesized that the scarcity of interruptions in the data prevented the models from capturing the prosodic variability characteristic of turn-taking transitions. By augmenting the number of interruptions, we would expect the models to assign greater relative importance to the acoustic features. Table 3 shows how the relative importance assigned by the models to the acoustic features changes after adding new instances. The first column shows that, under all three strategies, features measuring pitch level rank higher on average. Notably, features capturing the IPU-final pitch slope rank much higher when using the AUG1 and AUG1+2 techniques. No improvement in ranking is observed for the remaining acoustic features (voice quality and intensity slope, in the third column).

	Pitch level	Pitch slope	Other acoustic
INIT	23.3	25.3	25.6
AUG1	22.4	18.1	26.6
AUG2	22.6	27.6	26.6
AUG1+2	22.6	21.6	27.3

Table 3: *Mean ranking of acoustic feature types according to the feature-importance scores of the random-forest models. Here, a lower number means a higher importance.*

6. Discussion

In this paper, we have introduced three data augmentation techniques designed to increase the number and variety of interruption instances available for training machine-learning models that classify turn-taking transitions. The results show clear benefits: all strategies improve the F1 scores of the two interruption labels of interest, PI and I. These gains indicate that the models are better able to capture the prosodic variability that distinguishes different turn-taking transitions.

The feature-importance analysis further supports this interpretation. After augmentation, acoustic feature types associated with utterance-final prosodic cues — particularly pitch level and final pitch slope — receive substantially higher rankings, suggesting that the models rely more heavily on prosodic information when sufficient training material is available. This reinforces the idea that interruptions are not only rare but also prosodically diverse, and that expanding this portion of the dataset helps models learn more informative prosodic patterns.

In this work, we have focused exclusively on prosody as a source of disambiguating cues. However, turn-taking is also shaped by linguistic content, and lexical or syntactic signals may provide complementary information about when a speaker is likely to yield or hold the floor. Exploring linguistic and multi-modal cues is an active research topic (e.g., [16, 17]).

Finally, we deliberately restricted our augmentation techniques to methods with low computational cost and minimal modeling overhead, ensuring that they are easy to integrate into existing machine-learning pipelines. However, a promising avenue for future work involves leveraging large language models (e.g., [18]) to generate fully synthetic interruptions, which could further enrich the training data available to the models.

7. Acknowledgements

The authors would like to thank Pablo Brusco and Pablo "Bauna" Nussembaum for their valuable help and guidance.

8. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[3] P. Brusco and A. Gravano, "Automatic offline annotation of turn-taking transitions in task-oriented dialogue," *Computer Speech & Language*, vol. 78, p. 101462, 2023.

[4] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon *et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[5] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[6] E. A. Cutler and M. Pearson, "On the analysis of prosodic turn-taking cues," in *Intonation in Discourse*, C. Johns-Lewis, Ed. San Diego, CA: College-Hill, 1986, pp. 139–156.

[7] A. Hjalmarsson, "On cue – Additive effects of turn-regulating phenomena in dialogue," in *Diaholmia*, 2009.

[8] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[9] A. Gravano, P. Brusco, and S. Benus, "Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish." in *INTERSPEECH*, 2016, pp. 1265–1269.

[10] P. Brusco, J. M. Pérez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in american English and Argentine Spanish." in *INTERSPEECH*, 2017, pp. 2351–2355.

[11] P. Brusco, "Estudio translingüístico de pistas del manejo de turnos en diálogos hablados," Ph.D. dissertation, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 2021.

[12] A. Gravano, J. E. Kamienkowski, and P. Brusco, "UBA Games Corpus," Consejo Nacional de Investigaciones Científicas y Técnicas, Tech. Rep., 2023. [Online]. Available: https://ri.conicet.gov.ar/handle/11336/191235

[13] P. Brusco, J. Vidal, Š. Beňuš, and A. Gravano, "A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool," *Speech Communication*, vol. 125, pp. 24–40, 2020.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[15] A. Gravano and J. Hirschberg, "A corpus-based study of interruptions in spoken dialogue," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012, pp. 855–858.

[16] E. Ekstedt and G. Skantze, "How much does prosody help turn-taking? investigations using voice activity projection models," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Edinburgh UK: Association for Computational Linguistics, 2022, pp. 541–551.

[17] Y. Lin, Y. Zheng, M. Zeng, and W. Shi, "Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 15 310–15 322.

[18] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed, and E. Dupoux, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023. [Online]. Available: https://aclanthology.org/2023.tacl-1.15/