A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora

Ramiro H. Gálvez[1,2], **Lara Gauder**[1,2], **Jordi Luque**[3], **Agustín Gravano**[1,2]
[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Instituto de Ciencias de la Computación, CONICET-UBA, Buenos Aires, Argentina
[3] Telefonica Research, Spain
{rgalvez, mgauder, gravano}@dc.uba.ar, jordi.luqueserrano@telefonica.com

Abstract

Acoustic/prosodic (a/p) entrainment has been associated with multiple positive social aspects of human-human conversations. However, research on its effects is still preliminary, first because how to model it is far from standardized, and second because most of the reported findings rely on small corpora or on corpora collected in experimental setups. The present article has a twofold purpose: 1) it proposes a unifying statistical framework for modeling a/p entrainment, and 2) it tests on two large corpora of spontaneous telephone interactions whether three metrics derived from this framework predict positive social aspects of the conversations. The corpora differ in their spoken language, domain, and positive social outcome attached. To our knowledge, this is the first article studying relations between a/p entrainment and positive social outcomes in such large corpora of spontaneous dialog. Our results suggest that our metrics effectively predict, up to some extent, positive social aspects of conversations, which not only validates the methodology, but also provides further insights into the elusive topic of entrainment in human-human conversation.

1 Introduction

A phenomenon that has been repeatedly documented in human-human conversations is the tendency of partners to become more similar to each other in the way they speak. This behavior, known in the literature as *entrainment*, has been shown to occur along several dimensions during human-human interaction (see Pardo, 2006; Brennan and Clark, 1996; Reitter et al., 2011; Levitan et al., 2015; Gravano et al., 2015; Fandrianto and Eskenazi, 2012, inter-alia), being one of these dimensions the behavior of *acoustic-prosodic* (a/p) features (see, for example, Ward and Litman, 2007; Levitan and Hirschberg, 2011).

A/p entrainment has been associated with multiple social aspects of human-human conversations, such as the degree of success in completing tasks (Nenkova et al., 2008; Reitter and Moore, 2014), the perception of competence and social attractiveness (Street, 1984; Levitan et al., 2011; Štefan Beňuš et al., 2014; Michalsky and Schoormann, 2017; Schweitzer and Lewandowski, 2014), and the degree of speaker engagement (De Looze et al., 2014; Gravano et al., 2015). Nonetheless, empirical evidence also points toward these relations being quite complex. As an example, *disentrainment* (speakers actively adapting to become more dissimilar to each other) has also been associated with positive social aspects in conversations (see, for example, Healey et al., 2014; De Looze et al., 2014; Pérez et al., 2016).

In spite of these advances, research on the effects of a/p entrainment is still preliminary. In first place, because the way a/p entrainment in conversations is modeled is far from standardised. As an illustrative example, when estimating a/p entrainment metrics, some studies first approximate the evolution of each speaker's a/p features and then use these approximations to calculate a/p entrainment metrics (Gravano et al., 2015; De Looze et al., 2014; Kousidis et al., 2009; Pérez et al., 2016); others study the correspondence between adjacent inter-pausal units (IPUs) — defined as speech segments separated by a pause — from different speakers and derive metrics from it (Levitan and Hirschberg, 2011; Weise et al., 2019); and still other studies measure a/p features in different sections of speech for later comparing these values to compute a/p entrainment metrics (see, for example, Savino et al., 2016). Moreover, studies commonly differ in which metrics are analyzed. For this reason, a reliable, simple, general, and flexible framework able to unify the estimation of different types of entrainment metrics is needed. In second place, research is still preliminary because most of the reported findings rely on small corpora, or on corpora collected in experimental setups, making it hard to extrapolate their conclusions to more general contexts. In this way, evidence is still needed on how a/p entrainment relates to social aspects of human-human conversation under different types of natural interactions.

The present article has a twofold purpose. First, it proposes a unifying approach for modeling a/p entrainment in conversations. The methodology is simple and flexible enough as to allow calculating adapted versions of several a/p entrainment metrics used in previous studies. Second, it evaluates three entrainment metrics derived from the proposed framework on two very different large corpora of spontaneous telephone interactions (the Switchboard corpus, in English, and a large collection of call-center conversations, in Spanish), testing whether these metrics predict positive social aspects of conversations. To our knowledge, this is the first article testing the relation between a/p entrainment and positive social outcomes in such large corpora of spontaneous natural dialog.

Overall, our results suggest that metrics derived from the proposed methodology effectively predict, up to some extent, positive social aspects in conversations, which not only validates the methodology, but provides further evidence suggesting that a/p entrainment relates to positive social aspects in human-human conversation under different types of natural settings. Additionally, insights on how a/p entrainment metrics relate to social outcomes predictions is provided.

The rest of the paper is structured as follows. Section 2 presents the proposed framework for modeling a/p entrainment. Section 3 details on how we empirically test for relations between metrics obtained using the proposed methodology and positive social aspects of conversations. Section 4 presents results from the empirical study. Section 5 provides discussion and concludes.

2 A unifying framework for modeling a/p entrainment

Here we present a methodology for modeling a/p entrainment. We divide the process in three steps: 1) extracting a/p features from IPUs, 2) estimating the speakers' a/p evolution functions, and 3) calculating a/p entrainment metrics from a/p evolution functions. The following sections describe each step.

2.1 Extracting a/p features from IPUs

First, for each speaker in a conversation (A and B for exposition) all of their IPUs are identified.[1] Then, for each IPU the value of their a/p features are extracted. In this study we used the Praat toolkit (Boersma and Weenink, 2019) to estimate the IPU's F0 maximum and mean; intensity max and mean; noise-to-harmonics ratio (NHR); and jitter and shimmer (computed over voiced frames only). We also extracted speech rate, measured in syllables per seconds.[2]

Figure 1 plots the F0 mean values for all IPUs in a sample Switchboard conversation. Each horizontal segment represents an IPU, graphically indicating its beginning and end times.

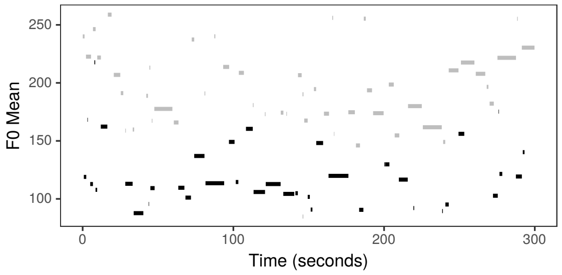

Figure 1: Estimated F0 mean values for all IPUs in a sample Switchboard conversation. *Note*: speaker A in gray, speaker B in black.

2.2 Estimating speakers' a/p evolution functions

Since speakers do not speak during the entirety of a conversation, the evolution of a/p features is undefined for several portions of a conversation. This stands as a challenge when modeling a/p entrainment. Previous research has dealt with this issue in multiple ways; for example, by pairing speakers adjacent IPUs (Levitan and Hirschberg, 2011; Weise et al., 2019) or by means of sliding windows (Kousidis et al., 2009; De Looze et al., 2014; Pérez et al., 2016). Instead, we propose filling these gaps by fitting a continuous function to approximate the evolution of a given a/p feature during a conversation. We do this by fitting a k-nearest neighbors

[1] For the Switchboard corpus we used the MS-State transcripts (Deshmukh et al., 1998), where IPUs are annotated. For the call center conversations we used the output of an in-house automatic speech recognition system (Cartas et al., 2019), defining an IPU as a continuous segment of speech without a pause larger than 200 ms.

[2] Syllables were estimated using the *Pyphen* package (Pyphen, 2019).

(KNN) regression model to each speaker's a/p feature values. Where, for each IPU, its x value is defined as its middle point in time (i.e., its start time plus its end time, divided by two). We refer to these estimated functions as f^A and f^B below.[3] As we show below, adapting existing a/p entrainment metrics to take these functions as input is straightforward.

A few considerations should be made regarding the way to do these approximations. Due to the presence of outliers, before fitting these functions, all IPUs having an associated value more than three standard deviations away from the mean are dropped (the corresponding mean and standard deviation are measured at the conversation level for each speaker). Second, f^A is defined for the interval $[t^A_{min}, t^A_{max}]$, where t^A_{min} is the start time of A's first non-outlier IPU, and t^A_{max} is the end time of A's last non-outlier IPU (analogously for f^B). Third, we define the *common support* as all time values that go from $t^- = max(t^A_{min}, t^B_{min})$ up to $t^+ = min(t^A_{max}, t^B_{max})$ (i.e., all values of t where both functions are simultaneously defined). Fourth, approximations for speakers that do not have at least k non-outlier IPUs in a conversation are not computed (being k the number of neighbors used to estimate the functions).

Figure 2 plots the estimated approximation function for the IPUs plotted in Figure 1.

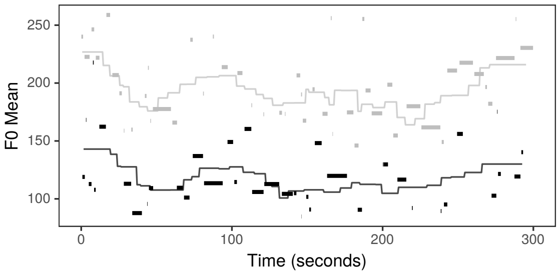

Figure 2: Estimated F0 mean evolution during a conversation. *Notes*: speaker A in gray, speaker B in black. The number of neighbors in the KNN regressions equals 7.

2.3 Calculating a/p entrainment metrics from a/p evolution functions

Existing a/p entrainment metrics can be easily adapted to take these functions as input. In this work we adapt and empirically test the three metric types presented in Levitan and Hirschberg (Levitan and Hirschberg, 2011): 1) *proximity* (a/p features having similar mean values across partners over the entire conversation), 2) *convergence* (a/p features increasing in similarity across partners over time), and 3) *synchrony* (speakers adjusting the values of their a/p features in accordance to those of their interlocutor).

2.3.1 Proximity

Proximity between f^A and f^B ($prox^{A,B}$) can be measured as the negated absolute difference of the mean values of f^A and f^B, that is:

$$-|\bar{f^A} - \bar{f^B}|$$

where, in general, $\bar{g}$ stands for the mean value of function g over the common support, and is calculated as:[4]

$$\bar{g} = \frac{1}{t^+ - t^-} \int_{t^-}^{t^+} g(t)dt$$

Values of $prox^{A,B}$ close to zero indicate that f^A and f^B are on average close to each other, while values far from zero indicate that they are distant.

2.3.2 Convergence

Convergence between f^A and f^B ($conv^{A,B}$) can be measured as the Pearson correlation coefficient between $-|f^A - f^B|$ and t, which can be calculated as:

$$\frac{\int_{t^-}^{t^+} (D(t) - \bar{D}) \cdot (t - \bar{t})dt}{\sqrt{\int_{t^-}^{t^+} (D(t) - \bar{D})^2 dt \cdot \int_{t^-}^{t^+} (t - \bar{t})^2 dt}}$$

where $D(t)$ stands for $-|f^A(t) - f^B(t)|$. Positive/negative values of this metric indicate that f^A and f^B become closer to/further apart from each other as the conversation advances.

2.3.3 Synchrony

Synchrony between f^A and f^B ($sync^{A,B}$) can be measured as the Pearson correlation coefficient between f^A and f^B. Given that speakers are not expected to adapt to the other instantaneously, several studies consider a lag factor (δ) when calculating synchrony (see, for example, Kousidis et al., 2009; Pérez et al., 2016). In this study we also incorporate lags, which is a small departure from Levitan and Hirschberg (Levitan and Hirschberg, 2011). Concretely, we calculated $sync^{A,B}$ as:

[3] As we will see in Section 3.2, k (the number of neighbors) can be treated as a hyperparameter to be tuned during the model selection procedure.

[4] In our empirical study, integrals are calculated using the Monte Carlo integration method.

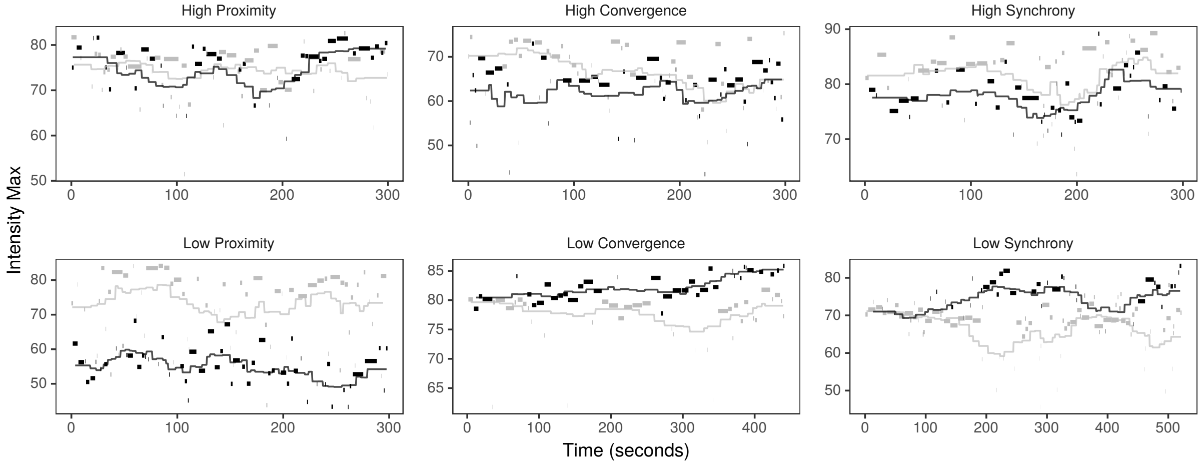

Figure 3: Sample conversations with different values of the estimated a/p entrainment metrics on intensity max. *Notes*: speaker A in gray, speaker B in black. The number of neighbors in the KNN regressions equals 7.

$$\frac{\int_{t^-}^{t^+}(f^A(t+\delta)-\bar{f^A})\cdot(f^B(t)-\bar{f^B})dt}{\sqrt{\int_{t^-}^{t^+}(f^A(t+\delta)-\bar{f^A})^2dt\cdot\int_{t^-}^{t^+}(f^B(t)-\bar{f^B})^2dt}}$$

where, in order to capture lags in synchrony between both functions, we test values of $\delta \in \{-15,-10,-5,0,5,10,15\}$ (being δ expressed in seconds). We take as the final value of $sync^{A,B}$ the one associated with δ resulting in the maximum value of $|sync^{A,B}|$.[5] Positive values of $sync^{A,B}$ indicate that f^A and f^B evolve in synchrony with each other, while negative values indicate that they evolve in opposite directions (e.g., when one goes up the other goes down).

To illustrate the kind of behavior captured by these metrics, Figure 3 plots sample conversations with high and low values of these three a/p entrainment metrics calculated on the 'intensity max' feature.

3 Empirical study materials and methods

Next, we ran a series of machine learning experiments aimed at investigating whether the metrics derived from the proposed methodology have any predictive power over positive social aspects of conversations. This section describes the corpora and the methodology used.

3.1 Corpora

3.1.1 Switchboard corpus

The Switchboard Corpus (SWBD) (Godfrey et al., 1992) is a collection of 2,438 recordings of spontaneous two-sided telephone conversations among 543 speakers (both female and male) from all areas of the United States. During collection, a robot operator system handled the calls, gave the caller appropriate recorded prompts, selected and dialed the callee, introduced one of about 70 topics for discussion (internally referred as IVIs), and recorded the whole speech from the two subjects into separate channels. Each conversation was annotated for degree of naturalness on Likert scales from 1 (very natural) to 5 (not natural at all).[6] For this corpus, **perceived naturalness** is the target social outcome to predict.

After dropping from the analysis a few conversations for which naturalness scores were missing, we were left with a total of 2,426 conversations (average conversation length: 382.3 seconds; SD: 124.8 seconds). To make the analysis and results more interpretable (more on this in Section 3.2), we dichotomized the naturalness scores in the following way: We treated values 1 and 2 as *high* scores (which we set equal to 1 — 88.4% of all conversations) and values from 3 to 5 as *low* scores (which we set to 0 — 11.6% of all conversations).

In addition to the proposed metrics, our experiments included features referred to as *external features*, which are expected to be linked to the naturalness score, but are unrelated to the speakers' a/p features. These features are used for building baseline models to compare against. For the case of SWBD these variables are: IVI indicator

[5] Note that shifting a series slightly modifies the common support, something that should be taken into account.

[6] More details on naturalness annotations available at https://catalog.ldc.upenn.edu/docs/LDC97S62/swb1_manual.txt.

variables (indicating the IVI used as the conversation topic), conversation length (in seconds), dialect area indicator variables (indicating whether at least one speaker belonged to a given dialect area, and whether both speakers belonged to the same dialect area), three gender indicator variables (both-female-speakers, both-male-speakers, mixed-gender-speakers), and the absolute value of the age difference between the speakers (in years).

3.1.2 Call center corpus

The call center corpus (CCC) is a collection of of 19,832 inbound call center conversations between clients and representatives of a telephone company (for further details see Llimona et al., 2015; Luque et al., 2017) (average conversation length: 551.7 seconds; SD: 432.9 seconds). It was collected throughout one month and comprises a huge variety of interactions. All conversations are in Latin American Spanish. At the end of each call, the customer was called back and gently asked to complete a brief service quality survey. Concretely, they had to indicate their overall satisfaction with respect to their previous call center call. To do so, they had to press from 1 (very dissatisfied) to 5 (very satisfied). For this corpus, **self-reported customer satisfaction** is the target social outcome to predict.

We again dichotomized the target variable in the following way: 4 and 5 were treated as *high* scores (which we set equal to 1 — 80.5% of all conversations), and values from 1 to 3 were treated as *low* scores (which we set to 0 — 19.5% of all conversations).

For anonymity reasons, the availability of external variables was more limited for CCC. Thus, only conversation length and the three gender indicator variables (Llimona et al., 2015) were included as external features.

3.2 Testing for associations between a/p entrainment and social outcomes

To test if the proposed entrainment metrics predict the outcomes, we ran a series of machine learning experiments. For each corpus we trained several XGBoost models (Chen and Guestrin, 2016)[7] using different feature sets, and evaluated their predictive performance. For example, one such model used only the synchrony metrics computed on the 8 a/p features described in Section 2.1 (F0 max, F0 mean, intensity max, intensity mean, NHR, jitter, shimmer, speech rate); other model considered all 24 entrainment metrics (8 a/p features × 3 metric types); other model considered only external features; and so on.

As the evaluation metric, we used the area under the receiver operating characteristic curve (AUC) (see Alpaydin, 2020). AUC goes from 0 to 1, where an AUC value equal to 0.5 indicates an equal-than-chance performance, while larger values indicate that the learning model effectively predicts the outcomes, up to some extent. To obtain our *out-of-sample* performance estimates we ran 10-fold cross validation experiments (see James et al., 2014). We tuned the hyperparameters following a random search strategy (Bergstra and Bengio, 2012): For each value of $k \in \{3, 5, 7, 9\}$ (number of neighbors used in the functional approximations) we tested 60 randomly sampled combinations of seven XGBoost hyperparameters.[8] The chosen hyperparameters are those for which the model had the higher cross validation performance.

3.2.1 Model interpretability

Comparing performance across models provides valuable information regarding feature importance. However, further valuable information can be obtained by interpreting the models' inner workings. To do so, several strategies have been proposed (see Molnar, 2019). In our analysis we made use of the Shapley additive explanations (SHAP) technique (Lundberg and Lee, 2017). SHAP values are calculated for each observation and predictive feature in the dataset used to train the model being analyzed. Concretely, a given SHAP value $\phi_{i,j}$ estimates, for observation i, how feature j contributes to push the model output (in logit scale) from its base output (being the base output equal to the average model output over the training dataset). In this way, SHAP values can be used to estimate feature importance for a given feature j by calculating $\sum_i |\phi_{i,j}|$. They can also characterize how the outputs diverge from the base output as feature j grows, by using SHAP feature dependence plots (that is, plotting $\phi_{i,j}$ against all observed values of feature j).[9]

[7] XGBoost is an open-source and efficient software implementation of the gradient boosting framework (Friedman et al., 2001). XGBoost has the additional advantage of dealing with missing values, which, in our analysis, were present both in the external features and the a/p entrainment metrics.

[8] Number of trees; tree depth; step size shrinkage coefficient; minimum loss reduction required to make a further partition; minimum child weight; number of columns sampled in each tree; and number of observations sampled in each tree.

[9] It is important to stress that any pattern derived from the model interpretability analysis does not imply that a feature

4 Empirical study results

4.1 Predictive performance

We trained models on eight different sets of inputs: 1) only proximity metrics, 2) only convergence metrics, 3) only synchrony metrics, 4) all a/p entrainment metrics, 5) external features and proximity metrics, 6) external features and convergence metrics, 7) external features and synchrony metrics, and 8) external features and all a/p entrainment metrics. The rest of this section presents the performances obtained for each corpus.

4.1.1 Switchboard corpus results

Table 1 presents the estimated performance for each set of input features. The top panel presents results excluding external features, while the bottom one includes them. Within each panel, models are sorted in descending AUC order.

Input Features	**AUC**
Excluding external features	
Only synchrony	**0.575**
All a/p entrainment metrics	0.566
Only proximity	0.561
Only convergence	0.547
Including external features	
External and synchrony	**0.641**
External and convergence	0.631
External and all a/p entrainment metrics	0.630
Only external	0.627
External and proximity	0.624

Table 1: Switchboard corpus AUC results

Table 1 shows that the trained models are able to predict up to some extent perceived naturalness. In all cases the obtained results are higher than chance (i.e., AUC > 0.5). But not all features have the same predictive performance. Synchrony entrainment metrics obtain the best results. Training on just synchrony metrics results in an AUC of 0.575, while using only convergence or proximity metrics leads to lower AUC values. Training on all a/p entrainment metrics results in an AUC of 0.566, lower than the one obtained with synchrony metrics.

Training only on external features results in an AUC of 0.627, higher than all values presented in the top panel. However, adding synchrony metrics to the external features is the combination that leads to the best overall results.

4.1.2 Call center corpus results

Input Features	**AUC**
Excluding external features	
All a/p entrainment metrics	**0.582**
Only synchrony	0.560
Only convergence	0.556
Only proximity	0.548
Including external features	
External and all a/p entrainment metrics	**0.582**
External and proximity	0.568
External and synchrony	0.564
External and convergence	0.560
Only external	0.537

Table 2: Call center corpus AUC results

For CCC, Table 2 shows that the trained models are also able to predict up to some extent self-reported customer satisfaction. However, we observe that combining all 24 entrainment metrics leads to better results (AUC $= 0.582$) than including just the synchrony ones (AUC $= 0.560$).

In this case the external features have low predictive power when compared to the entrainment metrics. Adding the external features to the model considering all a/p entrainment metrics yields exactly the same results as the ones obtained by the model trained only on all a/p entrainment metrics.

4.2 Model interpretability results

Switchboard corpus	
Input Feature	Feature Importance
Both-female-speakers	100.0
Speech-rate-synchrony	45.7
Conversation-length	27.8
Intensity-mean-synchrony	19.1
Jitter-synchrony	18.8
F0-max-synchrony	16.4
Age-difference	16.0
Shimmer-synchrony	9.9
F0-mean-synchrony	9.1
Intensity-max-synchrony	6.3
Call center corpus	
Input Feature	Feature Importance
Speech-rate-proximity	100.0
Speech-rate-synchrony	62.3
Conversation-length	54.9
Intensity-max-convergence	47.3
Jitter-convergence	36.6
F0-mean-proximity	30.2
Jitter-proximity	26.5
Speech-rate-convergence	23.8
F0-mean-convergence	19.9
NHR-convergence	19.9

Table 3: Feature importance ranking. *Note*: values are scaled such that the score associated to the most important feature equals 100.

has a causal relationship with the outcome. It merely indicates that a given feature causes **the model** to predict the outcome in a particular way (see Molnar, 2019).

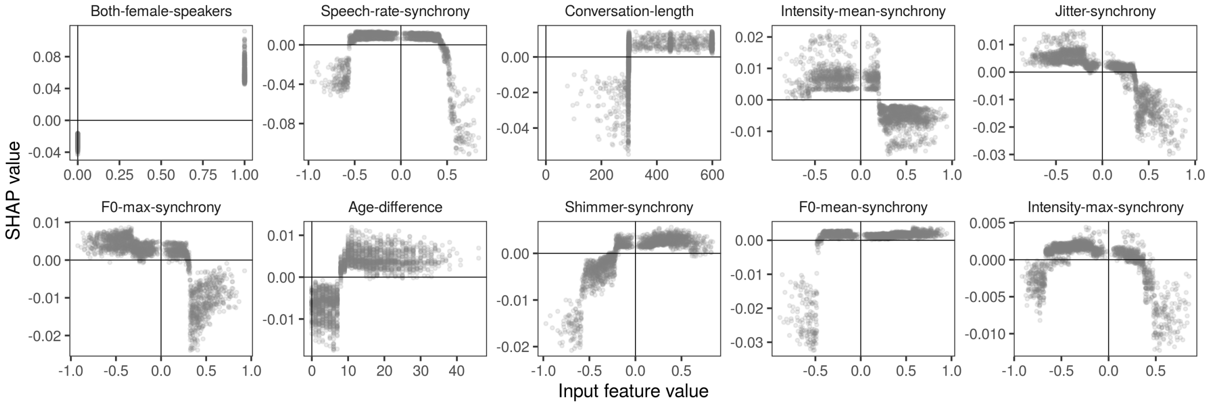

Figure 4: SHAP feature dependence plots for SWBD.

Table 3 contains the estimated feature importance values for both corpora analyzed. Each panel shows the top 10 features detected as most important in the best performing models (*external and synchrony* for SWBD; *external and all a/p entrainment metrics* for CCC). Both models were trained using the entirety of their respective corpus and making use of the best set of hyperparameters previously found.

For SWBD, Table 3 shows that the *both-female-speakers* indicator variable was by far the most important feature, followed by *speech-rate-synchrony*. For the case of CCC, speech rate entrainment metrics dominate the importance ranking, being *speech-rate-proximity* the most important one. Notably, no gender related feature is included in the ranking of the top 10 most important features for CCC.

Feature importance values are interesting in and of themselves, but say little of the way models make use of these features. To tackle this issue, Figure 4 presents SHAP feature dependence plots for SWBD's 10 most important features. Horizontal lines centered at SHAP $= 0$ serve as a reference; values appearing above/below this line indicate that the model output tends to increase/decrease relative to the base output. Regarding the external features, Figure 4 shows that conversations in which both speakers are female were associated to higher values of predicted naturalness, that short conversations are predicted to be less natural (speakers were instructed to speak for at least five minutes, but were allowed to speak longer), and that large age differences lead to higher naturalness predictions.

It is interesting to note that high synchrony values are **not** necessarily associated with higher perceived naturalness predictions. It is the case for *shimmer-synchrony* and *F0-mean-synchrony* (to a lesser extent). However, the opposite is observed for many a/p features. Moreover, *speech-rate-synchrony* and *intensity-max-synchrony* show an inverted U pattern, where extremely low or high entrainment values are associated to lower predicted values of the outcome.

Figure 5 presents a similar analysis for CCC. Regarding *conversation length*, the only high-ranked external feature, extremely short conversations are associated to lower predictions of self-reported satisfaction. Regarding a/p entrainment metrics, once again higher entrainment does not necessarily lead to higher predictions of the outcome variable. In fact, this is not the case for all entrainment metrics related to speech rate. Only *intensity-max-convergence* shows a positive relation between a/p entrainment and predicted satisfaction. Once again negative relations and inverted U patterns are observed (although the latter are less noticeable than in SWBD).

5 Discussion

In this work we proposed a unifying framework for modeling different types of a/p entrainment in natural conversations. We also tested on two very different corpora whether three metrics derived from our framework provide valuable information for predicting positive social outcomes in conversations (perceived naturalness in SWBD and self-reported customer satisfaction in CCC). Our results suggest that these metrics effectively relate to positive social outcomes. However, several remarks should be made.

First, the fact that the achieved AUC scores are greater than chance not only validates the proposed

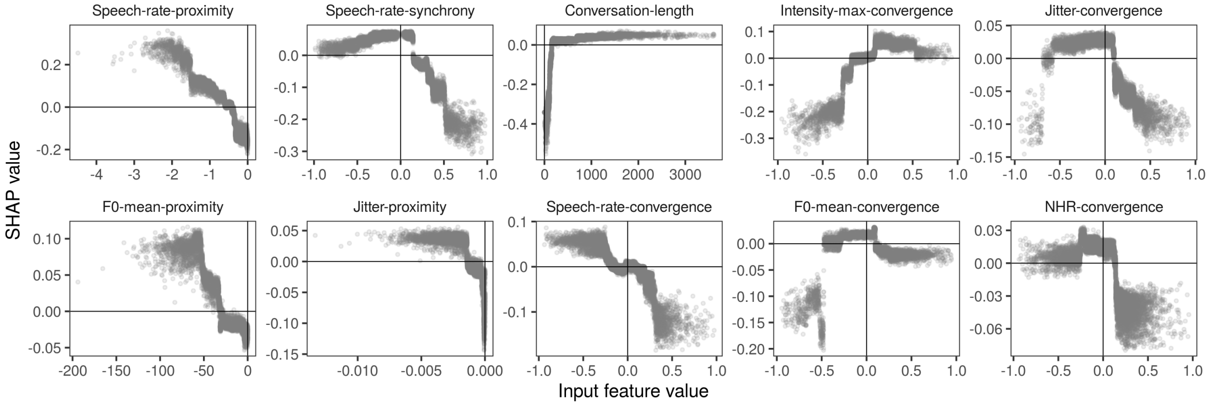

Figure 5: SHAP feature dependence plots for CCC.

metrics, but strongly suggests that a/p entrainment is related to positive social outcomes. Importantly, this result was not found on a single corpus, but on two independent ones. Both corpora had not only different positive social outcomes attached, but also differed in their domain and even in their language (English and Spanish). Future research should focus on testing if these results prevail across broader domains and for further social variables.

Second, even when the obtained results are higher than chance, they are far from being exceptionally high. This suggests that a/p entrainment metrics by themselves — at least the ones tested — may not contain enough predictive information as to achieve competitive results. Probably a competitive model should incorporate information regarding the semantic content of conversations and/or, for the case of a corpus like CCC, customer relationship management related information. However, this does not mean that the proposed metrics are of no use. Future research should focus on studying whether the information provided by a/p entrainment metrics complements the one provided by other sources. In our experiments we tested this up to some degree. In particular, we observed that a/p entrainment metrics complement the information contained in the external features. This effect is very strong for CCC and less strong for SWBD.

Third, the fact that the best set of features differs across corpora, suggests that which features predict positive social outcomes depends on the outcomes being predicted and the corpus itself. Note that a similar pattern was observed in Pérez et al. (Pérez et al., 2016) where, even on the same corpus, the significance of synchrony metrics calculated on different a/p features varied across different social outcomes.

Fourth, SHAP dependence plots suggest that the manner in which predictive models make use entrainment metrics is quite complex. First of all, not always are higher entrainment values associated to higher predicted values of a positive social aspect. Rather, two more patterns are observed: a negative relation between a/p entrainment and positive outcomes, and an inverted U pattern. Additionally, in line with the third remark, it is interesting to note the effects of a given a/p entrainment metric are not the same across corpora, again suggesting heterogeneity across tasks and corpora. An illustrative case are the patterns observed for *speech-rate-synchrony*, for which an inverted U pattern is observed in SWBD and a negative relation is observed in CCC.

Finally, the reason why people do entrain is still unknown (see, for example, Natale, 1975; Giles et al., 1991; Chartrand and Bargh, 1999; Pickering and Garrod, 2004, 2013). Consequently, metrics such as the ones tested in this work, albeit noisy and imperfect, are likely to be capturing part of some more complex phenomenon that we do not fully understand yet. Further research on the causes of entrainment in human speech is still needed.

References

Ethem Alpaydin. 2020. *Introduction to Machine Learning*, 4 edition. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305.

Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia

Hirschberg. 2014. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 71:3 – 14.

Paul Boersma and David Weenink. 2019. Praat: doing phonetics by computer [computer program]. Version 6.1.08, retrieved 5 December 2019 from `http://www.praat.org/`.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Alejandro Cartas, Martin Kocour, Aravindh Raman, Ilias Leontiadis, Jordi Luque, Nishanth Sastry, Jose Nuñez Martinez, Diego Perino, and Carlos Segura. 2019. A reality check on inference at mobile networks edge. In *Proceedings of the 2nd International Workshop on Edge Systems, Analytics and Networking*, EdgeSys '19, page 54–59, New York, NY, USA. Association for Computing Machinery.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Céline De Looze, Stefan Scherer, Brian Vaughan, and Nick Campbell. 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11 – 34.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. 1998. Resegmentation of switchboard. In *ICSLP-1998*.

Andrew Fandrianto and Maxine Eskenazi. 2012. Prosodic entrainment in an information-driven dialog system. In *INTERSPEECH-2012*, pages 342–345.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. *Accommodation theory: Communication, context, and consequence.*, Studies in emotion and social interaction., pages 1–68. Editions de la Maison des Sciences de l'Homme, Paris, France.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *ICASSP-92*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

Agustín Gravano, Štefan Beňuš, Rivka Levitan, and Julia Hirschberg. 2015. Backward mimicry and forward influence in prosodic contour choice in standard american english. In *INTERSPEECH-2015*, pages 1839–1843.

Patrick G. T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PLOS ONE*, 9(6):1–6.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Spyros Kousidis, David Dorran, Ciaran Mcdonnell, and Eugene Coyle. 2009. Time series analysis of acoustic feature convergence in human dialogues. In *SPECOM-2009*, St. Petersburg, Russian Federation.

Rivka Levitan, Stefan Benus, Agustín Gravano, and Julia Hirschberg. 2015. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*.

Rivka Levitan, Agustín Gravano, and Julia Hirschberg. 2011. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 113–117, Portland, Oregon, USA. Association for Computational Linguistics.

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

Quim Llimona, Jordi Luque, Xavier Anguera, Zoraida Hidalgo, Souneil Park, and Nuria Oliver. 2015. Effect of gender and call duration on customer satisfaction in call center big data. In *INTERSPEECH-2015*, pages 1825–1829.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Jordi Luque, Carlos Segura, Ariadna Sánchez, Martí Umbert, and Luis Angel Galindo. 2017. The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls. In *INTERSPEECH-2017*, pages 2346–2350.

Jan Michalsky and Heike Schoormann. 2017. Pitch convergence as an effect of perceived attractiveness and likability. In *INTERSPEECH-2017*, pages 2253–2256.

Christoph Molnar. 2019. *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172, Columbus, Ohio. Association for Computational Linguistics.

Jennifer S. Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Pyphen. 2019. Pyphen is a pure python module to hyphenate text. retrieved: 2019-01-24.

Juan M. Pérez, Ramiro H. Gálvez, and Agustín Gravano. 2016. Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement. In *INTERSPEECH-2016*, pages 1270–1274.

David Reitter, Frank Keller, and Johanna D. Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive Science*, 35(4):587–637.

David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29 – 46.

Michelina Savino, Loredana Lapertosa, Alessandro Caffò, and Mario Refice. 2016. Measuring prosodic entrainment in italian collaborative game-based dialogues. In *Speech and Computer*, pages 476–483, Cham. Springer International Publishing.

Antje Schweitzer and Natalie Lewandowski. 2014. Social factors in convergence of f1 and f2 in spontaneous speech. In *ISSP-2014*.

Richard L. Street. 1984. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169.

Arthur Ward and Diane Litman. 2007. Measuring convergence and priming in tutorial dialog. *University of Pittsburgh*.

Andreas Weise, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. 2019. Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, 115:78 – 87.