Neural Representations of Dialogical History for Improving Upcoming Turn Acoustic Parameters Prediction

Simone Fuscone, Benoit Favre, Laurent Prevot

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Neural representations of dialogical history for improving upcoming turn acoustic parameters prediction

Simone Fuscone[1,2]*, Benoit Favre*[2] *and Laurent Prévot*[1,3]

[1] Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[2] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[3] Institut Universitaire de France, Paris, France

simone.fuscone@univ-amu.fr benoit.favre@lis-lab.fr laurent.prevot@univ-amu.fr

Abstract

Predicting the acoustic and linguistic parameters of an upcoming conversational turn is important for dialogue systems aiming to include low-level adaptation with the user. It is known that during an interaction speakers could influence each other speech production. However, the precise dynamics of the phenomena is not well-established, especially in the context of natural conversations. We developed a model based on an RNN architecture that predicts speech variables (Energy, F0 range and Speech Rate) of the upcoming turn using a representation vector describing speech information of previous turns. We compare the prediction performances when using a dialogical history (from both participants) vs. monological history (from only upcoming turn's speaker). We found that the information contained in previous turns produced by both the speaker and his interlocutor reduce the error in predicting current acoustic target variable. In addition the error in prediction decreases as increases the number of previous turns taken into account.
Index Terms: convergence, prediction, acoustic features, prosody

1. Introduction

Throughout the course of a conversation the conversants, the one who has the floor alias the *-'speaker'-* and his partner the *-'interlocutor'-*, are constantly changing their speech production and its underlying features. These changes are partly due to convergence phenomena, the tendency of conversants to coadjust their speaking styles. Such phenomena have been established at various levels including acoustic and prosodic level (Energy) [1], [2]; (Fundamental Frequency) [3], [4]; Speech Rate [5].

Previous works have explored the correlation between the speech values produced by the speaker and his interlocutor, trying to explain the influence that a speaker has on its interlocutor's production and vice versa. We aim to expand the work that have been done in this direction. Our approach is to build a regression problem that consists in predicting some acoustics parameters of the upcoming turn using information contained in previous turns. Our method to study convergence consists in the estimation of the influence that the speech style of the interlocutor has on the speech style of the speaker in the upcoming turn.

The understanding of convergence mechanisms is crucial in the development of virtual agents for human robot interaction. Developing virtual agents that mirror human behavior could improve the success of communication between humans and virtual agents. Past literature showed that convergence is higher with human peer than a simple virtual agent ([6, 7]) and that a system that converges to the human speaker increases the success in accomplishment of a task ([8]) or that ([9]) speakers tend to ask advice mostly to systems that converge with them.

In this paper we introduce an exploratory methodology to study convergence by evaluating whether using information contained in the previous turns produced by both *speaker* and *interlocutor* leads to have better prediction of upcoming turn acoustic parameters than using information of previous turns produced by just the speaker. The paper starts with a review of related works (Section 2) that focused on the influence that conversants play on each other. Then we describe the model, the data and the feature extraction methods in Section 3. Using the Switchboard data set we present the experiments and results we obtained using as target the mean energy, pitch range and speech rate (Section 4). Finally, we discuss possible improvement and extensions of this approach (Section 5).

2. Related work

The target variables we scrutinize in this study -**Energy (E), Pitch range (F0)** and **Speech Rate (SR)**- were object of study in previous works. E is the speech variable regularly cited that exhibits convergence effects between speakers in both experimental [10], [11], [12] [13] [14] and natural conversations [15], [16]. Alongside [17, 18] describe convergence in F0 max for successfully interactions while [19] observe convergence both in average and range F0. Besides studies that measure convergence looking at the distance between the conversants some authors, at the best of our knowledge, focused on the influence between the previous productions of speaker and his interlocutor with predictive paradigms. Cohen et al. [20] use a linear mixed model to estimate average SR in a conversation using the average SR of his interlocutor. Similarly, Cohen & Sanker [21] use apply this approach to F0. In a more fine-grained approach Schweitzer and colleagues [22, 23] used SR of previous turn values to predict SR of upcoming turn using a linear mixed model. These methods account to get the correlation between the same variable (here, SR) but did not consider the relation that other speech features may have on the variable studied. We expand these studies to question the influence that each conversant has on his partner by looking if the information contained in previous history of the conversation helps to predict the evolution of acoustic features.

3. Methods

3.1. The Model

The regression problem is illustrated in Figure 1. The model takes as input the representation vector of the N ($1 \leq$

http://dx.doi.org/10.21437/Interspeech.2020-2785

$N \leq 10$) previous turns (overlapping/non-overlapping and consecutive/non-consecutive turns) that feeds as many Long Short Term Memory (LSTM) cells. The output of the hidden state of the LSTM propagates in a linear net that has one layer output of the same size. The output computes the predicted target value (mean E, F0 range and SR) of the upcoming turn. The Figure 2 presents the architecture of the model.

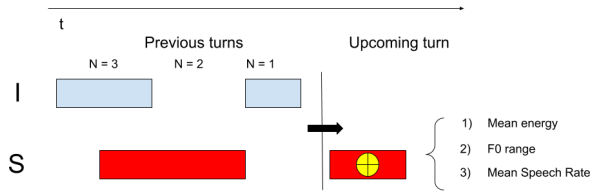

Figure 1: *A graphic representation of the task. The regression problem consists in predicting the target variable (mean E, F0 range, SR) of the next turn. The speaker (S) is the conversant that takes the floor in the next turn, while the interlocutor (I) is his partner.*

After a feature selection process, we compare the case in which we use vector representations of previous turns that belong to the **speaker (S)** of the upcoming turn and the case of using previous turns from both the conversants **speaker + interlocutor (S+I)**. In addition we compare our model with a baseline that is a *linear regression* that takes as input the average values of the target variable in previous turns.

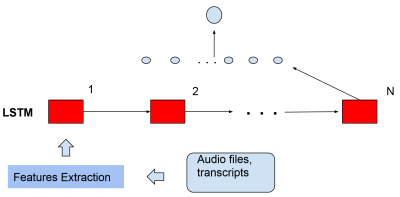

Figure 2: *LSTM layer fed with features extracted by the audio and transcripts files. The layer has many cells as N, the number of previous turns. The hidden state is then fed into a linear net that has a one layer output, that gives the predicted value of the target variable.*

3.2. Data

The Switchboard Corpus [24] is formed by spontaneous telephonic conversations in American English. The corpus consists of 2430 conversations (average duration of 6 minutes) for a total of 240 hours, involving 543 speakers. The corpus has audio, time aligned transcripts and a segmentation into *turns*. Moreover, 642 conversations have been segmented and annotated for dialogue acts (DA) [25], we use this version of DA as it contains alignment to the transcripts. A few dialogue acts (Statement: 36%, Acknowledgment: 19%, Opinion: 13%) are dominating the DA distribution (See [26] for the details). We used turn segmentation as provided by Switchboard [24], we did not apply any filtering on the turn segmentation, therefore taking into account overlapping and non overlapping turns. The distributions of turns is shown in Table 1.

Table 1: *Distribution of turns: we report for each N, the percentage of turns produced by S and I. The $S > I$ means that the S produced more turns than I, the opposite for $S < I$ while $S = I$ that S and I produced the same numbers of turns.*

N°turns	$S > I$	$S < I$	$S = I$
1	**32%**	**68%**	**-**
2	**11%**	**22%**	**67%**
3	**41%**	**59%**	**-**
4	**19%**	**30%**	**51%**
5	**44%**	**56%**	**-**
6	**25%**	**34%**	**41%**
7	**45%**	**55%**	**-**
8	**28%**	**34%**	**38%**
9	**46%**	**54%**	**-**
10	**31%**	**36%**	**33%**

3.3. Feature extraction

E and F0 are computed from the audio files with *openSMILE* audio analysis tool [27] while SR is computed using time aligned transcripts. In this section we will describe more in details the extraction and computation of the target features and the other features that are taken into account as input variables in the representation vector.

Energy (E): The mean value per each turn is computed as the average of values that have been sampled every 50 ms. To handle the distance mouth-microphone, which could vary during a telephone conversation affecting the voice intensity, we introduce a normalization factor consisting of dividing each speaker E value by the average E produced by that speaker in the entire conversation. In addition, to reduce the environmental noises, we computed the average E using the temporal windows where the probability of voicing is above 0.65.

Pitch range (F0 range): It is the distance between the max and min of F0 that were sampled every 50 ms, adopting the same filtering procedure applied for E.

Speech Rate (SR): We used the approach proposed by Cohen Priva [20] that defines SR for an utterance as the ratio between the actual duration of the utterance and its expected duration (computed by estimating every word duration into the whole corpus, for all speakers). Values above / below 1 correspond respectively to fast / slow speech compare to the average of the corpus.

Duration (Δ): It refers to the temporal length of each turn, as provided by the segmentation of the SW corpus.

Dialog acts (DA) type: We used as predictors the kind of speech activity that indicates the type of turn. From the NXT SWitchboard (SWB-NXT) [28], we developed a DA-tagger to cover the whole data set.

We simplify the tagging task by considering only 3 categories resulting from the merging of the 42 original ones: *Statement+Opinion* (STA+OPI), *Backchannel+Agreement* (BAC+AGR) and *Other* (OTH) which includes all the other DA. This grouping was obtained by first considering only the DA which dominates the distribution. Then we manually inspected many examples of each dialogue act and figured out that, although functionally different, *statements* and *opinions* on the hand *backchannel* and *Agreement* on the other hand correspond to very similar conversational activities. More precisely, the former have clear *main speaker* feeling with a lot of semantic content while the later have a much more *listener*

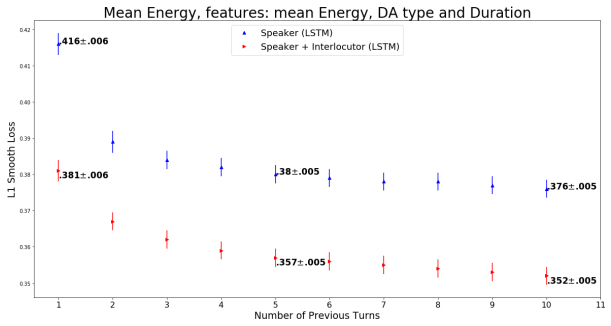

Figure 3: *L1 smooth loss while predicting mean E of upcoming turn for the setup S and S+I by the use of LSTM model with the richer representation (mean E, DA and Δ).*

nature with various kind of feedback related lexical items.

We used as train, development and test set the SWB-NXT corpus that contains annotated DA for 642 conversations. As the DA don't match the turn segmentation, we label each turn of the corpus by assigning one of the majority class, among the DA tags that forms the turn. The distribution results to be formed by 52% of STA+OPI, 25% of BAC+AGR and 23% of OTH. The model we used is described in ([29]) and inspired by the model of ([30]). It is a two levels hierarchical Neural Network *(with learning rate = 0.001, batch size = 32, max length of each turn = 80, embeddings words dimension = 200)*. In the first level each turn is treated singularly taking into account the words that form the turn while the second level is used to take into account the whole turn in the context of the conversation. Each level is a bidirectional Long Short Term (LSTM). We used 80% of switchboard data as training set, 10% for development and 10% for the test set. The F1 score of the DA tagger is BAC+AGR = 86%, STA+OPI = 87% and OTH = 55%. The F1 score of the class OTH, as expected, is low compared to the other 2 classes considering that it is formed by heterogeneous DA acts.

4. Experiments and Results

4.1. Training the Model

For each target and each setup (*S vs S+I*), we applied a random search grid to chose the learning rate of the Adam optimizer and the hidden size of the output of LSTM. We evaluate the performances on a validation set using the L1 smooth Loss *(E: lear. rate = 0.0022, size layer = 36; F0 range: lear. rate = 0.0025, size layer = 40, SR: lear. rate = 0.0022, size layer = 36)*.

4.2. Energy

Feature Selection. Our first aim is to select the variables of speech production described in 3.3 *(Energy, Pitch, Speech Rate, Duration, Dialogue Act type)* to build the representation vector that will be used as input in our model. As criterion of selection we chose the subset of variables that has the better performance over the turns. Table 2) shows that the vector representation formed by E, Δ and DA is the subset that improves significantly the performances in predicting the mean E of the upcoming turn compared to the other variables, for both the *S* and *S+I* setups. The use of the complete representation vector doesn't lead to a significant improvement of performance. We used a K-fold approach with K = 10 and use a dependent *t-test* to compare the different subsets applying the recursive features elimination method.

Table 2: *Features selection: L1 loss in predicting mean E, F0 range and SR. We report the average on the N turns for the different subsets of features described in 3.3.*

	E	
Features	S + I	S
E	0.386 ± 0.007	0.392 ± 0.009
E + DA	0.373 ± 0.007	0.389 ± 0.008
E + DA + Δ	0.360 ± 0.008	0.384 ± 0.010
E+DA+Δ+F0+SR	0.357 ± 0.008	0.382 ± 0.009
	F0 (Range)	
Features	S + I	S
F0	0.345 ± 0.023	0.346 ± 0.023
F0 + DA	0.326 ± 0.023	0.333 ± 0.023
F0 + DA + Δ	0.318 ± 0.021	0.330 ± 0.020
F0+DA+Δ+E+SR	0.317 ± 0.021	0.328 ± 0.022
	SR	
Features	S + I	S
SR	0.339 ± 0.008	0.339 ± 0.008
SR + DA	0.323 ± 0.005	0.324 ± 0.005
SR + DA + Δ	0.321 ± 0.008	0.323 ± 0.008
SR+DA+Δ+E+F0	0.321 ± 0.005	0.323 ± 0.006

Speaker vs Speaker + Interlocutor. When we include interlocutor's history, results show that the use of a representation vector formed by the selected features of both speaker and interlocutor brings to a significant ($p < 10^{-8}$) decrease of L1 loss than just using turns produced by the speaker (Figure 3). Secondly, using our model has better performances in predicting the mean energy of the upcoming turn compared to the use of the linear regression *(Speaker estimate* = 0.318 ± 0.004*, Interlocutor Estimate* = -0.149 ± 0.003*)*. As expected conversants follow the trend of the more recent turns history (see Table 3).

4.3. Pitch Range

Feature Selection Similarly to the experiment about energy we compare the different subsets of features (See 3.3). Here again, results show that the richer representation improves significantly the performances in predicting the F0 range of the upcoming turn compared to the use of only F0 range of the previous turns (Table 2) but it is not significantly better than the subset formed by just F0 range (the target), DA and Δ. We used a K-fold approach, with K = 10, and use a dependent t-test to compare the subsets of features, per each turn and per each setup (*S* and *S+I*).

Speaker vs Speaker + Interlocutor As for energy, using of a representation vector formed by the selected features of both speaker and interlocutor leads to a significant decrease ($p < 0.05$) of L1 loss than just using turns produced by the speaker (see Figure 4 and Table 3). The use of our model improves performances in predicting the F0 range of the upcoming turn compared to the use of the linear regression. The L1 loss is the same in the case of the baseline for both *S* and *S+I* as the estimate of the linear regression for interlocutor is close to zero (estimate = -0.024 ± 0.003 while the one of the speaker is 0.469 ± 0.004). Adding more turns to the previous history causes the error decrease both in our model and in the baseline.

Speech Rate For SR we don't observe any significant difference between the *S* and *S+I* setups (Figure 5 and Table 3).

Table 3: *Results for mean E, F0 range and SR in the case of our model (upper) and the Baseline (lower) for N = 1, 5, 10. ** means that p-value* $< 10^{-8}$ *while * that p-value* < 0.05.

	E		F0		SR	
N	S + I	S	S + I	S	S + I	S
1	$\mathbf{0.381 \pm 0.006}$	$\mathbf{0.416 \pm 0.006}$**	$\mathbf{0.371 \pm 0.011}$*	$\mathbf{0.390 \pm 0.011}$*	0.334 ± 0.005	0.337 ± 0.005
5	$\mathbf{0.357 \pm 0.005}$	$\mathbf{0.380 \pm 0.005}$**	$\mathbf{0.313 \pm 0.009}$*	$\mathbf{0.324 \pm 0.009}$*	0.321 ± 0.004	0.322 ± 0.005
10	$\mathbf{0.353 \pm 0.005}$	$\mathbf{0.377 \pm 0.005}$**	$\mathbf{0.299 \pm 0.009}$*	$\mathbf{0.311 \pm 0.009}$*	0.315 ± 0.004	0.318 ± 0.004

Baseline, Linear Regr.						
	E		F0		SR	
N	S + I	S	S + I	S	S + I	S
1	0.416 ± 0.006	0.422 ± 0.006	0.401 ± 0.009	0.401 ± 0.006	0.341 ± 0.005	0.341 ± 0.005
5	0.405 ± 0.005	0.406 ± 0.005	0.344 ± 0.009	0.344 ± 0.009	0.331 ± 0.005	0.331 ± 0.005
10	0.413 ± 0.006	0.413 ± 0.006	0.335 ± 0.010	0.335 ± 0.010	0.327 ± 0.005	0.327 ± 0.004

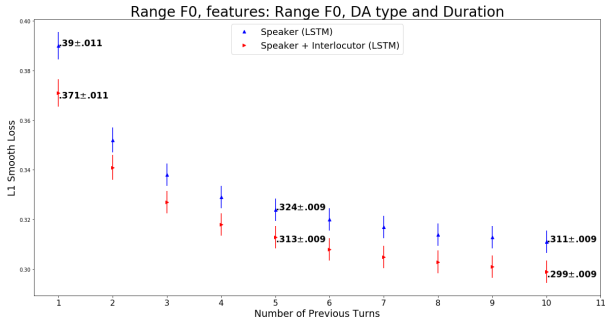

Figure 4: *L1 smooth loss while predicting F0 range of upcoming turn for the S and S+I setups by the use of LSTM model with the richer representation (F0 range, DA and* Δ*).*

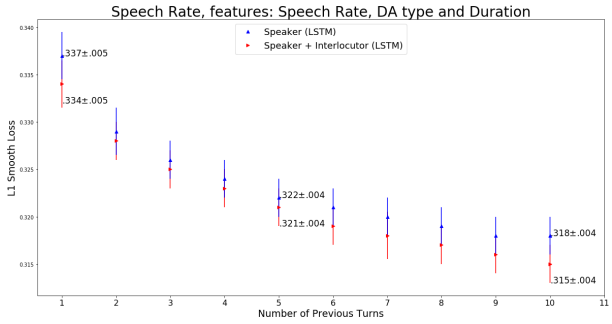

Figure 5: *L1 smooth loss while predicting SR of upcoming turn for the S and S+I setups by the use of LSTM model with the richer representation (SR, DA type and* Δ*).*

The use of our model decreases the error compared to the baseline *(speaker estimate* = 0.142 ± 0.002*, interlocutor estimate* = -0.013 ± 0.003*)* in both cases. A possible explanation is that the method we applied to compute SR is sensitive to the length of the words and this effect is amplified for short turns.

5. Discussion and Conclusions

In this paper we presented a new method to evaluate convergence, consisting in predicting some acoustic parameters of the upcoming turn in a conversation. The idea is to evaluate if the speech style of a speaker is influenced by the speech style of his interlocutor. We introduced a first simple model to predict the value of energy, F0 range and speech rate in the upcoming turn. We found that rich representation of the turn history for both *S* and *S+I* leads to reduce the error in prediction compared to the use of the linear baseline for all the scrutinized variables. Moreover, for energy and F0 range the use of turns history of speaker and interlocutor reduces the error in prediction compared to the case of use turns produced by the speaker only. For E the decrease is highly significant (p-value $< 10^{-8}$) while significant for F0 range (p-value < 0.05) for all the N turns we explored. This result is in agreement with past studies that claim energy to be a variable that exhibits strong convergence effects (due to a kind of automatic changed of energy as stated by [31]) while F0 shows a weaker of interlocutor influence. Even thought past literature assessed convergence for SR, in this study we do not have evidence that speaker SR is influenced by his interlocutor. The reason could be that the measure of SR we adopted is sensitive to words length and the computation is very noisy in case of short turns. We plan to use another measure of SR in the future.

The subset of features that we selected per each target variable turned out to be the same. It contains the target, the Δ and DA. This confirms that an important variable that controls the speech production and the reciprocal influence of speaker and interlocutor is the structure of the conversation and the type of DA, as [26] explained the lexical information are important to determine the type of DA. As part of these information is contained in DA, we plan to add lexical information as input to explore the influence that they could have on the prediction task.

In addition we observe that the loss decreases as the number of turns of previous history increases for both F0 and SR. This is in agreement with past literature ([20, 32]) that states speakers mainly tend to converge to their baseline (average value that they have in other conversations). On the other hand the trend of E is different as it depends essentially by the most recent turns than the antecedent history.

Even though our goal is not to build a system that adapts to human interlocutor such results should be of interest either as justification or inspiration for anyone interested in building artificial systems able to adapt to the user speech characteristics. Our future goal is to generalize the approach to other variables as well as to test the approach on bigger data sets (like the Fisher corpus [33]). It allows to refine the tagger of DA (adding more categories and improving the F1 score of the classification) and improve our model, adding an attention layer.

6. References

[1] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." *Journal of Personality and Social Psychology*, vol. 32, no. 5, p. 790, 1975.

[2] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech 2011*, 2011.

[3] A. Gravano, Š. Beňuš, R. Levitan, and J. Hirschberg, "Three tobi-based measures of prosodic entrainment and their correlations with speaker engagement," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 578–583.

[4] H. Giles and P. Powesland, "Accommodation theory," in *Sociolinguistics*. Springer, 1997, pp. 232–239.

[5] R. L. Street, "Speech convergence and speech evaluation in fact-finding iinterviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[6] K. Bergmann, H. P. Branigan, and S. Kopp, "Exploring the alignment space–lexical and gestural alignment with real and virtual humans," *Frontiers in ICT*, vol. 2, p. 7, 2015.

[7] J. Brandstetter, C. Beckner, E. B. Sandoval, and C. Bartneck, "Persistent lexical entrainment in hri," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 63–72.

[8] J. Lopes, M. Eskenazi, and I. Trancoso, "Automated two-way entrainment to improve spoken dialog system performance," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8372–8376.

[9] R. Levitan, S. Beňuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar." in *INTERSPEECH*, vol. 16, 2016, pp. 1166–1170.

[10] E. Székely, M. T. Keane, and J. Carson-Berndsen, "The effect of soft, modal and loud voice levels on entrainment in noisy conditions," in *Interspeech*, 2015.

[11] C. Sanker, "Comparison of phonetic convergence in multiple measures," Cornell Working Papers in Phonetics and Phonology, Tech. Rep., 2015.

[12] S. Kousidis, D. Dorran, C. Mcdonnell, and E. Coyle, "Convergence in human dialogues time series analysis of acoustic feature," 2009.

[13] Q. Ma, Z. Xia12, and T. Wang, "Absolute and relative entrainment in mandarin conversations."

[14] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 113–117.

[15] A. Weise and R. Levitan, "Looking for structure in lexical and acoustic-prosodic entrainment behaviors," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, vol. 2, 2018, pp. 297–302.

[16] K. P. Truong and D. Heylen, "Measuring prosodic alignment in cooperative task-based conversations," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[17] J. M. Pérez, R. H. Gálvez, and A. Gravano, "Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement." in *INTERSPEECH*, 2016, pp. 1270–1274.

[18] C. De Looze, C. Oertel, S. Rauzy, and N. Campbell, "Measuring dynamics of mimicry by means of prosodic cues in conversational speech," 2011.

[19] B. Vaughan, "Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] U. Cohen Priva, L. Edelist, and E. Gleason, "Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 2989–2996, 2017.

[21] U. Cohen Priva and C. Sanker, "Convergence is predicted by particular interlocutors, not speakers," 2019, (submitted).

[22] A. Schweitzer and M. Walsh, "Exemplar dynamics in phonetic convergence of speech rate." in *INTERSPEECH*, 2016, pp. 2100–2104.

[23] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech." in *INTERSPEECH*, 2013, pp. 525–529.

[24] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[25] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[26] A. Stolcke, E. Shriberg, R. Bates, N. Coccaro, D. Jurafsky, R. Martin, M. Meteer, K. Ries, P. Taylor, C. Van Ess-Dykema *et al.*, "Dialog act modeling for conversational speech," in *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, 1998, pp. 98–105.

[27] F. Eyben and B. Schuller, "opensmile:): the munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[28] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[29] J. Auguste, R. Perrotin, and A. Nasr, "Annotation en actes de dialogue pour les conversations d'assistance en ligne," in *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, 2018, p. 577.

[30] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[31] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[32] U. Cohen Priva and C. Sanker, "Distinct behaviors in convergence across measures," in *Proceedings of the 40th annual conference of the cognitive science society. Austin, TX*, 2018.

[33] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.