Speech Communication 125 (2020) 24–40

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool

Pablo Brusco [*,a,b], Jazmín Vidal [a,b], Štefan Beňuš [c,d], Agustín Gravano [a,b]

[a] *Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*
[b] *Instituto de Ciencias de la Computación, CONICET-UBA. Buenos Aires, Argentina*
[c] *Constantine the Philosopher University in Nitra, Slovakia*
[d] *Institute of Informatics, Slovak Academy of Sciences, Slovakia*

ARTICLE INFO

Keywords:
Dialogue
Prosody
Slovak
English
Spanish
Machine learning

ABSTRACT

In dialogue, speakers produce and perceive acoustic/prosodic turn-taking cues, which are fundamental for negotiating turn exchanges with their interlocutors. However, little of the temporal dynamics and cross-linguistic validity of these cues is known. In this work, we explore a set of acoustic/prosodic cues preceding three turn-transition types (hold, switch and backchannel) in three different languages (Slovak, American English and Argentine Spanish). For this, we use and refine a set of machine learning techniques that enable a finer-grained temporal analysis of such cues, as well as a comparison of their relative explanatory power. Our results suggest that the three languages, despite belonging to distinct linguistic families, share the general usage of a handful of acoustic/prosodic features to signal turn transitions. We conclude that exploiting features such as speech rate, final-word lengthening, the pitch track over the final 200 ms, the intensity track over the final 1000 ms, and noise-to-harmonics ratio (a voice-quality feature) might prove useful for further improving the accuracy of the turn-taking modules found in modern spoken dialogue systems.

1. Introduction

Corpus-based computational linguistics studies have opened new opportunities for answering questions about how human dialogue flows. In addition to more recent state-of-the-art prediction techniques, data-based studies have allowed us not only to study complex speech phenomena, but also to use the resulting knowledge in the creation of more natural spoken dialogue systems. Turn-taking management is a very good area for exploring these issues in depth.

Research articles over the last decades have shown that information about what the next *turn-transition* is going to be seems to be present not only in *what* we say, but also in *how* we say it. In addition to the textual cues (lexical, syntactic, pragmatic), prosodic cues also play a role in perceiving how the dialogue will unfold. In particular, Gravano and Hirschberg (2011) identify a group of seven **turn-yielding cues** and six **backchannel-preceding cues** in American English. They compute these cues through the use of statistics over certain acoustic/prosodic features that are automatically extracted from speech signals several hundreds of milliseconds before pauses in conversations. They also provide supporting evidence for the Duncan's theory, which establishes that the sum of turn-taking cues impact on the subject's agreement on how the dialogue will unfold (Starkey and Fiske, 1977).

Nevertheless, more research is needed to better understand the **amount of information** these cues contain and the **dynamics over time** of these acoustic/prosodic cues – i.e., how informative each cue is and how its informativeness varies over time. Additionally, we know little about the **cross-linguistic validity** of these findings. Revealing aspects of how turn-taking cues affect human-human conversations in different languages may not only help the community understand the underlying process of communication, but also allow improvements in human-computer interfaces in languages in which annotated data may not be available.

In the present article, we study the similarities and differences of how acoustic/prosodic turn-taking cues are produced in three typologically different Indo-European languages: Slovak (Slavic), American English (Germanic), and Argentine Spanish (Romance). We address this task in a data-driven approach in which we use **machine learning** techniques to model turn transitions based on hours of labeled, naturally-spoken

[*] Corresponding author at: Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina.
E-mail addresses: pbrusco@dc.uba.ar (P. Brusco), jvidald@dc.uba.ar (J. Vidal), sbenus@ukf.sk (Š. Beňuš), gravano@dc.uba.ar (A. Gravano).
URL: http://www.habla.dc.uba.ar/pbrusco.com (P. Brusco).

https://doi.org/10.1016/j.specom.2020.09.004
Received 29 October 2019; Received in revised form 10 April 2020; Accepted 22 September 2020
Available online 29 September 2020

dialogue from the **Objects Games Corpus collection** – a series of conversations with no visual contact in which 38 pairs of subjects collaborate in simple object-positioning games. We expect to validate our findings by modeling data taken from the same experimental setup, and by using the exact same methodology, in all three languages. Therefore, the contribution of this work is twofold: 1) it brings novel evidence regarding the variation of turn-taking cues over time and their comparison across different languages, and 2) it provides insights about the use of machine learning as a tool for describing speech corpora.

1.1. Turn taking cues

Studies in turn-taking have traditionally been interested in the way interlocutors engage in dialogue and the dynamics of speaker change. In a seminal work, Sacks et al. (1974) propose that turn-taking allocation is controlled by a set of fixed but flexible rules that allow an indeterminate number of participants into a conversation with no interruptions or overlaps. Starkey and Fiske (1977), suggest that participants produce a number of prosodic, syntactic and even gestural cues that, in combination, contribute to the flow and naturalness of turn-taking in conversations. While some studies argue the non-relevance of acoustic/prosodic cues and claim that lexical and syntactic information are sufficient for turn-management, others (Duncan, 1974; Ford and Thompson, 2010; Ferrer et al., 2002; Wennerstrom and Siegel, 2003; Gravano and Hirschberg, 2011; Hjalmarsson, 2011; Bögels and Torreira, 2015; Ward, 2019, inter alia), including the aforementioned studies, show evidence suggesting that acoustic/prosodic cues based on pitch and duration, and syntactic features such as the position of a word in an utterance play a key role in the turn-allocation mechanism.

While many, especially earlier, studies analyze English, turn-taking management has been explored also in other languages. In Hjalmarsson (2011), the author performs a series of experiments for understanding how turn-taking cues affect the perception of the interlocutor in Swedish conversational dialogues. She reinforces the importance of the additive effect of cues on the perception of turn-transitions about to come. Even though some studies claim that cultures strongly deviate in different turn-taking systems (Watson-Gegeo et al., 1976), others argue that some kind of 'universals' exist (Schegloff, 2006). For instance, in Stivers et al. (2009), the authors analyze ten different languages and explore the variability in the response offsets in turn transitions. They arrive to a series of cross-culturally valid observations; for example, in all of the analyzed languages, speakers provide answer responses to questions significantly faster than non-answer, and confirmation answers are delivered faster than non-confirmation ones. In a cross-linguistic study in the perception of prosodic cues in Slovak and Argentine Spanish, Gravano et al. (2016) test the subjects' predictions regarding turn-taking transitions in the two languages and show that some prosodic cues provide similar information in both languages, thus contributing to the aforementioned turn-taking 'universals'. Closely following them, the present study contributes to this line of work and helps filling the gap of knowledge in turn-taking behavior though comparing Germanic and non-Germanic languages.

From a modeling perspective, recent research has shown that acoustic/prosodic features can be used for the construction of turn-transition predictive models (Skantze, 2018; Maier et al., 2017; Hara et al., 2018; Roddy et al., 2018, inter alia). For instance, in Skantze (2018) the author predicts the future speech activity in dialogues using LSTMs – recurrent neural network models especially designed to learn contextual representations from temporal series. They use both interlocutors' pitch, intensity, and spectral stability tracks together with a voicing mask every 50 ms; and, also present a system for detecting turn-taking transitions (in particular turn continuation and switches) by following heuristics for automatically labeling turn transitions based on speech activity labels. However, these techniques are still not easy to analyze in terms of what they learn, thus making the underlying knowledge base for these aspects of human-human turn-taking management inaccessible for the moment. Contrary to these approaches, we intend to use **machine learning as a descriptive tool** to obtain information about a complex phenomenon through the exploration of models built from data. In this way, we intend to facilitate further advances in the research community in discovering and validating new findings.

1.2. Chosen languages

The three chosen languages provide a good testing ground for studying which prosodic features (and their development over time) might be cross-linguistically valid and which might present language-specific cues in turn-taking management. On the one hand, the prosodic systems of the three languages differ. For example, according to Hualde (2013), the most important difference between the intonation system of Spanish (and of other Romance languages) and that of English (and of other Germanic languages) is the flexibility found in this second group of languages in the placement of the nuclear accent. In English, the position of the nuclear accent can move to indicate focus on various constituents. In Spanish, on the contrary, the position of the nuclear accent is practically fixed, and, except in cases of narrow focus, it falls on the last syllable with a lexical accent. Slovak is a prototypical example of a hybrid system, that is characteristic of the Slavic languages (e.g. Jasinskaja (2016)), and that combines the Romance and Germanic ones above: the information structure is expressed jointly by intonation and movable nuclear accent (like in Germanic) as well as by a flexible word-order and the tendency to move the focused element to the end of utterances (like in Romance). Since the location and type of pitch accents influence prosodic contours to a great extent, these differences might also participate in the predictions the contours have in turn-taking management.

On the other hand, the three languages share many characteristics in how prosody participates in information structure signalling the intentions and mutual beliefs of the speakers. Graham (1978) argues that Spanish and English share certain intonation patterns such as rising pitch at the end of (polar) questions, which is certainly common in Slovak as well. However, some narrow-focus Slovak polar questions might also be realized with a plateau, or gradual fall, following the nuclear rising pitch. Additionally, the notion of the 'continuation (rise)' in the literature on English intonation is related to the notion of 'incompleteness' in the Romance and Slavic traditions of intonation descriptions. Hence, rising pitch followed by a pause should be interpreted in a forward-looking fashion that either the speaker wishes to continue or that a response from the interlocutor is expected. While continuations in English are typically related to pitch rises, incompleteness in Slovak/Spanish has been linked to more variability in contours (e.g. Quilis, 1993; Král, 1988).

1.3. Machine learning as a descriptive tool

Typically, in the literature of turn-taking, studies that use confirmatory data analysis require a hypothesis to be specified before the design of the dataset and also need assumptions about the generation of the data by a given stochastic data model. Nevertheless, as described in Lin et al. (2007, p. 243), interpreting the results of methods outputting *p*-values or R-scores in high dimensional data can be difficult and misleading. For example, in the case of time series, researchers typically study specific pre-selected time frames or collapse the data dimension by averaging over time. In this way, they gain statistical power and avoid multiple comparisons problems at the expense of losing temporal detail and other information.

In contrast with classical statistics, machine-learning algorithmic models are built assuming data generation mechanisms are unknown. Methods such as random forests, support vector machines, and neural networks, among others, are known to produce powerful predictive models, are designed to handle variable interactions, and generally capture non-linearities in high-dimensional data.

Nevertheless, the vast majority of machine learning-based models are currently used as a powerful tool for achieving state-of-the-art results in **predicting** turn exchanges. To our knowledge, only a small number of studies observe and analyze information to allow the scientific community to **explain** the reasons of why a prediction is made. In this work, we focus our attention on an approach in which models are created to provide new evidence that may help reject or reinforce linguistic hypotheses, sometimes to the detriment of the models' prediction power.

In the past, when model interpretability was needed, simple and transparent models such as linear regression or decision trees have been used to understand complex phenomena. Yet, it is essential to clarify that up to this day, interpretable models tend to be less powerful than fully black-box models, such as neural networks, especially when enough labeled data is available. Simple models usually suffer from high bias or high variance problems. That is, models underfit or overfit the data due to their design characteristics, which leads to low predictive power or low stability in the obtained results.

However, as explained in Breiman (2001b), transparency is not the only way of getting information from machine learning models. Methods such as *Trees Impurity Importance* (Breiman, 2001a), *Permutation Feature Importance* (Breiman, 2001a), *Partial Dependence Plots* (Friedman, 2001), *LIME* (Ribeiro et al., 2016), and *Shapley Additive Explanations* (Lundberg and Lee, 2017) have allowed the exploration of fully black-box models to a certain level. The goal of these methods is exploring not the internal structure of the model but how the model generates predictions. Unfortunately, especially in the area of speech processing — in which models predict based on processing and combining high-dimensional temporal series — the interpretation of complex models still remains an open problem.

In this work we explore the use of a random forest classifier, a robust and competitive supervised classifier along with a modified version of the permutation feature importance method. First proposed by Breiman (2001a), the random forest algorithm proposes to build an ensemble of decision trees classifiers whose predictions are individually produced and then combined for a final decision. This algorithm has shown competitive prediction power with almost no tuning effort. It manages complex and non-linear relations between inputs and outputs while avoiding overfitting at the same time. Random forests are not as simple and transparent as linear models or decision trees; therefore, some effort must be put into how they are explored. See Biau and Scornet (2016) for a full description of the method.

1.4. *About this article*

The article is divided into two experiments. The first one addresses question Q1, how do the acoustic/prosodic features of speech compare in Argentine Spanish, Slovak, and American English just before a turn transition? The second one addresses question Q2, how much information do acoustic/prosodic features carry and what is their relative contribution when preceding a turn-taking transition?

Section 2 introduces the speech corpora on which we based the experiments, with particular detail on the annotations we created. In Section 3, we analyze the corpora by visualizing different acoustic/prosodic features over time and across languages; we compare the results with previous works and show the difficulties of working in high dimensional data. In Section 4, we address the problem of automatically classifying turn-taking events, paying special attention to revealing which features contain the most relevant information over time. We also test the stability of our results by varying the way features are extracted. Finally, in Section 5 we discuss the research results and present the outlines of future work.

2. Materials: the object games corpora

We used three versions of the Objects Games Corpus (first described in Gravano and Hirschberg (2011)), in American English, in Argentine Spanish, and in Slovak. In each, a collection of spontaneous task-oriented dyadic conversations elicited from native speakers playing Objects Games was gathered. Subjects were paid to play a series of collaborative computer games requiring verbal communication. Experiments took place in soundproof booths, each participant using a different laptop computer, and separated from the other by an opaque hanging curtain. The subject's speech was not restricted in any way and it was emphasized that the game was not timed.

During the game, each subject's laptop displayed a game board with 5–7 objects. Both players saw the same set of objects at the same position on the screen except for one, the target. For one player, the Describer, the target appeared in a random location on the screen; for the other, the Follower, it appeared at the bottom. The Describer was instructed to describe the position of the target object on her screen to make the Follower match the position perfectly on his own.

Subjects could discuss freely about the location of the target object. After the negotiation, they were awarded 1–100 points based on how well the location of the target object on the Follower's screen matched its location on the Describer's. Each session consisted of a minimum of 10 and a maximum of 14 instances of the Objects Game, with subjects alternating in the Describer and Follower roles. At the end of the session, subjects were paid a fixed amount of money for their participation, plus a bonus based on the number of awarded points.

The English Corpus was collected and annotated jointly by the *Spoken Language Group* at Columbia University and the Department of Linguistics at Northwestern University. A total of 13 subjects (6 female, 7 male), ages between 20 and 50 (M = 30.0, SD = 10.9), participated in the study in New York City in October 2004. Eleven of the subjects participated in two sessions on different days, each time with a different partner. All subjects were native speakers and lived in the New York City area at the time of the study. A total of 4.5 h of dialogue were recorded in the English corpus.

The Spanish Corpus was recorded at the *Laboratorio de Investigaciones Sensoriales* (Hospital de Clínicas, Universidad de Buenos Aires)[1], in November-December, 2012. A total of 14 subjects (7 female, 7 male), ages 19 to 59 years (M = 28.6, SD = 12.7), participated. All subjects were native speakers of Argentine Spanish, lived in the Buenos Aires area at the time of the study. The Spanish corpus contains a total of 6.4 h of dialogue.

The Slovak Corpus was recorded at the sound-treated room at the Institute of Informatics, Slovak Academy of Sciences over several months in 2012. A total of 11 subjects (5 females 6 males), ages 21–67 years (M = 32.6, SD = 15.7) participated. 7 of the subjects played the game twice, each time with a different partner and 4 subjects only played once. All subjects were native Slovak speakers and lived in Bratislava at the time of the study, but their dialects varied. A total of 6.3 h of dialogue were collected in the Slovak corpus.

The collection process was the same in all three languages, and we observed no remarkable differences in the resulting corpora.

2.1. *Annotations: Interpausal units and turn transitions*

A team of trained annotators orthographically transcribed the sessions at the word level and manually time-aligned each word to the speech signal. False starts, filled pauses and speech errors were also marked. Words were automatically transformed into phonetic transcriptions through the use of phonetic dictionaries in all three languages.

We define an inter-pausal unit (IPU) as a maximal sequence of words surrounded by silence longer than 50 ms. IPUs were generated by

[1] http://www.lis.secyt.gov.ar

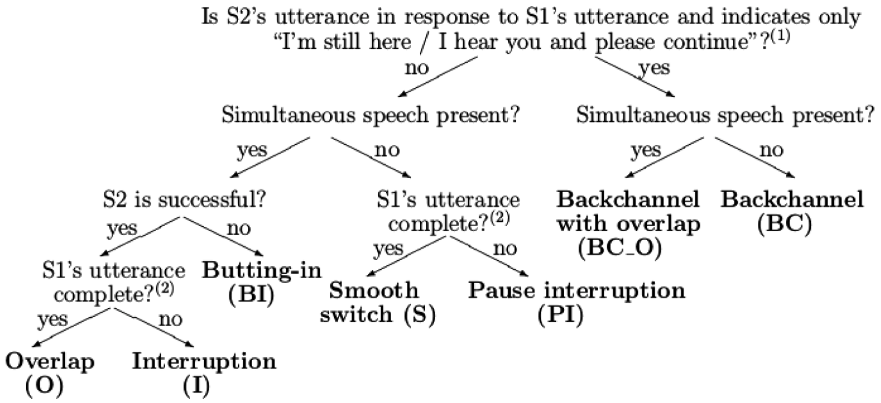

Is S2's utterance in response to S1's utterance and indicates only "I'm still here / I hear you and please continue"?(1)
no
yes
Simultaneous speech present?
Simultaneous speech present?
yes
no
yes
no
S2 is successful?
S1's utterance complete?(2)
Backchannel with overlap (BC_O)
Backchannel (BC)
yes
no
yes
no
S1's utterance complete?(2)
Butting-in (BI)
Smooth switch (S)
Pause interruption (PI)
yes
no
Overlap (O)
Interruption (I)

Fig. 1. Turn-taking labeling guidelines, as presented in Gravano and Hirschberg (2011).

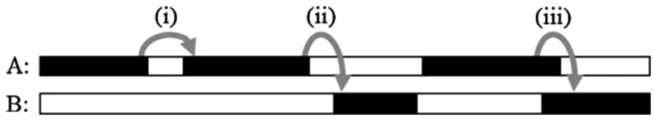


Fig. 2. Turn-transition types. Black segments represent speech; white segments, silence. The arrows show three different types of turn transitions: (i) *Hold*, when speaker A keeps the turn after a short pause; (ii) *switch, pause interruption* and *backchannel*, when speaker B produces a speech segment after a pause from speaker A; and (iii) *overlap, backchannel with overlap, interruption* or *butting-in*, when speaker B starts while the current speaker is still taking.

joining contiguous chunks of words without intermediate silences of the expected size. Next, we define a TURN as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. Turns were automatically delimited on the time-aligned orthographic annotations.

Finally, TURN TRANSITIONS were manually labeled. In each language, two trained annotators separately labeled the whole corpus following a set of instructions as shown in Fig. 1. Levels of agreement were high, reaching a Cohen's κ score (Cohen, 1960) of 0.81, 0.88, and 0.91[2] for Slovak, Spanish and English respectively. Fig. 2 shows a compact representation of the different labeled turn-exchanges.

In this work, we focus on three specific types of turn transitions, HOLD TRANSITIONS **(H)**; when the current speaker continues talking after a short pause; SWITCH TRANSITIONS **(S)**; when the interlocutor takes the floor after the speaker finishes the turn; and, BACK-CHANNELS **(BC)**; short utterances such as *yeah* or *uh-huh* used to display attention and invite the current speaker to continue. Table 1 shows the amount of IPUs extracted from each corpus for each of the turn transition types under study. We see that BC, S and H transitions represent approximately 5%, 20% and 50% of all turn transitions respectively with a larger presence of BC on the Spanish data.

3. Study 1: visualization of acoustic/prosodic features

In this first study, we address the question of how the acoustic/prosodic features of speech compare in Argentine Spanish, Slovak, and American English just before a turn transition.

To this end, we perform for each language a series of exploratory analyses with descriptive visualizations on how a number of acoustic/prosodic features behave on IPUs immediately preceding each turn-transition condition – namely, a turn exchange or switch (S), a turn continuation or hold (H), or a backchannel (BC) from the interlocutor.

Corpus-based studies on turn-taking, beyond variations in focus, tend to share a common set of prosodic and non-prosodic features. In this work, acoustic features were chosen closely following the work in Gravano and Hirschberg (2011). These include pitch, intensity, jitter, shimmer, and noise-to-harmonics ratio; and prosodic features include previous IPU's duration and speech rate. To our knowledge, only a few new features have been reported in recent articles on turn-taking and turn-ending prediction. For example, Truong et al. (2010) include pause information, Morency et al. (2010) investigate the use of multi-modal features such as gaze and transcribed speech in combination with prosody. In languages like Finnish and German, characteristics of creaky voice and glottal stops are included in Szczepek Reed (2014). From a more general perspective, Eyben et al. (2016) argues against the proliferation of brute-force parameter sets in the field. Instead, and to share standards, they propose a minimalist set of features.

However, whereas Gravano and Hirschberg (2011) model data points extracted from a given portion of the signal, this study focuses on the **temporal aspect** of such features. Our goal is to use modern visualization techniques for making a contribution to our understanding of turn-taking cues present in the acoustic signal. Also, we will explore the novel relationships between features and turn-transition types and test their **cross-linguistic validity**.

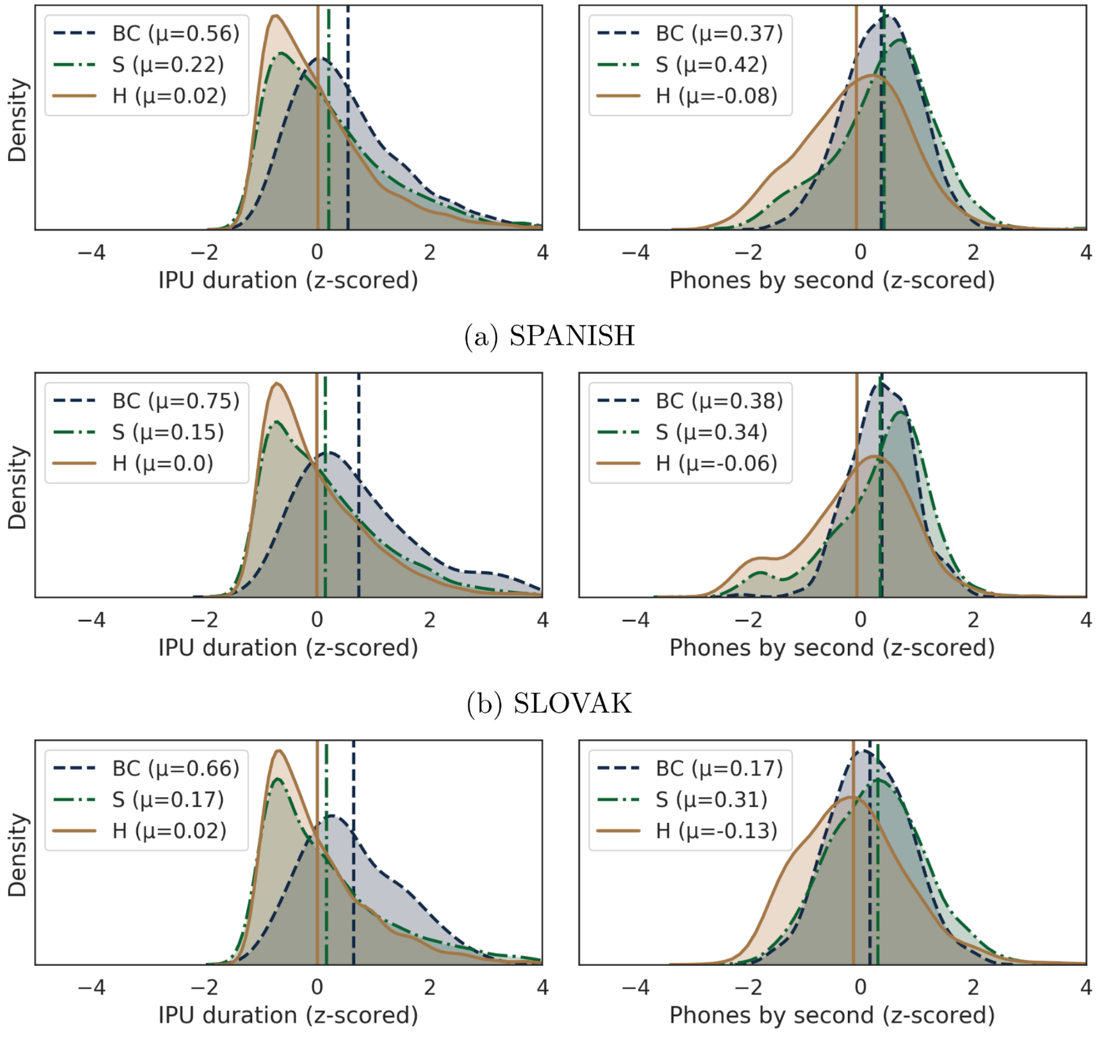
3.1. Whole-IPU and momentary acoustic features

We first extracted from the IPUs preceding each turn transition in our corpora two features that we call **whole-IPU features**: IPU DURATION, measured in milliseconds, and IPU SPEECH RATE, measured in phones per second.[3]

Skantze (2018) uses a set of "momentary" features: voice activity, pitch, intensity, and spectral stability. Following this nomenclature, we extracted what we call **momentary acoustic features**: time series (or tracks) of different acoustic/prosodic features. These features include INTENSITY, measured as the mean of squared signal values multiplied by a Hamming window; PITCH, calculated as the smoothed fundamental frequency contour; SHIMMER, the amplitude deviations between pitch periods; JITTER, the deviations in the pitch period length; and LOGHNR, the logarithm of the Harmonics-to-Noise ratio. Shimmer, Jitter and logHNR are computed only on voiced frames. To make results comparable, we normalized each feature through *z*-scores. The normalization process was performed for each speaker-session pair in our dataset, instead of

[2] The English κ score does not include the identification of backchannels, performed by different annotators as described in Gravano et al. (2007). Including the identification of backchannels is expected to reduce the agreement to numbers near the Spanish or Slovak level.

[3] It is important to note that, when an IPU has 1–3 short words, this definition of speech rate is problematic, because the effect of intrinsic phone durations might greatly affect the overall rate. In this analysis we decided not to take this problem into account in favor of having comparable data to Gravano and Hirschberg (2011).



(a) SPANISH

(b) SLOVAK

(c) ENGLISH

Fig. 3. Approximated probability density functions of *z*-scored duration and speech rate of IPUs preceding different turn-transition types. The vertical lines indicate the average value for each type. See Table 1 for the amount of data points available.

Table 1
Amount of IPUs preceding each turn-transition type. The *other* category includes overlapping transitions (O, I, BI and BC_O), non-overlapping interruptions (PI), and transitions in which there was no agreement among the annotators.

	Spanish	Slovak	English
BC	842 (7.6%)	272 (2.9%)	393 (4.6%)
S	1935 (17.5%)	1937 (20.7%)	1659 (19.6%)
H	5299 (47.8%)	4976 (53.2%)	4283 (50.6%)
other	3008 (27.1%)	2175 (23.2%)	2123 (25.1%)
total	11,084	9360	8458

isolated speakers, to avoid environmental conditions. For each feature, mean and standard deviation were computed from all the speech produced by a speaker in a particular session.

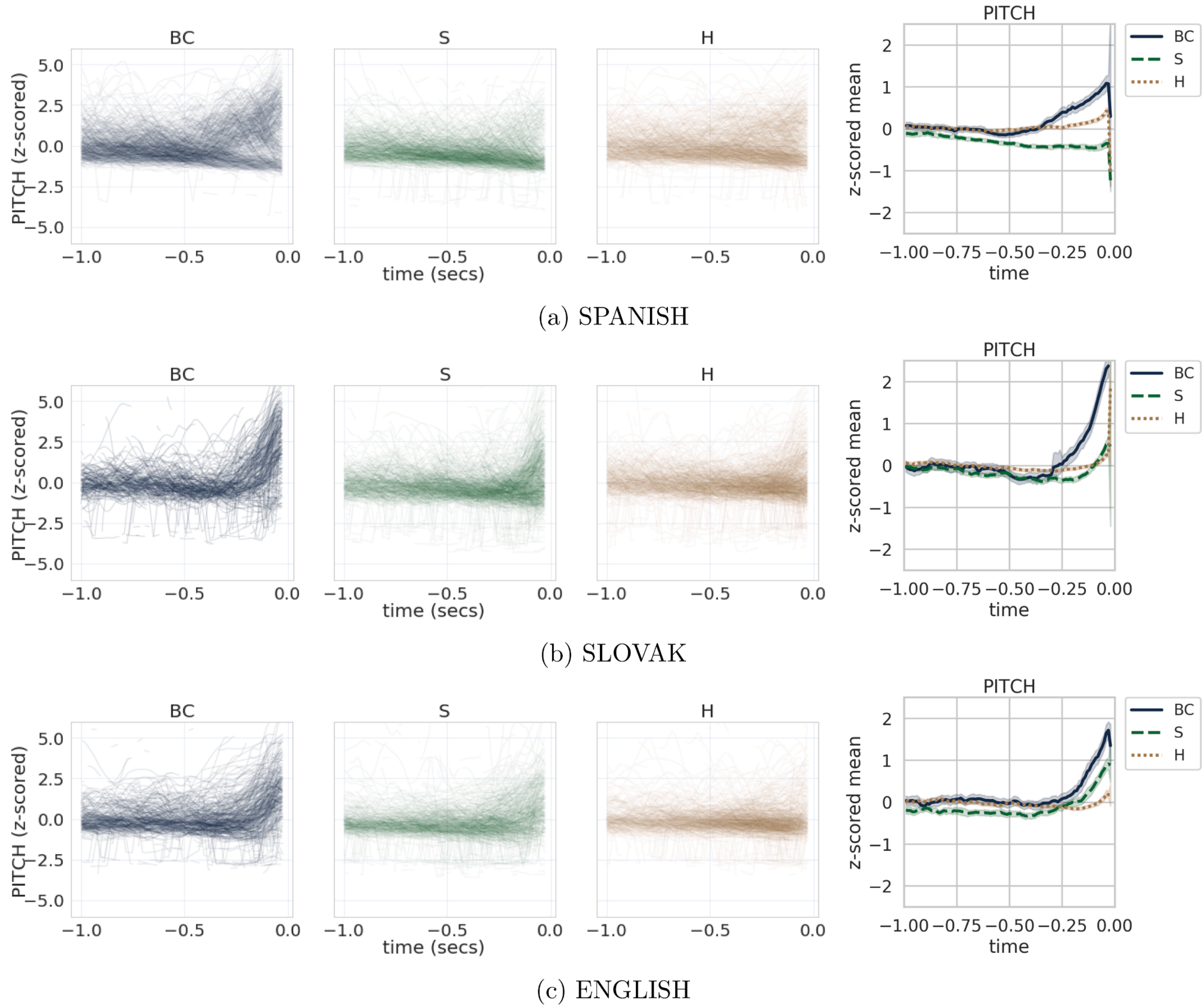
All momentary acoustic features were automatically computed using the openSMILE open-source toolkit (version 2.2) (Eyben et al., 2013). We used part of the *INTERSPEECH 2010 Paralinguistic Challenge feature set*, which contains feature tracks extracted from overlapping windows of 50 ms every 10 ms. The specific configuration file can be found at https://git.io/JvSg1. See Eyben et al. (2016) for a detailed explanation of how features are extracted under this configuration.

3.2. IPU Duration and speech rate

In this section, we examine how the set of whole-IPU features (IPU duration and IPU speech rate) compare across languages. We focus on the characteristics of IPUs preceding each turn-transition under study (H, S, BC).

Fig. 3 shows probability density approximations for IPU duration (left panel) and IPU speech rate (right panel) for each language.[4] The left plots in the figures show that IPUs preceding BC tend to be longer than IPUs preceding S; and that IPUs preceding S, in average, tend to be longer than IPUs preceding H (as shown in the vertical line). Moreover, differences in length between BC and the others get more pronounced in English and Slovak than in Spanish. The results show similarities in mean and distribution across languages, providing support for the universal hypothesis introduced in Stivers et al. (2009), and adding to their findings.

[4] Densities were estimated through Kernel Density Estimation (KDE) — a non-parametric technique that estimates the unknown probability distribution of a random variable based on a sample of points taken from that distribution. This method is known as a continuous version replacement for discrete histograms. For a more detailed explanation see Silverman (2018).



(a) SPANISH

(b) SLOVAK

(c) ENGLISH

Fig. 4. *Shadow plots* of *z*-scored pitch tracks extracted from the final second of all IPUs preceeding each turn-transition type. These plots were built using a maximum of 1000 randomized samples. All pitch tracks are aligned to IPU ending. The *average plots* on the right hand side show the pitch mean and standard deviation over time.

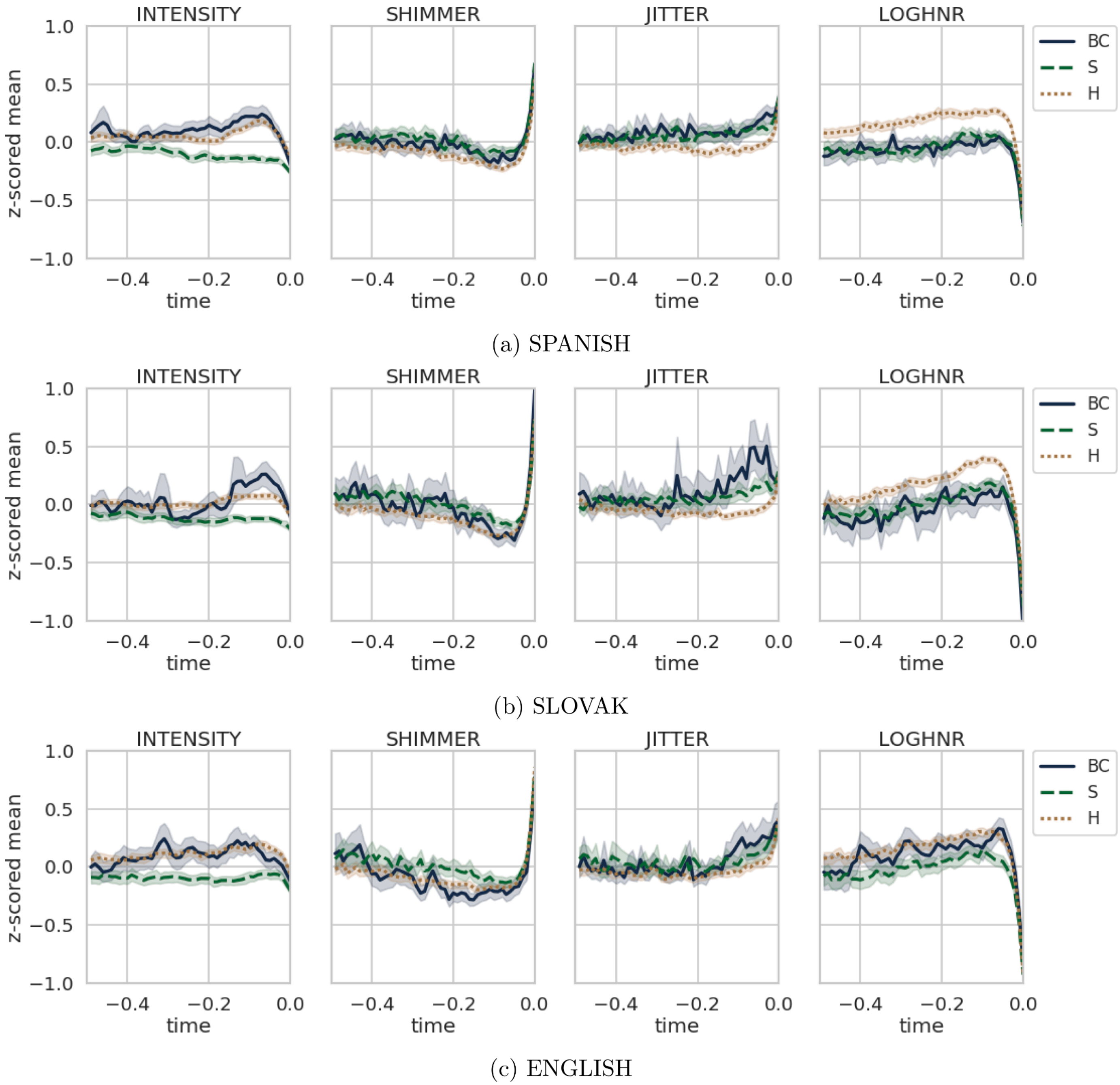
When analyzing speech rate, we see an increase for IPUs preceding S and BC with respect to H, with very similar distribution shapes across languages. This means that, before a H, speakers tend to speak slower. When listening to the data we perceive that, similar to BC, H seems to appear mostly at the end of a phrase or in points of syntactic incompleteness, but, contrary to BC, in places where speakers seem to be still planning what's coming ahead. Simply put, it is observed in all three languages that planning slows down speech right before holds.

This may be explained as a consequence of the type of interactions in play. We examine task oriented dialogues where conversations are highly structured and roles and goals are clear beforehand. On the one hand, to participate, speakers need to convey large amounts of information, such as lists of instructions and long descriptions within a more or less shared context and no access to non-verbal communication. Hence a specific pattern of language rises in which expressions tend to be fully descriptive and syntactically long making IPUs before BC longer, and BC a frequent strategy from the listener to show support to the speaker without interrupting long interventions. On the other hand, fluently switching roles and making short pauses are golden rules in verbal interactions. In fact, pauses between turns lasting more than 200 ms are considered to carry negative connotations (Sacks et al., 1974; Stivers et al., 2009; Levinson, 2016). Moreover, task oriented dialogues are hierarchically divided into units of action and incrementally constructed towards a goal (Tolins and Fox Tree, 2014). From this standpoint word lengthening before H, seems to appear as a strategy from the speakers to hold the floor and keep the rhythm of the conversation while planning how to continue in terms of conversational goals.

3.3. Pitch analysis

We continue to analyze what we call momentary acoustic features. In particular, how they vary over time. First, we analyze the pitch tracks taken from the last second of IPUs.

Fig. 4 shows what we call *track's shadow plots* and *average plots*. A shadow plot (a variant of the *bitmap clustering* method as presented in Heldner et al. (2008)) shows the last second of *z*-scored pitch tracks from 1000 randomly selected IPUs preceding each turn-transition category. In the case of BC, we plot all IPUs (393 for English, 272 for Slovak and 842 for Spanish). The rightmost plots show the averages over the whole corpus, without any type of sampling.



(a) SPANISH

(b) SLOVAK

(c) ENGLISH

Fig. 5. Average plots for intensity, shimmer, jitter and logHNR tracks. The lines and their shadows represent the mean value and the corresponding standard deviation at each time point. All tracks are aligned to the last voiced frame.

A closer look at average plots shows unusual falls and raises in the final portion (last 10 ms) of every line. We attribute these falls to the fact that at the end of the IPU, the number of unvoiced frames increases, therefore, making the averages less reliable. To validate this observation we performed an additional experiment in which we aligned the IPUs to the last voiced frame, and observed the same general patterns without extreme falls or raises in the last frames.

A first glance at the plots shows no distinguishable patterns between 1.0 and 0.5 seconds before the IPU ending. Nevertheless, the last hundred milliseconds contain some interesting common patterns to be analyzed. Examining separately each category, we see that (a) most IPUs preceding BC finish in high-rising pitch, except in Spanish where a bimodal distribution appears and IPUs end either in a high rise or in a plateau; (b) IPUs before S end in falling pitch in Spanish, high-rising pitch in English, and a bimodal pitch shape in Slovak; (c) for H in Spanish, most of the IPUs end in a falling shape with a fair amount of rises, while in Slovak and English they end primarily in a plateau.

Intonation conveys relationships between the propositional content of previous and subsequent utterances, and segments discourse as well. For example, according to Pierrehumbert and Hirschberg (1990), in English, a rising boundary tone may indicate that the speaker wishes the hearer to interpret the utterance with particular attention to the following ones; and, by doing it in a prominent way, may elicit a response from the interlocutor.[5] In our analysis, we see indeed that the presence of high boundary tones work as a turn-ending indicator and,

[5] Patterns that we also observe in Spanish and Slovak.

thus, allow for S and BC especially in Slovak and English; while sustained signals or plateau, indicate the speaker's intention to continue talking and results in H.

In Spanish, IPUs preceding S end mostly in falls. This may be explained as a consequence of a typical strong falling intonational pattern present in Argentine Spanish (Vidal de Battini, 1964). Regarding BC, Quilis (1993) (p. 460–475) classifies declarative utterances in Spanish as either complete or incomplete. Within the latter, three final intonations are possible: two ascending and one plateau. The first ascending configuration is coincidental with those found in questions; the second, preceded by a *circumflex* movement (ascending-descending), is used for emphasis; and the third, the plateau, occurs whenever the speaker doubts or does not know how to finish. If we understand task-oriented dialogues as a collection of semantically incomplete utterances within a hierarchically structured goal, this classification may serve to explain the emerging bimodality.

3.4. Intensity and voice quality features

Fig. 5 shows average plots for all tracks; this time, aligned to the last voiced frame to avoid the previously mentioned artifacts at the end of the figures. As for Intensity, it can be seen that, for all three languages, the average for S is consistently lower than for BC and H. Also, there is a small increase followed by a decrease near the utterance end for Spanish and Slovak BC, and also for Spanish H. Intensity is a way of measuring how loud a person's speech is. Together with pitch rises, it serves to establish focus over specific portions of an utterance. In the context of turn-taking/turn-yielding it may be a way of capturing the interlocutor attention and, thus, maintaining the turn either in H and BC.

In the case of voice quality features, jitter shows differences in the last 150 ms where the H signal goes below the others, especially in Slovak; shimmer shows no clear patterns in each category; and logHNR, shows a distinction between H and the rest of the signals, especially in Spanish and Slovak. Although the described voiced quality patterns alone are not very clear, in Gravano and Hirschberg (2011) it is explained that a combination of these seems to have relevant information that is not seen individually. More recently, in Heldner et al. (2019) the authors conclude that higher level features, such as *cepstral peak prominence smoothed* (CPPS), may be better suited for capturing voice quality than jitter, shimmer and HNR.

3.5. Summary of study 1

The first approach to this study shows that the three analyzed languages share remarkably similar aspects in the acoustic/prosodic realization of turn-yielding cues. The results are consistent with previous ones in the literature and with the theory of universality that states that minimal cultural variability exists in turn-taking systems. In English, as reported in Gravano and Hirschberg (2011), our results show that IPU duration works not only as a turn-yielding cue but also as a backchannel-inviting cue, where IPUs preceding BC have higher duration than in the other type of turn transitions. Also, we show that speech rate is higher before S than before H and even more so before BC than before any of the other conditions. Our experiments not only validate these previous findings but also extend the analysis to Argentine Spanish and Slovak in which these results replicate.

In addition, as we observed for the BC category, Spanish backchannel-preceding cues present a clear bimodality between a high-rising final intonation and a plateau. Nevertheless, the strong multimodal pattern in the pitch track does not occur in English or Slovak, thus suggesting a difference in the signaling of BC produced by Spanish speakers.

Clear differences are found for the mean value of the temporal series for pitch, jitter, shimmer, and logHNR between S and H (i.e., turn-yielding cues), and also for intonation, pitch and intensity levels, IPU duration, and logHNR between BC and H (i.e., backchannel-inviting cues), as described in Gravano and Hirschberg (2011). Still, we consider these results to be incomplete and, sometimes, misleading. Analyzing averages over time (as we are showing) or the mean value of a 200, 300 or 500 ms-window at the end of the IPU (as the mentioned work does) in a univariate fashion, may lead to inaccurate conclusions. As seen in the pitch analysis, using simple averages for exploring possibly-bimodal temporal series may turn out to be insufficient.

In those cases, it may be useful to isolate the different signals that match with one type of pattern and explore if the signals belong to different turn-transition categories. This way, categories will be divided into more specific ones. For example, in the case of back-channels, two different patterns might be detected through listening.

Nevertheless, this issue exposes the problem of using univariate analysis in isolation. In the study described in the following section, we take a different approach. We train learning algorithms to model the instant decision a person makes with only a couple of seconds of partial information — the acoustic and prosodic information present in turn-transition preceding IPUs. This way, we are able to explore the relative influence of turn-taking cues when used in combination on the decision of the next turn-transition.

4. Study 2: learning turn-transitions

We know from Starkey and Fiske (1977) and from quantitative measures in Gravano and Hirschberg (2011), that turn-transition cues have an additive effect. Turn-yielding and turn-holding cues do not occur in an isolated way: the more cues signaling turn-hold or turn-yield, the higher the agreement among the listeners in identifying the subsequent turn type. Analyzing these cues separately does not give us the full picture. Additionally, it is not clear to what degree these patterns are indispensable and which can be replaced with others in relation to the information needed by the interlocutor to anticipate the next turn-transition.

This second study addresses the questions of the amount of information carried by acoustics and prosody, and the **relative contribution** of these features when preceding a turn-taking transition. Through a series of machine learning experiments, we build models capable of learning the relation between features (duration, speech rate, pitch, intensity, shimmer, jitter, and logHNR) and the subsequent turn-transition type (hold, switch, and back-channel). In particular, we train and inspect multi-class random forest classifiers for each of the three languages under study (Spanish, Slovak, and English). By finding and examining accurate models, we aim at describing with greater depth the interactions that the selected features have on the production of turn-transition cues.

The two main goals of this second study are a) to model the acoustic information available to a person when listening to IPUs before turn transitions; and b) to find out which aspects of the features and which time intervals are informative and essential for a system to predict the type of turn transition.

As an important remark, in this study, we only measure how the acoustic information relates with the subsequent turn transition. We do not claim that people use the same information these models do. The human auditory system and brain may capture information dispreferred by or even unavailable to the models, and classifiers may use information that human brains tend to discard. To understand the relative importance of the incidence of these cues, perceptual studies need to be conducted such as in Hjalmarsson (2011) where utterances are synthesized to understand how different prosodic cues affect the decision of the interlocutor. In the following sections, we describe how we build, train and inspect our machine learning models.

4.1. Model choice

As pointed out by Skantze (2018), a common approach in the turn-taking literature is to calculate a brief window of interest just

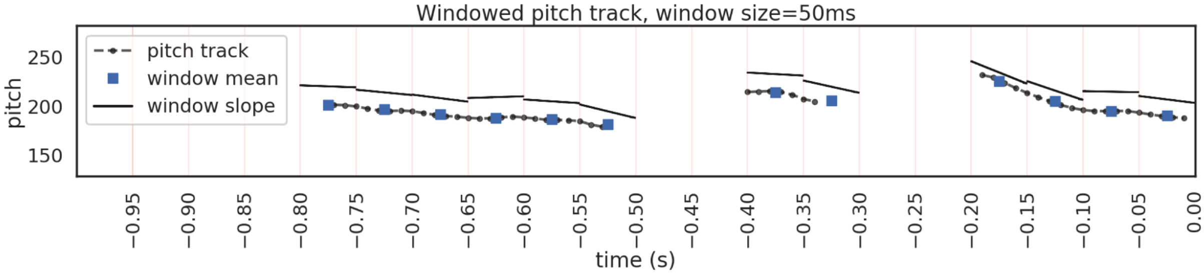


Fig. 6. Illustration of how mean and slope values are computed over z-scored pitch tracks for a given instance of our datasets. The dotted line represents the pitch track of the instance; the vertical lines delimit the different windows; the blue squares represent the mean value of each window; and the black lines show a shifted version of the linear fit to the points from which the slope is obtained. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

before each pause. In this paradigm, the dynamic of features and the contextual information in previous events in conversation need to be represented as numbers through the introduction of heuristics and careful feature engineering. Skantze points out that this way of treating features constitutes a clear drawback since context (i.e., previous speech activity) or signal dynamics (i.e., rapid changes on the pitch track) is lost if not taken explicitly into account. In that direction, he proposes to use LSTM networks for processing input vectors on a frame-by-frame basis. These state-of-the-art models take context into account and capture the dynamics of multivariate signals.

Unfortunately, today's techniques for interpreting deep learning models remain immature. LSTMs and other state-of-the-art techniques are considered to be obscure in the way they process information. Understanding how such models predict in terms of the original features, what information they pay attention to, and how features interact with each other is still a difficult task, especially over multivariate time series data.

Consequently, we choose to experiment with random forest models which are considered accurate and simple enough in terms of complexity (understood as the amount of required training data, tuning effort, and computing power). This method has been widely-used as a means of understanding phenomena even in high-dimensional problems with complex interactions. For example, Lunetta et al. (2004) shows that random forests can be used to detect relevant genetic marker interactions more efficiently than univariate screening methods like Fisher Exact test. For these reasons, we consider these models to be well suited to our purpose of understanding in further detail how speakers produce turn-taking cues.

4.2. Machine learning task definition

The INSTANCES of our machine learning tasks are inter-pausal units (IPU, see Section 2.1). In particular, we keep IPUs preceding hold (H), switch (S) and backchannel (BC) transitions. From each instance, we compute a FEATURE VECTOR — a fixed-size vector used as input for the machine learning models. These vectors contain the normalized versions of the whole-ipu features (IPU duration and speech rate) and also a normalized representation of the momentary acoustic features (intensity, pitch, jitter, shimmer, and logHNR).

To compute the representation of the momentary acoustic features, we align each time series to the end of the IPU and then slide a fixed-size window over the last second of it. If an IPU is shorter than a second, we left-pad the time series with NaN (not a number) values. Windows are 50 ms wide, with a 50 ms step with no overlap.[6] From each window, we compute each feature's mean value given all available samples within the limits of the window and also their slope as an attempt to capture the dynamics of the features over time. Note that the time series of momentary acoustic features may contain missing values due to the presence of unvoiced frames in which pitch track and voice quality features are undefined. When fewer than two values are defined in the window, we set mean and slope values to be NaN. Fig. 6 serves as an illustration of the mean and slope values computed over the z-scored pitch track for a given instance of the datasets.

By running the sliding window process, we obtain a 40-dimensional vector $x_f^{(i)}$ that contains the mean and slope values of a given momentary acoustic feature f computed over the 20 intervals in the last second of the IPU. Finally, all vectors are concatenated along with the two whole-IPU features obtaining $\boldsymbol{x}^{(i)} = \boldsymbol{x}_{pitch}^{(i)} \oplus \boldsymbol{x}_{intensity}^{(i)} \oplus \boldsymbol{x}_{logHNR}^{(i)} \oplus \boldsymbol{x}_{jitter}^{(i)} \oplus \boldsymbol{x}_{shimmer}^{(i)} \oplus \boldsymbol{x}_{wholeIPU}^{(i)}$, a 202-dimensional feature vector.

After defining the instances and the feature extraction process, we proceed to build models $\hat{f}(\mathbf{x})$ that, given a feature vector $\boldsymbol{x}^{(i)}$, predict the turn-transition type $y^{(i)}$ that follows. The possible TARGET CLASSES are H, S and BC.

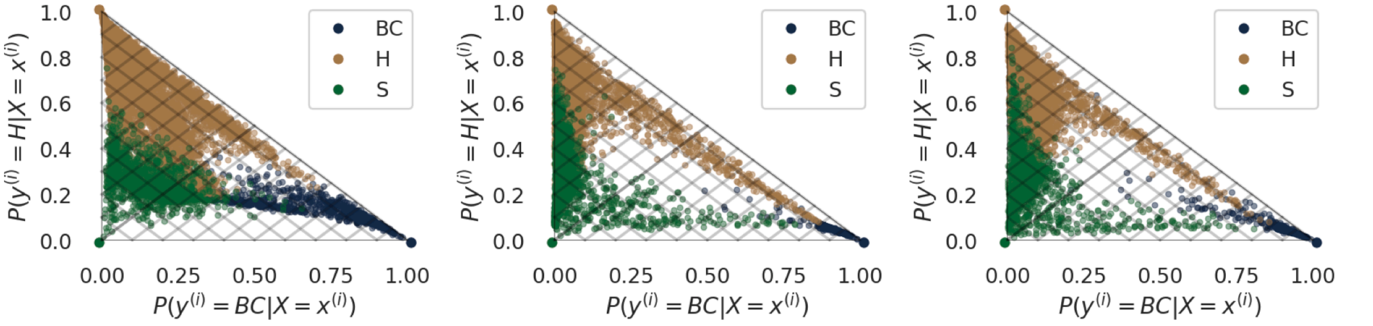
The selected LEARNING ALGORITHM was the scikit-learn implementation of random forest (Pedregosa et al., 2011). We then ran the MODEL SELECTION step, in which we measured different hyperparameter combinations. For each language, we ran the scikit-learn randomized search procedure for 100 different random combinations. Given a combination of hyperparameters, we measure its performance by running a 10-fold leave-one-group-out cross validation using macro-averaged F1 score as the performance metric. A **group** consists of all the IPUs of at least one speaker, and all speakers belong to only one group. Appendix A shows how the hyperparameter search was made, the cross validation mechanism, the selected model's performance, and some other implementation details.

Since the scikit-learn random forest implementation does not support undefined values, we substitute each missing value with a fixed number. For this, we choose a value lower than the feature's overall minimum (after the normalization process). Given the way decision trees and random forests work, the learning algorithm should be able to automatically create specific sub-trees for feature-value combinations that are below or above a threshold. Therefore, replacing NaN with a value below the minimum should make it possible for trees to specifically deal with such cases.[7]

This decision of *imputing* (i.e., filling in missing values) a constant number below the minimum, rather than the more standard approach of

[6] These values were selected based on a trade-off between temporal granularity (we want the time series to be informative) and feature robustness (we want to have enough measurements in each interval to be able to compute robust slopes).

[7] In practice, it is important to study how the algorithms search for the optimal cuts, since there are various heuristics. An heuristic that does not consider the specific cut that leaves the minimum values in a region, will not be able to distinguish these from other small values of the feature.



(a) Spanish (b) Slovak (c) English

Fig. 7. Posterior probabilities given by the selected models to the training data. Each plot summarizes the scores assigned by a model to all the training instances in a corpus. The three vertices of a triangle represent perfect predictions – the closer a point is to a vertex, the better its score given by the model.

using the variable mean, is related to the fact that the existence itself of a missing value may comprise useful information. As an example, the lack of pitch at the end of an IPU may be part of a cue for the interlocutor to take the floor. Strategies such as discarding instances with missing values, interpolating values based on other features, or replacing missing values with a value with high probability (according to the feature's distribution) may contribute to unrealistic and sub-optimal models not able to distinguish between real and synthetic values, thus introducing new, unwanted biases. Also, we discarded the idea of using dummy columns as indicators for imputed values. This approach not only doubles the number of features, but also introduces a conceptual issue related to the selection of features that the random forest algorithm performs at each node. Selecting features at random may result in the separation of the original variable from its dummy column, thus impeding trees from determining whether it is an imputed value or not.

The last step of the process consists in re-fitting the selected models using all the available data for each language. In other words, after selecting the combination of hyperparameters, we train the models on all the data. This way, we exploit all the corpora at the expense of losing the ability to correctly estimate how models would perform in a real-world setting. Again, our goal is not to build competitive models, but interpretable ones to explore the dataset.

4.3. *Measuring feature contributions*

Once the models have been trained, we compute a variant of the *permutation importance* method for measuring the contribution of each feature to the model. Proposed by Breiman (2001a), the permutation importance method proposes to estimate the contribution of each feature by measuring the decrease in the performance of the model when that specific feature is not available. Since we are dealing with a multiclass problem in which classes are unbalanced, we implement a variation of the method that measures how predictions shift for instances of each class independently. The complete procedure is as follows.

First, we compute the scores that each model assigns to each instance. In the case of a random forest, these scores are an estimated probability computed by averaging the prediction of each individual tree. For example, for instance i, the model may assign $P(Y = H|X = x^{(i)}) = 0.3$, $P(Y = BC|X = x^{(i)}) = 0.05$, and $P(Y = S|X = x^{(i)}) = 0.55$, meaning, in the first place, that the instance is likely to precede a switch transition and, in the second, to precede a hold. Fig. 7 shows how the three selected models assign scores for every instance of our corpora, colored by class. This figure summarizes the way models split the data. Points close to a vertex of the same color represent perfectly classified instances. Misclasification occurs when the distance of a point to its true vertex is greater than the distance to a different vertex. Despite not being perfect (F1-score between 0.7 and 0.73 on this data), these triangles visually show that instances were assigned with higher probabilities to the right class in most of the cases. See Appendix A for further details on the performance of the models.

Second, we measure how the models behave when a feature is "removed". Since when making a prediction a trained random forest model expects all features to be present, Breiman (2001a) proposes to mimic the deletion of a variable by replacing feature values with random noise drawn from the original feature's distribution, breaking in this way the relationship that may exist between an instance feature value and the expected target value. Therefore, the last step consists in shuffling the column that contains a specific feature and computing how the assigned scores change. If the feature significantly contributes to determining the probability of certain classes, we will expect a substantial decrease in the correct class score; otherwise, we will expect smaller shifts in scores.

For measuring how much the scores shift, we computed what we call Mean Class-Probability Decrease (MCPD) defined as follows. Given a feature j and a label y, the feature's MCPD is:

$$\mathrm{MCPD}_y^j(X) = \frac{1}{N_y} \sum_{x^{(i)} \in X^{(Y=y)}} \mathrm{ProbDecrease}_y^j\left(x^{(i)}\right)$$

where $\mathrm{ProbDecrease}_y^j(x^{(i)}) = P(Y = y|X = x^{(i)}) - P(Y = y\big|X = x_{\pi j}^{(i)})$, and $x_{\pi j}^{(i)}$ corresponds to the ith instance of class y with the value for feature j replaced with a random value (drawn from all values in the jth column of X). In other words, this metric determines the mean decrease in posterior probabilities assigned to all instances of a given class when a feature value is replaced by other from the same distribution.

Note that ProbDecrease is only used for instances of the class being evaluated. Therefore, it is expected that the result of this function is a positive number, and the more this attribute is used by the model, the greater its value. In any case, ProbDecrease might produce a very small or even negative number due to the method's variance, the noise in the data, or the chance involved in the permutation of X_j. These small fluctuations should cancel each other out and, when experimenting, we did not observe important effects as a result of this property. In general, the true effects of important features were of a different order of magnitude.[8]

We utilize MCPD as our measure of feature contribution and interpret higher values as an indicator of higher feature importance. Finally, we compute the MCPD for all whole-IPU and momentary acoustic features for the three selected models.

[8] Also note that this issue is also present in the technique on which we based this method, Breiman's permutation importance.

Table 2
MCPD ranking for whole-IPU features (IPU duration and speech rate) among all features. Results are shown for each model (M_{50}, M_{100} and M_{200}) in every language.

		M_{50} (202 features)			M_{100} (102 features)			M_{200} (52 features)		
		Spanish	Slovak	English	Spanish	Slovak	English	Spanish	Slovak	English
IPU Duration	BC	4	4	3	4	4	3	3	4	3
	H	3	5	7	4	7	5	4	5	8
	S	4	4	5	5	5	5	4	4	6
Speech Rate	BC	1	5	7	2	6	6	2	5	6
	H	1	1	2	1	1	1	1	3	3
	S	1	1	1	1	1	1	1	1	2

4.4. Results: Contribution of whole-IPU features

The first three columns of Table 2 (M_{50}) show how the whole-IPU features are ranked among all 202 features. In all three languages, speech rate and IPU duration appear among the 7 most contributing features. In particular, speech rate seems to be the feature most affecting H and S predictions. This pattern coincides with the analysis in study 1: at the acoustic level, features of IPUs preceding H and S are similar in most respects, except in speech rate, which is consistently lower before H in every language. Therefore, slow speech rate before a pause highly increases the likelihood of a turn continuation.

It is important to point out that comparing scores assigned to the two whole-IPU features against scores assigned to the other 200 momentary acoustic features can be unfair — when a whole-IPU feature is removed, all of its information is lost; on the contrary, if a momentary acoustic feature is removed, its temporal neighbors plus the slope of the previous timepoint will still be contributing similar information. For a more fair comparison, we train two alternative models for each language, both identical to the previously detailed ones, but using a window width of 100 or 200 ms instead of 50 ms for momentary acoustic features, thus exploring widths (and therefore feature sets) of different orders of magnitude. These new models handle fewer features at the expense of a lower temporal precision. For the sake of organization, we call them M_{50}, M_{100} and M_{200} to indicate the width of the feature window.

We now test how whole-IPU features rank based on these new models. Since the correlation effect is attenuated, we expect whole-IPU features to rank lower. Columns 4 to 10 in Table 2 show how whole-IPU features rank for models M_{100} and M_{200}. Speech rate still maintains the highest positions in the ranking, again among the top 3 most important features for discriminating H and S. This indicates that word final lengthening and other factors that cause a decrease in speech rate before pauses contain crucial information for detecting if a H or S transition is about to occur. IPU duration also maintains its position among the top 8 features for all three languages and turn taking types.

4.5. Results: contribution of momentary acoustic features

In the next paragraphs we direct our attention to the contributions of momentary acoustic features over time. We compare their relative importance against each other. Fig. 8 shows, for model M_{50}, the contribution of each turn transition type in each language. Each small square represents the MCPD value for a given feature in a given time interval using a chromatic scale. The darker the color, the more salient the contribution of a feature to the prediction of the subsequent transition. Additionally, we use Fig. 9 to compare the contribution of each feature in the M_{100} model, as defined in the previous section.

We first observe the contribution of pitch in the M_{50} model. In all three languages, it seems essential, especially, on the last 200 ms. In Slovak, the major importance appears to be in the last 50 ms, while in Spanish and English it goes further back around 100 ms before silence. Recalling the Slovak plots of Fig. 4, we notice that BC shows similar patterns in pitch to S and H, except for a high rise in the last hundreds of milliseconds. This difference makes the BC pattern very specific and, therefore, discriminative of the category. In English and Spanish, softer and less discriminative distinctions emerge. The contribution of features in the M_{100} case shows similar patterns.

Second, we focus on pitch slope. In English and Slovak, its contribution is comparable to that of pitch itself while in Spanish the pitch feature keeps the majority of the information. This is somehow expected since, as showed in Study 1, Spanish pitch presents a bimodality and, therefore, a near-zero slope does not contribute enough to determine whether the IPU precedes BC, H or S transitions. In Slovak and English, on the contrary, the slope seems to contribute complementary information.

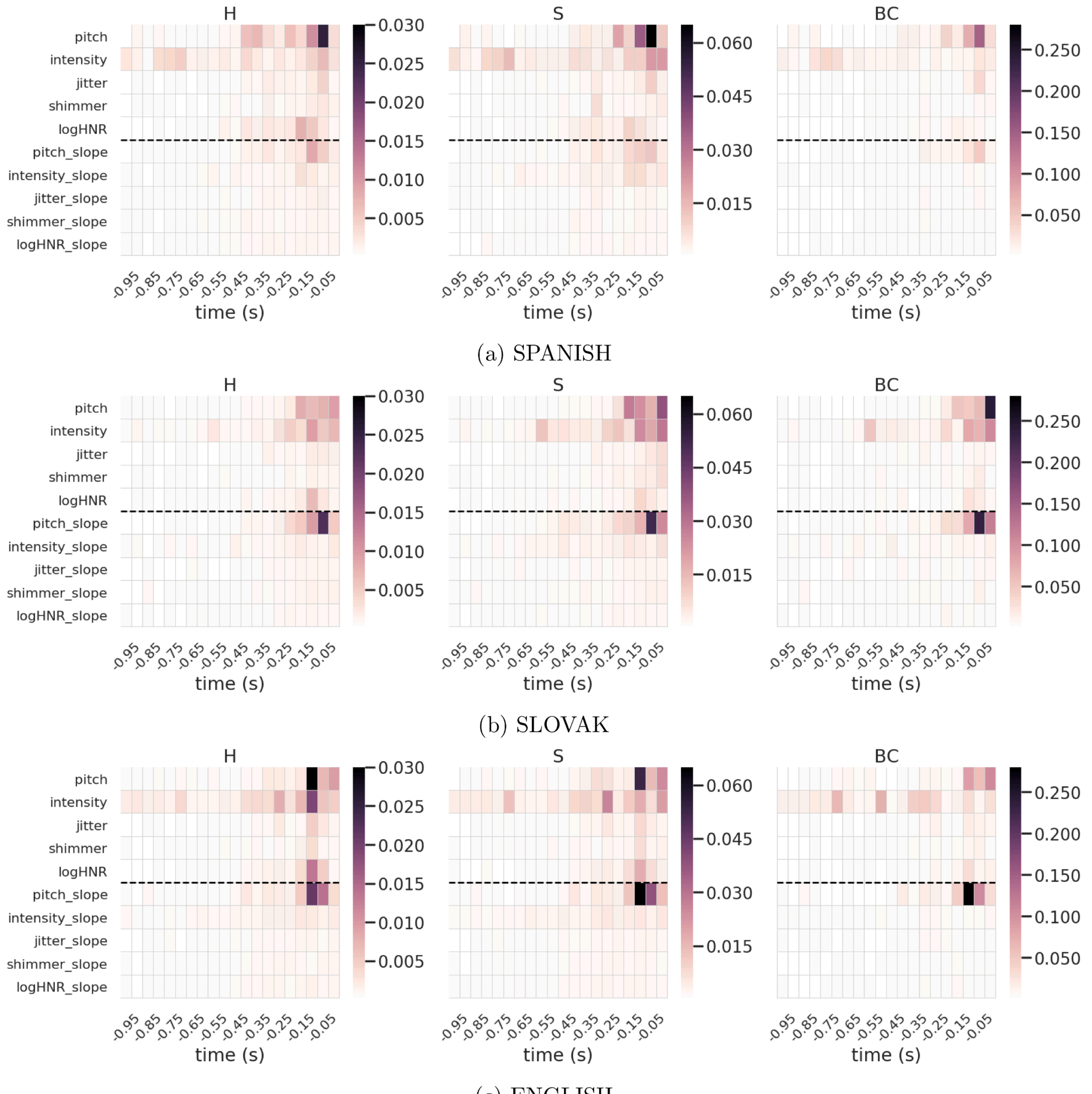
Third, the intensity track. In English and Slovak a peak can be seen around 150 ms before silence. In the case of Spanish this pattern is not as clear in the M_{50} model, as it is in the M_{100} model. This information matches the observed changes in intensity analyzed in the previous study, where clear patterns before S (low plateau shape) and before BC and H (higher values with some rises and falls near the end) could be seen in the average plots (Fig. 5) in the last 200 ms. Moreover, this feature shows higher contributions across time than the rest of the features, especially in the case of English and Spanish. Again, looking at the average plots, we observe that S intensity seems to keep a near constant distinguishable value across time in English and Spanish. The intensity slope does not seem to be as important as the intensity level showing that the dynamics of this feature do not affect the results as much as the overall level.

Fourth, we look at logHNR. This feature shows high contribution for H in the last hundred milliseconds before pauses in the three languages. In the case of Spanish, the contribution is spread in the interval between −400 and −100 ms before silence. In the case of Slovak and English, it concentrates between 100 and 200 ms before silence. Model M_{100} confirms these patterns. The average plots in Study 1 showed strong patterns of separation across the entire 500 ms under analysis in Spanish and Slovak, and in English to some extent. In contrast, the newly measured contribution seems to concentrate on a specific portion of the signal showing that complex interactions may be taking place when capturing the information provided by this feature.

Finally, shimmer and jitter show little contributions to the model predictions. An exception can be seen near the last 150 ms in which jitter in English and Spanish contributes to some extent. Regarding jitter and shimmer slope, some contribution is seen towards the end especially in English for the M_{100} model. To our knowledge, there is no linguistic reason for jitter and shimmer slopes to contain meaningful information and no clear patterns emerged in the previous study. We leave the exploration of these features open for future studies.

4.5.1. Feature correlations

In Genuer et al. (2010), authors present a series of simulations that portray how the importance of a variable is not stable. They do so by generating small perturbations in the hyperparameters of the random forest algorithm and varying the number of noisy and correlated variables. They then measure the impact this produces in the variable importance measures. Since the algorithm tends to select one variable over another by chance, the importance of the entire group of correlated variables decreases, sometimes making them indistinguishable from noisy variables. In Toloşi and Lengauer (2011) the authors define the



(a) SPANISH

(b) SLOVAK

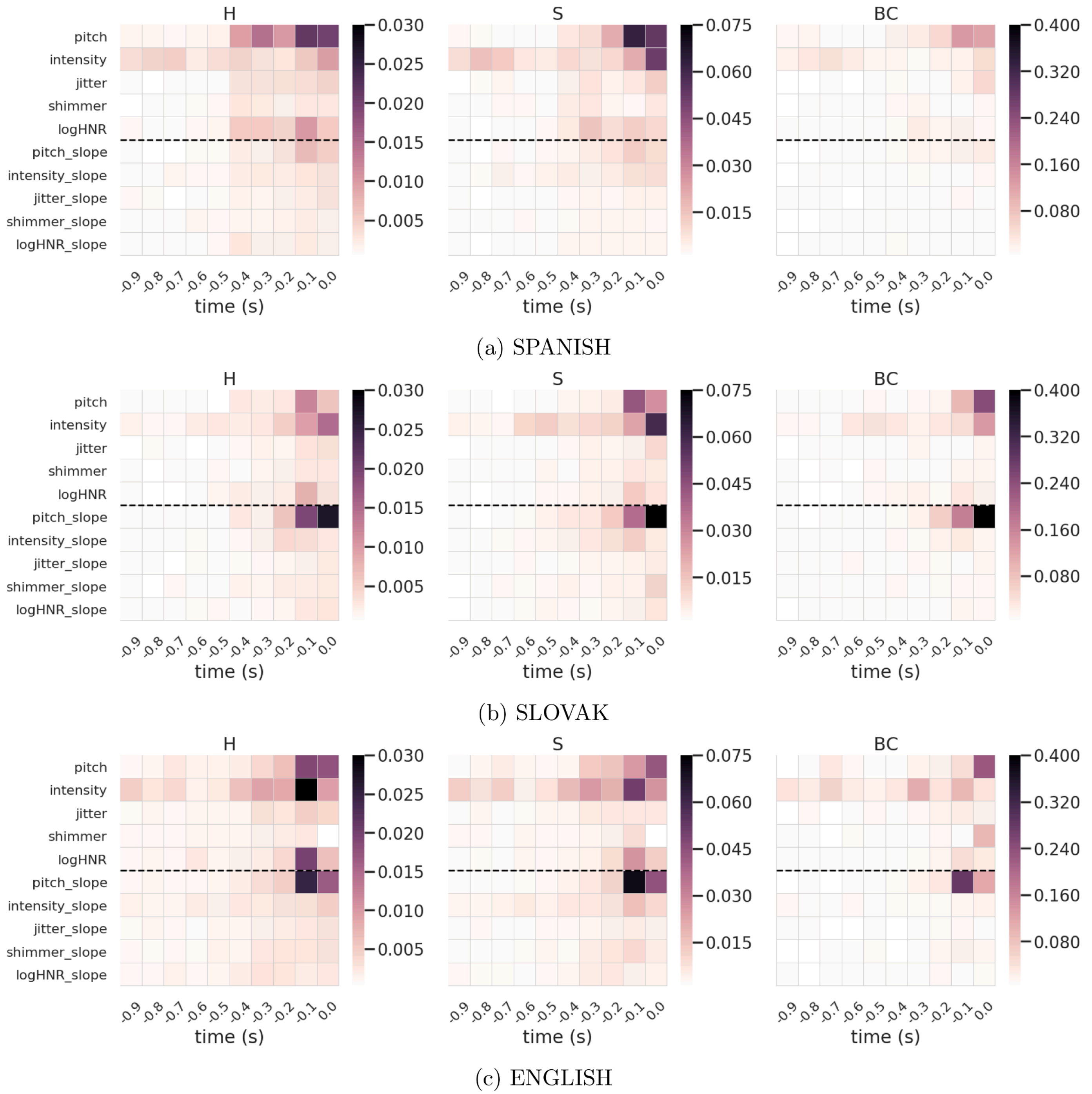
(c) ENGLISH

Fig. 8. Importance of momentary acoustic features (using a 50 ms window). The color plots show the contribution over time of each feature in the M_{50} model, for each language. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

term CORRELATION BIAS as the breach of specific desirable rules: (1) all variables in a correlated group should have the same importance; (2) the size of the group should not affect the importance of the individual variables and (3) the importance of the variables reflects the magnitude of the effect of the corresponding process on the outcome. They run simulations in which they show that random forest models suffer from correlation bias. If the way variable contribution is measured does not take the correlation bias problem into account, **unstable results** may be produced, affecting the generalizability and reproducibility of the results under description.

Our version of the permutation importance method, as well as most feature selection and feature importance techniques, suffers from the variable correlation problem. If several variables are correlated and the estimator uses them all equally, the importance of the permutation can be small for all these features. Also, the elimination of one of the features may not affect the result, since the estimator still has access to the same information from other features.

In this work, we addressed this problem by choosing our features so that they would have small pairwise correlations. In preliminary experiments, we computed our features using overlapping windows and observed that each time we trained a model results differed significantly. After taking this problem into account, we concluded that non-



(a) SPANISH

(b) SLOVAK

(c) ENGLISH

Fig. 9. Importance of momentary acoustic features (using a 100 ms window). The color plots show the contribution over time of each feature in the M_{100} model, for each language. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

overlapping windows were the best solution.

Fig. 10 shows the correlation matrix of all the features extracted from each of the corpora. In this 200 by 200 matrix, each value shows the correlation value between two momentary acoustic features in the different time intervals. For illustration purposes, we only used the momentary acoustic features of the Spanish M_{50} model, computed from a 50 ms sliding window with no overlap. This figure serves as an indicator of how much we can trust the permutation importance measure presented in the previous section. If features were highly correlated, it is more likely that the contribution will be lower and unstable across different model training steps.

It can be seen for example that pitch is one of the most autocorrelated features over time. Also, it has a considerable correlation with intensity. On the other hand, pitch against its slope and intensity against its slope also have high correlations. Finally, logHNR, shimmer and jitter also seem to share information. These facts have to be taken into account when making claims about feature importance and about how features interact with each other.

4.6. Summary of study 2

In this second study we observed that, in all three languages, various features contribute to the production of different types of turn transitions in a similar way; and that these results are consistent across

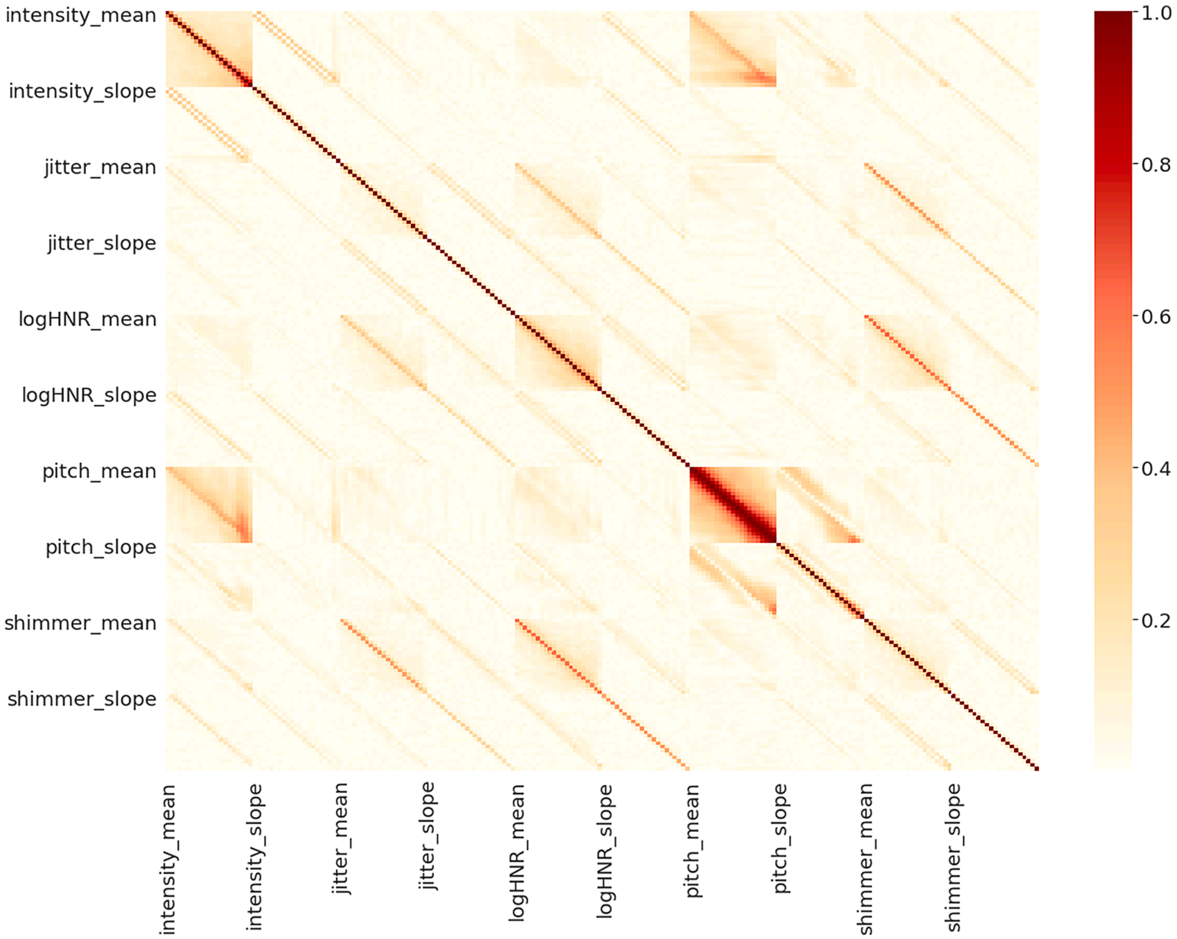


Fig. 10. Correlation matrix for momentary acoustic features for the Spanish corpus (50 ms window).

different time windows. However, because features show correlation to some extent, the results need to be interpreted carefully. We used MCPD (Mean Class-Probability Decrease), a novel measure of feature importance to estimate the contribution of what we call whole-IPU and momentary acoustic features.

As a first result, we saw that whole-IPU features, speech rate and IPU duration, appeared among the seven most contributing features in all languages, speech rate being the one that most affects H and S predictions. However, while the information present in a whole-IPU feature disappears as soon as it is removed, in a momentary feature such information could be retrieved from its context. To attenuate these correlation effects we built and tested three different models for each language, with varying time windows. Again, results showed whole-IPU features, especially speech rate, to be among the most important features to discriminate H and S.

Regarding momentary features, pitch was the most important, especially on the last 200 ms of speech in all languages, only followed by pitch slope, which turned to be more important in English and Slovak than in Spanish. Intensity, showed the bigger contribution over time, concentrating its effect on the (−200 ms, -100 ms) interval for all languages. LogHNR showed to be important for discriminating H and S in the last hundred milliseconds before pauses. Finally, shimmer and jitter showed scarce contributions to the model predictions, with an exception in English and Spanish over the last 400 ms.

This analysis contains lots of condensed information that make comparisons somewhat difficult to summarize. Therefore, we release a folder containing all feature importance calculations for the reader to explore: https://github.com/pbrusco/turn-taking-SPECOM.

5. Conclusions

We conducted a number of experiments to explore similarities and differences between American English, Slovak and Argentine Spanish in the production of acoustic/prosodic cues before turn exchanges. We analyzed the speech in three corpora of spontaneous dyadic conversations, first through a series of visual explorations, and second by using machine learning techniques to predict turn transitions based on features from pause-preceding units.

In the first study, we saw in detail how IPU features (pitch, intensity, duration, speech rate, and voice quality features) vary over time and comparably across the three languages. Also, we considered the problems associated with the use of averages instead of raw signals. In the second study, we defined a new metric to measure feature importance per class, by simulating the deletion of a feature in a chosen model and analyzing its effects on its predictions.

After the experiments, we found that, generally speaking, the three languages under study share acoustic/prosodic resources to signal turn transitions. We were also able to rank the features in each language by their contribution to the separation of turn transition classes. We consider this information as useful for both the linguistics community and the spoken dialogue systems community alike. If one builds a prediction system, a good approach is to start with the most informative features. In particular, we recommend:

1. To use speech rate as a feature. This feature showed to be fundamental to distinguish turn-holdings from turn-yieldings. Word lengthening, too, shows to be important in the domain of

collaborative dialogue and may be important as well in more general dialogue systems.

2. To include the entire pitch track (especially the last 200 milliseconds) in the feature set. Pitch is well known to be essential as a turn-yielding and backchannel-inviting cue. In particular, pitch slope over time has shown to provide useful complementary information in cases where models do not automatically take into account the dynamics of the signal.
3. To include the entire intensity track in the feature set. Intensity level itself, rather than its dynamics, has shown to contain useful information for hold, switch and backchannel predictions. In particular in English and Spanish, where the model showed a contribution through time for longer periods than in Slovak.
4. To use logHNR as a feature. This feature showed particular importance between 200 and 100 ms before pauses, especially for switch and hold predictions. We advise to include at least the last 200 ms of information to capture these cues.

There are several possible directions for future research. The first one is to understand to what extent a spoken dialogue system may use contextual cues such as turn-initial prosodic cues. For example, Sicoli et al. (2015) show that speakers use a boosted initial pitch to signal questions. Yet, our models only explore turn-final IPUs, and turn-initial prosodic information might be missing. For a better understanding of these phenomena further experiments need to be conducted.

Second, there seems to be some margin for improving the models and exploring what information could be missing to reach a prediction performance as perfect as possible. We also plan to redesign the MCPD metric for taking into account the direction in which predictions shift and the original prediction value.

A third direction may consist in testing the classifiers on data from a perception study. In the present study we only make claims about the information rendered important by the proposed algorithms to classify different types of turn transitions. We make no assumptions on how the human brain processes such information. However, it may be interesting to see the similarities or differences between the mistakes made by system and humans.

Fourth, we plan to extend this work with the analysis of other types of turn transitions such as interruptions and overlaps; as well as explore further the already known, for example, make an in-depth analysis of the final pitch contours in IPUs preceding BC that seem to produce a bimodal distribution in Spanish.

CRediT authorship contribution statement

Pablo Brusco: Conceptualization, Investigation, Software, Methodology, Visualization, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Jazmín Vidal:** Methodology, Validation, Data curation, Writing - original draft, Writing - review & editing. **Štefan Beňuš:** Resources, Data curation, Investigation, Writing - original draft. **Agustín Gravano:** Resources, Conceptualization, Investigation, Methodology, Data curation, Writing - original draft, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Argentina's National Agency for Scientific and Technological Promotion (ANPCYT), PICT-PRH 2009-0026 and PICT 2014-1561; the University of Buenos Aires, UBACYT 20020090300087 and 20020120200025BA; the Bilateral Cooperation Program between Argentina's National Scientific and Technical Research Council (CONICET) and the Slovak Academy of Sciences (SAS); the VEGA 2/0161/18 grant from the Slovak Scientific Granting Agency, and the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055.

Appendix A. Hyperparameter optimization

A model hyperparameter is defined as an external configuration whose value is not estimated from the data. Generally speaking, hyperparameters define the model's flexibility, and therefore play a central role in the bias-variance trade-off. Hyperparameters were selected by running a randomized search procedure with the following restrictions:

- `n_estimators`: The number of trees in the forest (an integer number between 50 and 500).
- `max_depth`: The maximum allowed depth for each tree (an integer number between 2 and 15).
- `max_features`: The number of features that are randomly selected at each node split (a percentage between 1% and 100% of the total number of features). The higher, the more similar the resulting forest trees will be to each other.
- `balance method`: We tried three different balancing strategies — **oversampling** (supplementing the training data with random samples of some of the minority classes); **undersampling** (restricting the amount of data from each class at each tree level); and **unbalanced** (in which no balancing is applied).

Given a combination of hyperparameters, we measure its performance by conducting a 10-fold leave-one-group-out cross validation using macro-averaged F1-score as the performance metric. In this procedure, subsets of data (*folds*) are assigned either to train a model (9 folds) or to measure its *validation performance* (the remaining, unseen fold). Validation folds are rotated to get 10 values of performance for each hyperparameter combination. Each fold contains a particular group of speakers that are not present in more than one fold. Therefore, by measuring the validation set score, we take into account the desired property of how the model generalizes to the new group of speakers without overestimating our results due to speaker-specific patterns.

At the point in which we train our models, we are interested in balancing the data so that all trees learn patterns from the same amount of H, S, and BC instances; i.e., we remove prior information about turn-transition probabilities. Removing prior information may not be the best option when building a real state-of-the-art system for predicting in a "real setting" since turn-taking transitions are naturally unbalanced. However, our goal is to build a model that can be explored for understanding the data, rather than building a state-of-the-art model for making predictions on new corpora. In the same direction, we did not set the random forest `class_weight` attribute since the bias it produces on the forest does not help for being fair in the contribution each feature has.

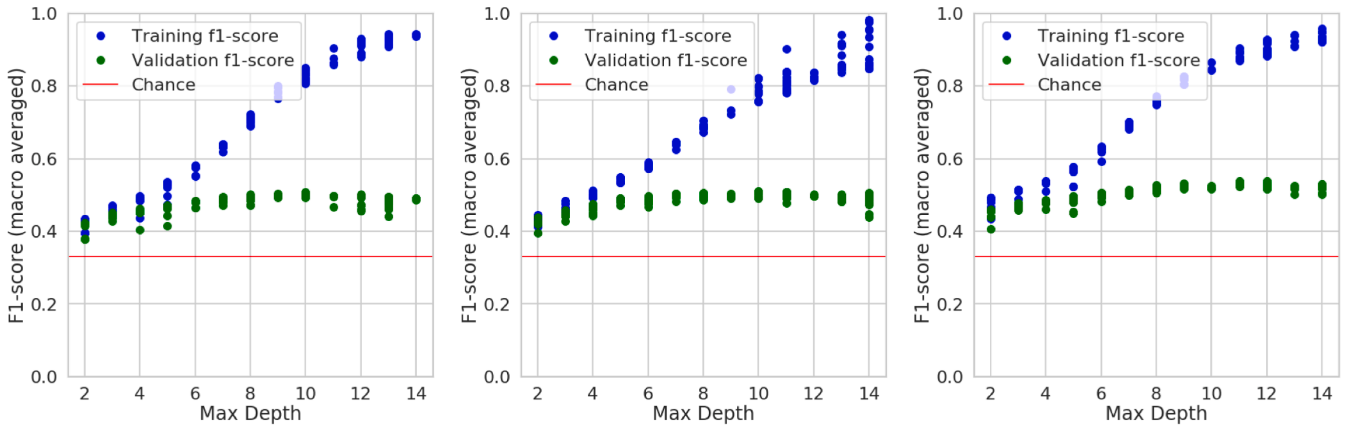
The **metric** we selected for measuring each model performance is the *macro-averaged* F-score. Given the posterior probabilities emitted by a model,

Table A.3
Hyperparameter search: Balance method results. The two center columns show the mean and standard deviation of the performance (F1-macro averaged score) of all hyperparameter combinations. The two rightmost columns show the performance of the selected models.

		All combinations		Selected Model	
	Balancing Method	Training	Validation	Training	Validation
Spanish	Oversampling	0.70(± 0.18)	**0.47**(± 0.02)	0.84	0.50
	None	0.54(± 0.24)	0.33(± 0.04)		
	Undersampling	0.61(± 0.24)	0.34(± 0.04)		
Slovak	Oversampling	0.71(± 0.15)	**0.48**(± 0.02)	0.82	0.51
	None	0.61(± 0.25)	0.35(± 0.04)		
	Undersampling	0.60(± 0.26)	0.35(± 0.04)		
English	Oversampling	0.73(± 0.16)	**0.50**(± 0.02)	0.84	0.52
	None	0.68(± 0.23)	0.39(± 0.05)		
	Undersampling	0.63(± 0.23)	0.37(± 0.06)		

instances are assigned to a class based on the most likely value: $prediction = arg\_max_k\{P(Y = k|X = x^{(i)})\}$. Once all instances have been assigned, recall and precision are computed for each individual class (H, S and BC). Next, the F score (the harmonic mean of precision and recall) is computed as $F_k = 2\cdot\frac{\text{precision}_k\cdot\text{recall}_k}{\text{precision}_k+\text{recall}_k}$. Third, "macro average" is the result of computing the metric for each class, and then compute their unweighted mean: $F = (F_H + F_{BC} + F_S)/3$. Again, we do not consider label imbalance at this point, since we are not interested in making predictions in a real domain, in which imbalance is very likely to occur. This oversampling method is performed only over training folds after splitting the original data to avoid information leaking across folds. Training scores are also computed since they can help understanding how our model fits the data.

After running the hyperparameter optimization process, we observe that the balancing method turned out to be the most determinant hyperparameter in which the oversampling method outperformed almost all other combinations. The first two columns of Table A.3 show a summary of these results. Having selected the balance method, the `max_depth` hyperparameter was the second most determining setting. Fig. A.11 shows how the performance is affected when varying the `max_depth` hyperparameter. These plots show the known bias-variance trade-off, in which the deeper the trees are allowed to grow, the more likely it is for such trees to overfit to the training data, and thus to increase their training score, in detriment of their generalization power (i.e. the validation performance). Also, this figure shows a remarkable similarity between how different combinations perform in the three languages. This desired property allows us to choose similar settings for the three classifiers and as a consequence, allows fairer comparisons across languages. By analyzing the group of higher ranked combinations for each language (i.e combinations without significant difference versus the best one), we selected the following hyperparameter combination for all languages: number of trees in the ensemble: 300, max depth: 10, features to select at each split: 50% and oversampling of data as the balancing method. The last two columns of Table A.3 show how the selected models perform on each corpus.



(a) Spanish (b) Slovak (c) English

Fig. A.11. Results of a randomized-search procedure conducted on oversampled data. The green dots indicate the F1-score performance of a combination of hyperparameters on validation sets; the blue dots, on training sets. The red lines indicate the performance of a majority-class (i.e., dummy) classifier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

Biau, G., Scornet, E., 2016. Rejoinder on: a random forest guided tour. Test 25 (2), 264–268. https://doi.org/10.1007/s11749-016-0488-0.

Bögels, S., Torreira, F., 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. J. Phon. 52, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., 2001. Statistical modeling: the two cultures. Stat. Sci. 16 (3), 199–215. https://doi.org/10.1214/ss/1009213726.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Duncan, S., 1974. On the structure of speaker-auditor interaction during speaking turns. Lang. Soc. 3 (2), 161–180. https://doi.org/10.1017/S0047404500004322.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andre, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P., 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417.

Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. Proceedings of the 2013 ACM Multimedia Conference. ACM, pp. 835–838. https://doi.org/10.1145/2502081.2502224.

Ferrer, L., Shriberg, E., Stolcke, A., 2002. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody. 7th International Conference on Spoken Language Processing, ICSLP 2002, pp. 2061–2064.

Ford, C.E., Thompson, S.A., 2010. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. Interact. Grammar 13, 134–184. https://doi.org/10.1017/cbo9780511620874.003.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.2307/2699986.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31 (14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Graham, R., 1978. Intonation and emphasis in spanish and english. Hispania 61 (1), 95–101.

Gravano, A., Benus, S., Hirschberg, J., Mitchell, S., Vovsha, I., 2007. Classification of discourse functions of affirmative words in spoken dialogue. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2007, pp. 1621–1624.

Gravano, A., Brusco, P., Benus, S., 2016. Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2016, pp. 1265–1269. https://doi.org/10.21437/Interspeech.2016-585.

Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Comput. Speech Lang. 25 (3), 601–634.

Hara, K., Inoue, K., Takanashi, K., Kawahara, T., 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2018, pp. 991–995. https://doi.org/10.21437/Interspeech.2018-1442.

Heldner, M., Edlund, J., Laskowski, K., Pelcé, A., 2008. Prosodic features in the vicinity of silences and overlaps. Proc. 10th Nordic Conference on Prosody, pp. 95–105.

Heldner, M., Włodarczak, M., Benus, S., Gravano, A., 2019. Voice quality as a turn-taking cue. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2019. The International Speech Communication Association (ISCA), pp. 4165–4169. https://doi.org/10.21437/Interspeech.2019-1592.

Hjalmarsson, A., 2011. The additive effect of turn-taking cues in human and synthetic voice. Speech Commun. 53 (1), 23–35. https://doi.org/10.1016/j.specom.2010.08.003.

Hualde, J.I., 2013. Los sonidos del español: Spanish Language Edition. Cambridge University Press.

Jasinskaja, K., 2016. Information structure in slavic. The Oxford Handbook of Information Structure, pp. 709–732.

Král, A., 1988. Pravidlá slovenskej vyslovnosti. Slovenské Pedag. Nakl.

Levinson, S.C., 2016. Turn-taking in human communication - origins and implications for language processing. Trends Cogn. Sci. 20 (1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010.

Lin, M., Fan, B., Lui, J.C., Chiu, D.M., 2007. Stochastic Analysis of File-Swarming Systems, vol. 64. Springer. https://doi.org/10.1016/j.peva.2007.06.006.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, vol. 2017-December, pp. 4766–4775.

Lunetta, K.L., Hayward, L.B., Segal, J., van Eerdewegh, P., 2004. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. 5 (1), 32. https://doi.org/10.1186/1471-2156-5-32.

Maier, A., Hough, J., Schlangen, D., 2017. Towards deep end-of-turn prediction for situated spoken dialogue systems. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2017, pp. 1676–1680. https://doi.org/10.21437/Interspeech.2017-1593.

Morency, L.P., de Kok, I., Gratch, J., 2010. A probabilistic multimodal approach for predicting listener backchannels. Auton. Agent Multi. Agent Syst. 20 (1), 70–84. https://doi.org/10.1007/s10458-009-9092-y.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P.R., Morgan, J., Pollack, M.E. (Eds.), Intentions in Communication. MIT Press, Cambridge, MA, pp. 271–311.

Quilis, A., 1993. Tratado de fonología y fonética españolas. Gredos, Madrid.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13–17-August-2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Roddy, M., Skantze, G., Harte, N., 2018. Multimodal continuous turn-taking prediction using multiscale RNNs. Proceedings of the 2018 International Conference on Multimodal Interaction. ACM, pp. 186–190. https://doi.org/10.1145/3242969.3242997.

Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Schegloff, E.A., 2006. Interaction: the infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In: Enfield, N.J., Levinson, S.C. (Eds.), Roots of Human Sociality. Berg, London, pp. 70–96.

Sicoli, M.A., Stivers, T., Enfield, N.J., Levinson, S.C., 2015. Marked initial pitch in questions signals marked communicative function. Lang. Speech 58 (2), 204–223. https://doi.org/10.1177/0023830914529247.

Silverman, B.W., 2018. Density Estimation: For Statistics and Data Analysis. Routledge, London. https://doi.org/10.1201/9781315140919.

Skantze, G., 2018. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pp. 220–230. https://doi.org/10.18653/v1/w17-5527.

Starkey, D., Fiske, D.W., 1977. Face-to-Face Interaction: Research, Methods, and Theory. Lawrence Erlbaum Associates.

Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.E., Levinson, S.C., 2009. Universals and cultural variation in turn-taking in conversation. Proc. Natl. Acad. Sci. U.S.A. 106 (26), 10587–10592. https://doi.org/10.1073/pnas.0903616106.

Szczepek Reed, B., 2014. Phonetic practices for action formation: glottalization versus linking of TCU-initial vowels in German. J. Pragmat. 62, 13–29. https://doi.org/10.1016/j.pragma.2013.12.001.

Tolins, J., Fox Tree, J.E., 2014. Addressee backchannels steer narrative development. J. Pragmat. 70, 152–164. https://doi.org/10.1016/j.pragma.2014.06.006.

Toloşi, L., Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27 (14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300.

Truong, K.P., Poppe, R., Heylen, D., 2010. A rule-based backchannel prediction model using pitch and pause information. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3058–3061.

Vidal de Battini, B.E., 1964. El español en la Argentina. Consejo Nacional de Educación, Buenos Aires.

Ward, N.G., 2019. Turn-Taking Constructions. Cambridge University Press, Cambridge, MA. https://doi.org/10.1017/9781316848265.011.

Watson-Gegeo, K.A., Bauman, R., Sherzer, J., 1976. Explorations in the Ethnography of Speaking, vol. 52. Cambridge University Press, Cambridge, MA. https://doi.org/10.2307/412740.

Wennerstrom, A., Siegel, A.F., 2003. Keeping the floor in multiparty conversations: intonation, syntax, and pause. Discourse Process. 36 (2), 77–107. https://doi.org/10.1207/s15326950dp3602_1.