On the Predictability of the Prosody of Dialog Markers from the Prosody of the Local Context

Anindita Nath, Nigel G. Ward

University of Texas at El Paso

anath@miners.utep.edu, nigelward@acm.org

Abstract

Dialog markers, such as *yeah* and *okay* generally seem to fit smoothly in the flow of dialog, with prosody that is natural and appropriate for the local context. We here examine this effect, specifically looking at the predictability of the prosody of dialog markers from the prosody of the local context. Using 72 prosodic features representing the local context, we built simple models able to predict the average pitch, log energy, cepstral flux, and harmonic ratio for the 12 most common dialog markers of American English. The model's predictions accounted for over a third of the variance in the observed prosody, showing a modest but meaningful context dependence.

Index Terms: prosody, computational modelling, discourse markers, fillers, backchannels, pitch, energy

1. Introduction

Dialogs often seem to have some degree of momentum, in the sense that the properties of a speaker's next utterance can be partly determined by what is appropriate for the local context, regardless of any specific communicative intention. To the extent that such momentum exists, we should study dialog as a thing in itself, not just as an accidental product of independent individual choices. Accordingly, exploration of the nature and strength of such momentum can inform how to best build models of dialog phenomena.

Our work also has a practical motivation. Dialog systems often exhibit awkward prosody. One cause is intrinsic unnaturalness, but this is being alleviated by recent advances leading to models that can produce prosodically-natural utterances [1, 2, 3], at least when judged in isolation. A second cause is prosody inappropriate for the intended meaning or function, and this topic has also been addressed by much work, leading to a good understanding of many prosody-meaning mappings. A third cause is prosody that is simply inappropriate for the context. This has received much less attention, but remains a major challenge for speech synthesis [4].

In this research we examine discourse markers occurring in spoken dialog, or "dialog markers" for short, and in particular, the predictability of their prosody from the local context. These are convenient for an initial exploration for several reasons: They are very common. They serve many important functions, including managing turn-taking, marking topic structure, and expressing stance [5, 6]. They typically are semantically semi-independent, standing outside the propositional content, or in other words, have a core procedural and not conceptual meaning [6]. And finally, their prosody is usually their own, being less often affected by the larger prosodic patterns that govern many word sequences.

Our hypothesis is that local prosodic context is informative for predicting the prosodic form of dialog makers.

	token count
huh	1417
now	5910
oh	14053
okay	3915
really	11009
right	12448
uh	52230
uh-huh	12155
um	16392
well	16701
yeah	33768
yes	3393

Table 1: *Dialog markers considered*

2. Related Work

Various aspects of the prosody of dialog markers have received significant attention. Much work has described the prosodic correlates of various different uses, for example, discourse *vs* sentential uses of *now* [7], direction *vs* acknowledgment uses of *right* [8], questioning *vs* reacting uses of *really* [9], different polarities and intensities of *yeah* [10], backchannel, topic shift, agreement marker and other uses of *okay* [11, 12, 13], and so on [14, 15, 16, 17, 18, 19, 20]. An extensive study on prosodically marked and unmarked *okay*s [21] revealed that *okay*s are more prosodically marked — with more extreme pitch, loudness, duration, timing, and overall vocal quality — when they used in the display of various orientations (such as disagreeing, displaying aggravation, treating others' actions as odd or bizarre, exuding happiness, and excitement) than when they are used simply to convey acknowledgment, acceptance, or assessment of the other speakers' actions. Other work has noted general prosody-function mappings present across many dialog markers [22]. Some prosody-pragmatic relationships have been shown to be present across many languages [23]. The study of how prosodic context directly affects dialog marker prosody has been very limited; we know of only two small-corpus studies leading to a set of handcrafted rules [24, 25], and investigations of the extent of entrainment [26, 27].

Considering more generally work aimed at implementing simple, direct responsiveness, based on local prosodic context, a large body of work on aligning to the prosody of the interlocutor's previous utterance [28], another large body of work showing success for turn-taking predictions [29], and work on choosing the emotional coloring of responses [30] and the form of backchannels [31, 32] and other words and turns [33, 34]. This paper extends this line of inquiry to explore the prospects for prosodic tailoring of dialog makers.

mean over the token	max over the token
log energy	log energy
cepstral flux	cepstral flux
pitch	pitch
harmonic ratio	harmonic ratio

Table 2: *Predicted Features*

	Past Windows,
	from the
	start (s) of the token
		Future Windows,
		from the
		end (e) of the token

log energy lengthening peak disalignment	s-3200 to s-800, and s-800 to s	e to e+800, and e+800 to e+3200
creakiness pitch lowness pitch highness narrow pitch wide pitch speaking rate	s-1600 to s-200, and s-200 to s	e to e+200, and e+200 to e+1600

Table 3: *Features used for prediction (context features). Times are in milliseconds relative to start (s) and end (e) of the dialog marker whose prosody is being predicted*

3. Data

We used the Switchboard corpus of American English telephone conversations [35]. After excluding recordings with poor audio quality or artifacts that bothered our pitch tracker, we considered 1900+ conversations involving 400+ speakers. We considered all audio spans bearing labels from the list in Table 1, according to the Picone transcriptions [36]. We did not use functional labels [37] or do any additional checks, so cases where the word was not actually being used as dialog marker were not excluded.

4. Prosodic Features Predicted

Ultimately, we would like to predict every detail of the prosody of each dialog marker token: the value for every feature at every frame of the token. However, for this study we predict only four features namely: i) loudness, as measured by its acoustic correlate *log energy*, ii) pitch, as measured by its acoustic correlate *fundamental frequency* or *f0*, estimated by the pitch tracker *fxrapt* [38] in the *VoiceBox* toolkit for MATLAB, iii) *cepstral flux*, as a measure of lengthening and reduction, and iv) the *harmonic ratio* [39] which is a proxy for harmonicity and, indirectly, other properties of voicing, including creakiness, breathiness, and devoicing, The relevance of pitch, energy, and timing is well known; we also included harmonicity since it appears to help differentiate among roles for some dialog markers [13]. For each of these four features, we did two experiments, one to predict the average over the entire token, and another to predict the maximum value, as both contribute to what is perceived. Table 2 summarizes. For pitch, frames with undefined values were excluded from the feature computations.

5. Context Features

As our aim is to explore, we sought neither a maximal set of features nor a minimal one. Rather we chose a set of 72 features that were diverse, convenient, reliable, and broadly covered the local context. We used contextual features for both speakers: the one who produced the dialog marker and the interlocutor. Together these features cover the time from 3.2 to 0 seconds before the start of the dialog marker and the time from 0 to 3.2 seconds after its end. We chose to consider also future information because we observed [13] that often the observed prosody of a dialog marker is suitable not only for what came before, but also for what is upcoming, either by the same speaker, or by the interlocutor, because the prosody of a dialog marker can guide or otherwise relate to the interlocutor's future behavior. However we also did experiments using only the past context, since that is more relevant for most use cases.

Specifically, for each speaker, we computed the 36 features shown in Table 3: 9 base features, each computed over 4 time spans. All were computed using the Midlevel Prosodic Features Toolkit [40]. The four pitch configuration features are used to enable everywhere-meaningful computation of pitch information, even over windows with few pitch points [41]. The "peak disalignment" feature is a measure of the displacement between energy peaks and pitch peaks [42]; for this data, this generally measures late peak (delayed peak) occurring in stressed syllables. The specific time windows were chosen based on some initial intuitions about the rate of local prosodic change relative to broader movements, and were not subsequently optimized or revisited. Together these features capture much about the local prosody and the local turn-taking state. Both the token features and the context features were z-normalized per track, to reduce the effects of intrinsic speaker differences.

6. Prediction Model

Our goal being insight rather than optimization, we used a very simple model: multivariate linear regression. This allowed us to trivially examine how the context features were affecting the dialog markers' prosody. We developed separate models for each dialog marker type, as we did not expect the same rules to work well for all, for example, for both *huh* and *okay*, though we did also experiment with an overall model.

7. Experiment Design and Results

We followed an intra-corpus evaluation approach. Each model, one for each of the 12 dialog markers, was evaluated with a disjoint train-test split of 70:30, chosen such that the test set contained no dialogs seen in the training set. The root mean squared error (RMSE) was used to evaluate the performance of each model. The utility of local context information was measured by the percent reduction in RMSE values for model predictions compared to the baseline of simply predicting the average over all instances of that type, for example, predicting the global average *yeah*.

Table 4 shows the quality for the baseline predictions and the model's predictions. The errors are lower with the model, with reductions ranging from 22% to 37%, showing that the local context is informative. The benefit is statistically significant, for all 4 predicted features in each case (matched pairs t-tests, $p < 0.001$).

This contextual dependency of prosody is demonstrated

	predicting mean features				predicting maximum features			
	le	cf	p	hr	le	cf	p	hr
Baseline RMSE	0.61	0.82	0.67	0.66	0.74	1.38	1.95	1.18
Model RMSE	0.44	0.64	0.52	0.46	0.49	0.92	1.34	0.74
Reduction, %	28.7	22.4	23.6	29.6	33.4	31.9	31.1	37.2

Table 4: *Prediction errors with the baseline and with the the model (using Linear Regression), and percent reduction for predicting mean (respectively maximum) features. Errors are the unweighted average of the RMSE values for each of the 12 dialog marker types. le is log energy, cf is cepstral flux, p is pitch, and hr is harmonic ratio.*

	predicting mean features				predicting maximum features			
	le	cf	p	hr	le	cf	p	hr
Baseline RMSE	0.61	0.82	0.67	0.66	0.74	1.38	1.95	1.18
Model RMSE	0.58	0.80	0.60	0.61	0.64	1.34	1.91	1.15
Reduction, %	6.0	2.9	10.2	7.7	12.6	2.6	2.5	2.8

Table 5: *Results for predictions using only past context.*

even more prominently if we compute the variance of prediction errors. It is seen that the overall average variance of the predicted prosody, 0.27 for mean features and 0.55 for max features were reduced by around 44% and 56%, respectively from their corresponding baseline averages.

We incidentally note that the reductions were greater for the maximum features than for the mean features, although this difference may be be largely due to outliers.

If momentum partly determines speaker behavior, then the prosody should be largely predictable from features representing the past context only. A model using only such features gave some benefit, as seen in Table 5, with RMSE reductions of 3% to 13%, but much less than those obtained when using also future context. This was a surprise to us, given that we had earlier observed many dependencies on past context [13], albeit for task-oriented dialogs. Perhaps the functions of dialog markers are relatively more forward-looking in casual conversation than in task-oriented dialogs.

To see whether individual models for each marker type were really necessary, we trained a general model, using linear regression, on the data from all 12 dialog marker types together. This generic model generally did not perform as well: for mean value prediction it gave less error reduction for log energy, 24.3%, cepstral flux, 19.2%, and harmonic ratio, 24.6% (c.f. Table 4), although for predicting mean pitch the generic model performed better, giving a 27.4% error reduction. This indicates that effects of context on dialog marker prosody are somewhat type-dependent, but not enormously so.

8. Analysis of the Models

This section illustrates the regularities that our models learned, and discusses the strengths and limitations of prediction from context alone.

For most dialog markers, the most predictive features were the pitch disalignment features. These features often had correlations of 0.20 or higher with high pitch, high volume, and high harmonicity. This is likely because peak disalignment often marks times of shared laughter, questions, and other high-engagement dialog acts [42], and these generally call for enthusiastic dialog markers. There was also a tendency to matching: more specifically, when the immediate past context exhibits higher volume or pitch, the prosody of the dialog marker often does too, for example when acknowledging new information.

Some specific dialog markers had additional unique tendencies. For example, for the word *now*, high pitch correlated with high pitch by the same speaker over the next few sections, likely due to its forward-looking role, as in introducing new subtopics. While most of the strong correlations were with contextual behavior by the person who produced the dialog marker, there were also interlocutor effects. For example, the word *okay* tended to be lower in pitch when in the context the interlocutor's cepstral flux was low, likely due to the use of lengthening and reduction marking a low density of new information and/or seeking only weak feedback.

For insight on why the model sometimes performed well and sometimes poorly, we start by considering Table 6. We note relatively high predictability for *uh*, and *huh*, likely because they usually have no independent prosody or meaning beyond their roles in the local context. We see low predictability for mean features of *really* and for both mean and max features of *right*, which are sometimes dialog markers, but sometimes just adverbs and adjectives, in which roles they likely have different prosodic tendencies. The prosody of *okay* was also hard to predict, perhaps because it often is deployed to convey a specific meaning or function, rather than just fitting passively in the context.

To further understand where our model succeeded and failed, we examined its performance on specific tokens: for each dialog marker type, the 5 for which the predictions were least accurate, and the 5 for which they were most accurate. This was done subjectively, relying on our perceptions and qualitative inductive methods.

Factors that were common when the model failed included: i) background noise in the audio segment. (Our feature computations were not robust to noise.) ii) long monologues (a dialog activity type uncommon in Switchboard, and likely rare in the training data). iii) one speaker with an unusual accent or perhaps a speech impediment, iv) incorrect annotations, for example, where the label was *um*, but the sound was more like *hmm*. (Our model for *um*, of course, had not been trained to predict the prosody of *hmm* tokens.) v) sequences of dialog markers,

	predicting mean features					predicting maximum features				
	le	cf	p	hr	Avg.	le	cf	p	hr	Avg.
huh	29	43	26	34	33	22	46	27	15	28
now	19	21	30	25	24	23	21	44	39	32
oh	17	27	15	24	21	35	37	25	37	34
okay	18	30	7	6	15	19	31	18	40	27
really	46	16	-1	15	19	37	39	35	48	40
right	37	2	7	25	18	31	14	19	39	26
uh	43	21	56	30	38	52	48	33	32	41
uh-huh	12	20	34	50	29	18	11	31	39	25
um	31	22	29	32	28	56	50	36	27	42
well	25	24	23	33	26	32	24	40	33	32
yeah	38	25	20	44	32	49	38	40	54	45
yes	30	19	37	37	31	28	22	26	45	30
Average	29	22	24	30	26	33	32	31	37	33

Table 6: *Results per Dialog Marker: Percent reduction in root mean squared error for predicting mean (respectively maximum) feature values using linear regression.*

such as *well, yeah* and *oh, okay*. (The prosody of markers in sequence is apparently different from those in isolation, the more common case in the training data.) vi) *okay* at the end of conversation, where it was short and breathy as part of the closing, and vii) *huh* when produced as a repair question or strong exclamation.

Cases where the model's predictions were most accurate included i) typical backchannel uses of *yeah*, ii) times the speaker and interlocutor shared happy or excited agreement, for example, *You're pretty Texan, yes . . . [interlocutor laughter]*, and iii) sympathetic productions of *really* in the context of talk about troubles or problems, as in *that can really be a problem*.

Overall, it seems that the model tends to perform well when the local dialog context is one that is common in the Switchboard genre.

9. Discussion and Future Work

We have found evidence that dialog markers' prosody can indeed be predicted directly from the prosody of the context, to some extent. This is true even with this very limited feature set and very simple model.

However the strength of "momentum" or flow as a factor in the choice of dialog marker prosody is fairly modest; instead their prosody depends more on what the speaker intends to say next.

Better performance should be obtainable using better models and more context features, including not only more prosodic features, but ideally also lexical information. Future work should also attempt more detailed predictions: not just of a token's averages, but also of contour parameters or even frame-by-frame values.

An important open question is the extent to which the prosody adjustments recommended by a context-sensitive model have actual value in dialog. Previous research suggests that improved responsiveness can increase perceived naturalness and responsiveness, and ultimately rapport, engagement, and user satisfaction [30, 43, 44, 45, 46, 27]. Prosodic entrainment has even been shown to lead to greater success in student learning in an intelligent dialog tutoring system [47]. Human subjects experiments are needed to establish whether such manipulations also have value for dialog markers.

This will pave the way to more responsive dialog systems, for example spoken language chatbots. Optimal exploitation of context-based prosody predictions will, however, likely require advances in speech synthesis, to support generation of tokens that exhibit fully appropriate prosody. Moreover, while simple context-based control of dialog marker prosody may be adequate for chatbots, where the aim is to keep the dialog flowing, use in task-oriented systems will bring further challenges. We would likely need an additional module trained to judge the similarity of the local dialog context to the context in the training data. We would also need methods to combine these context-based predictions with other factors that affect the prosody, such as the current dialog state and the communicative intent of the system [48].

More generally, future work should explore the possibility of predicting other aspects such as the prosody of full utterances in dialog based on local context. Such models could help not only improve dialog systems, but also provide knowledge that could be used very widely, for example to to help autistic people and language learners master the typical patterns of responsiveness in dialog, thereby helping improve their communication skills.

10. References

[1] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in *InterSpeech*, 2019.

[2] S. Shechtman and A. Sorin, "Sequence to sequence neural speech synthesis with prosody modification capabilities," in *10th ISCA Speech Synthesis Workshop*, 2019.

[3] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[4] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation: state-of-the-art assessment and suggestion for a novel research program," in *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, 2019.

[5] M. Louwerse and H. Mitchell, "Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account," *Discourse Processes*, vol. 35, pp. 199–239, 2003.

[6] B. Fraser, "What are discourse markers?" *Journal of Pragmatics*, vol. 31, no. 7, pp. 931–952, 1999.

[7] J. Hirschberg and D. Litman, "Empirical studies on the disambiguation of cue phrases," *Comput. Linguist.*, vol. 19, no. 3, pp. 501–530, 1993.

[8] K. Ward and D. Novick, "Prosodic cues to word usage," in *ICASSP*, vol. 1, 1995, pp. 620–623.

[9] C. Lai, "Perceiving surprise on cue words: Prosody and semantics interact on right and really," in *Interspeech*, 2009, pp. 1963–1966.

[10] V. Freeman, G.-A. Levow, R. Wright, and M. Ostendorf, "Investigating the role of 'yeah' in stance-dense conversation," in *Interspeech*, 2015.

[11] A. Gravano, S. Benus, H. Chávez, J. Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of 'okay'," in *ACL*, 2007, pp. 800–807.

[12] M. Van Zyl and J. J. Hanekom, "When "okay" is not okay: Acoustic characteristics of single-word prosody conveying reluctance," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. EL13–EL19, 2013.

[13] A. Nath, "Towards naturally responsive spoken dialog systems by modelling pragmatic-prosody correlations of discourse markers," in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 2020, p. 128–129.

[14] R. Gardner, *When Listeners Talk: Response tokens and listener stance*. John Benjamins, 2001.

[15] K. Truong and D. Heylen, "Disambiguating the functions of conversational sounds with prosody: The case of 'yeah'," in *Interspeech*, 2010, pp. 2554–2557.

[16] L. Lee, D. Jouvet, K. Bartkova, Y. Keromnes, and M. Dargnat, "Correlation between prosody and pragmatics: case study of discourse markers in French and English," in *Interspeech 2020*, 2020.

[17] G. Skantze, A. Hjalmarsson, and C. Oertel, "Turn-taking, feedback and joint attention in situated human–robot interaction," *Speech Communication*, vol. 65, pp. 50–66, 2014.

[18] C. Oertel, J. Gustafson, and A. W. Black, "On data driven parametric backchannel synthesis for expressing attentiveness in conversational agents," in *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, 2016, pp. 43–47.

[19] D. Neiberg, G. Salvi, and J. Gustafson, "Semi-supervised methods for exploring the acoustics of simple productive feedback," *Speech Communication*, vol. 55, no. 3, pp. 451–469, 2013.

[20] C. Rühlemann and S. T. Gries, "How do speakers and hearers disambiguate multi-functional words? the case of well," *Functions of Language*, vol. 28, pp. 55–80, 2020.

[21] W. Beach, "Using prosodically marked "okays" to display epistemic stances and incongruous actions," *Journal of Pragmatics*, vol. 169, pp. 151–164, 2020.

[22] N. Ward, "Pragmatic functions of prosodic features in non-lexical utterances," *Speech Prosody*, vol. 4, 11 2003.

[23] M. Heldner, M. Włodarczak, Štefan Beňuš, and A. Gravano, "Voice Quality as a Turn-Taking Cue," in *Interspeech*, 2019, pp. 4165–4169.

[24] N. Ward and W. Tsukahara, "A study in responsiveness in spoken dialog," *International Journal of Human-Computer Studies*, vol. 59, pp. 603–630, 2003.

[25] N. G. Ward and R. Escalante-Ruiz, "Using subtle prosodic variation to acknowledge the user's current state," in *Interspeech*, vol. 9, 2009.

[26] M. Heldner, J. Edlund, and J. B. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Interspeech*, 2010.

[27] N. Sadoughi, A. Pereira, R. Jain, I. Leite, and J. F. Lehman, "Creating prosodic synchrony for a robot co-player in a speech-controlled game for children," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2017, p. 91–99.

[28] R. Levitan, "Developing an integrated model of speech entrainment," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[29] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, p. 101178, 2020.

[30] J. C. Acosta and N. G. Ward, "Achieving rapport with turn–by–turn, user–responsive emotional coloring," *Speech Communication*, vol. 53, no. 9, pp. 1137–1148, 2011.

[31] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. G. Ward, "Prediction and generation of backchannel form for attentive listening systems." in *Interspeech*, 2016, pp. 2890–2894.

[32] C. Oertel, P. Jonell, D. Kontogiorgos, J. Mendelson, J. Beskow, and J. Gustafson, "Crowd-Sourced Design of Artificial Attentive Listeners," in *Interspeech*, 2017, pp. 854–858.

[33] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "Neural spoken-response generation using prosodic and linguistic context for conversational systems," *Proc. Interspeech*, pp. 246–250, 2021.

[34] S. Fuscone, B. Favre, and L. Prevot, "Neural representations of dialogical history for improving upcoming turn acoustic parameters prediction," in *Interspeech 2020*, 2020, pp. 4203–4207.

[35] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *ICASSP*, 1992, p. 517–520.

[36] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *ICSLP*, 1998, pp. 1543–1546.

[37] 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.

[38] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[39] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005, p. 304, iSBN: 978-0-470-09334-4.

[40] N. G. Ward, "Midlevel prosodic features toolkit (2016-2021)," 2021, https://github.com/nigelgward/midlevel.

[41] ——, *Prosodic Pattterns in English Conversation*. Cambridge University Press, 2019.

[42] ——, "A corpus-based exploration of the functions of disaligned pitch peaks in American English dialog," in *Speech Prosody*, 2018, pp. 349–353.

[43] N. Lubold and H. Pon-Barry, "Acoustic-prosodic entrainment and rapport in collaborative learning dialogues," in *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014, pp. 5–12.

[44] Y. Li, C. T. Ishi, K. Inoue, S. Nakamura, and T. Kawahara, "Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction," *Advanced Robotics*, vol. 33, no. 20, pp. 1030–1041, 2019.

[45] R. H. Gálvez, A. Gravano, Š. Beňuš, R. Levitan, M. Trnka, and J. Hirschberg, "An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars," *Speech Communication*, vol. 124, pp. 46–67, 2020.

[46] J. I. Choi and E. Agichtein, "Quantifying the effects of prosody modulation on user engagement and satisfaction in conversational systems," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 2020, pp. 417–421.

[47] J. Thomason, H. V. Nguyen, and D. J. Litman, "Prosodic entrainment and tutoring dialogue success," in *AIED*, 2013.

[48] N. G. Ward and D. DeVault, "Challenges in building highly interactive dialog systems," *AI Magazine*, vol. 37, no. 4, pp. 7–18, 2016.