Towards Continuous Estimation of Dissatisfaction in Spoken Dialog

Nigel G. Ward **Jonathan E. Avila** **Aaron M. Alarcon**
Computer Science
University of Texas at El Paso
nigelward@acm.org, jonathan.edav@gmail.com, aaronalarcon2368@gmail.com

Abstract

We collected a corpus of human-human task-oriented dialogs rich in dissatisfaction and built a model that used prosodic features to predict when the user was likely dissatisfied. For utterances this attained a $F_{.25}$ score of 0.62, against a baseline of 0.39. Based on qualitative observations and failure analysis, we discuss likely ways to improve this result to make it have practical utility.

1 Motivation

Accurate models of dialog quality are needed for many purposes, including closed-loop improvement of dialog systems (Walker et al., 2000; Möller et al., 2008; Lykartsis et al., 2018; Ponnusamy et al., 2020; Roller et al., 2020; Lin et al., 2020; Deriu et al., 2021). Spoken dialog includes much information that can be used to predict quality judgments, and successful prediction has been shown for many genres, and in particular in call-center analytics (Ang et al., 2002; Zweig et al., 2006; Morrison et al., 2007; Kim, 2008; Vaudable and Devillers, 2012; Pandharipande and Kopparapu, 2013; Chowdhury et al., 2016; Luque et al., 2017; Egorow et al., 2017; Irastorza and Torres, 2018; Abhinav et al., 2019; Cabarrão et al., 2019; Li et al., 2019).

While most work on dialog quality has focused on the quality of entire interactions, finer-grained quality estimates are more useful for many purposes. Casual observation suggests that in conversation people are often not shy about indicating, moment by moment, how they feel about things, both in terms of making progress towards their goal and in terms of how happy they are with the contributions and behavior of their interlocutor. To date, however, predictive modeling of quality at the level of turns has been rarely attempted, and has focused mostly on interaction quality and conversational proficiency, and in only a few dialog genres, both for human-machine and human-human dialogs (Ultes and Minker, 2014; Ultes et al., 2017a; Lykartsis et al., 2018; Bodigutla et al., 2019; Stoyanchev et al., 2019; Spirina et al., 2016; Ramanarayanan et al., 2019; Ando et al., 2020; Katada et al., 2020). In this work we attempt turn-level quality estimation in human-human dialogs in a new genre: short calls to an unknown merchant to make an appointment or arrange a simple transaction.

This paper presents the first publicly available corpus of (mock) customer-service calls, describes observations on how dissatisfaction occurs in conversations gone wrong, discusses prosodic and turn-taking indications, presents a simple model giving modest performance on the tasks of detecting dissatisfaction moment by moment and at the utterance level, and discusses what more is needed.

2 Scenario and Data

Among the many possible contexts in which to study aspects dialog quality, we chose to examine what happens when a person is trying to get something done and expects that it can be easily accomplished, but finds that it is not possible. We would have liked to study real commercial dialogs, where customers or users often have a goal that the agent or system may be unable or unwilling to satisfy, but there appear to be no datasets in this genre available for study. We therefore did our own data collection, with the details chosen to align with the goals of our sponsor, Google.

In some markets, Google enables users to find merchants by voice search, leading to the presentation of phone numbers to call. This is especially useful for illiterate users. Unfortunately, the ecosystem includes bad actors, who purchase adwords to entice callers, but then do not offer the expected

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 13–20
July 29–31, 2021. ©2021 Association for Computational Linguistics

service, offer it at an excessive price, or otherwise disappoint or trick callers. Google would like better ways to flag such abusive merchants, ideally from automatic analysis of behavior in the call itself. Unlike most conversations addressed in call analytics, there is no large reference corpus of good behavior in the domain, these callers have no previous relationship with the business, and, conveniently for our purposes, many confounds and complexities are reduced (Möller and Ward, 2008) and the causes of any negative feelings will be largely dialog-internal.

We accordingly collected a new corpus of telephone calls. Each participant was given rough instructions, for example, in the customer role, to call to arrange to get a flat tire patched for no more than $10, and, for the merchant, to get the customer's information and set an appointment time. In half the cases the two sets of instructions were aligned, so that the merchant was able to satisfy the customer's need (although often only after an attempt to upsell, to make things more realistic). In the other half, the merchant's instructions included constraints that precluded satisfying the customer's need. Thus, for example, they might be instructed to only make an appointment if the customer agreed to the $60 tire care package or accepted an additional $40 rush fee. Thus these calls were designed to reflect the behavior of abusive merchants, and to accordingly elicit the behavior of unsuspecting callers as they came to realize that they were dealing with a bad actor.

Wanting a wide sampling of customer-side behavior, we recruited participants for that role through a crowdsourcing site. These participants were given two to four tasks to accomplish, with a number to call for each. The base rate was $5 and they were incentivized with a $1 bonus for each call where they successfully made arrangements with a merchant within budget, but were told that this would not always be possible. The merchant-role participants were six trained confederates. The calls were in English, with the confederates mostly native speakers of American English and the customer actors, it turned out, mostly non-native speakers from European countries, with Poland and Portugal overrepresented. In total we collected 191 calls.

Most of the calls were, in our judgment, quite realistic, with each side trying hard to achieve their assigned goals. Indeed, some callers were able to get our confederates to deviate from instructions and agree to provide the requested service at the requested price; conversely, the confederates were sometimes able to wear down callers into agreeing to a price that violated their instructions. Excluding the latter category and other special cases, we had 52 "doomed" (bad-actor) calls and 62 fully satisfactory calls.

Calls were recorded in stereo. They were typically 1 to 4 minutes in length. Full documentation is available (Avila et al., 2021), and the corpus itself is freely downloadable (Avila, 2021b).

3 Subjective Observations and Annotation of Dissatisfaction

Callers in the doomed-to-fail dialogs reacted diversely. Often they showed surprise at the first indication that the merchant was not going to behave according to expectation. Often they attempted repair, usually by restating their goals, generally more assertively than the first time. Often they expressed annoyance or other negative assessment, although always politely, never with raw emotion. Occasionally callers engaged in other behaviors, including negotiating, pleading, and even displaying anger. Across these specific behaviors, there was often an underlying feeling of growing dissatisfaction. Doomed conversations also generally lasted longer (Miramirkhani et al., 2017) and lacked the warm and appreciative/grateful closings that were common in the control dialogs.

While most call analytics systems rely on speech recognition (Ando et al., 2020), this makes sense mostly for high quality audio, for languages where good speech recognizers exist, and for focusing on how to improve agents' behavior; none of these are the case in our sponsor's scenario. In particular, the bad actors strive to be indistinguishable from good actors, so we chose to focus on acoustic-prosodic features of the caller.

There are two lines of work that we might have built on: first, work identifying the prosodic correlates of specific dialog acts, including some relevant here (Selting, 1996; Ogden, 2010), but the variety of behaviors across speakers and calls would make it difficult to leverage this work; and second work on the prosodic correlates of emotion, but the behaviors observed here were more social and linguistic than visceral or paralinguistic, so we again decided not to attempt to leverage such findings. Instead, we chose to approach the problem as one of modeling undifferentiated dissatisfaction. We hoped that this would be generally, if weakly, detectable, using the same features across all contexts. Although dissatisfaction was often subtle to the point that we were unsure exactly when it was present, prosodic models are often able to exploit indications below conscious awareness, and we hoped that would also be the case here. Focusing on general dissatisfaction also aligns with our broader goal of better automatic quality judgments.

We accordingly labeled each utterance with **d** for those with indications of dissatisfaction, defined broadly, to include disappointment, annoyance, sadness, disengagement and so on, **n** for non-dissatisfied or "neutral" utterances, and **?** for those that were inaudible or otherwise impossible to classify (Avila et al., 2021). Initially 18 dialogs were annotated, each by four people, and, for frames within utterance spans labeled by all four, the Fleiss Kappa was 0.57. The weak agreement, illustrated in the Appendix, seemed to be mostly due to varying preferences for classifying borderline utterances as **d** versus **?** or **n**, rather than substantive differences in perception. Accordingly the rest of the corpus was labeled by only one annotator, and the results below are reported for these annotations.

4 Experiment Set-Up

We set ourselves two tasks: 1) Utterance-level prediction: distinguishing dissatisfied utterances from neutral utterances, and 2) Frame-level prediction: distinguishing moments within dissatisfied utterances from moments within non-dissatisfied utterances. For both tasks, the input was only those frames (or utterances) which had been given a **d** or **n** utterance; silent regions and ambiguous regions were thus excluded.

For the utterance-level and frame-level models, there are many more negative samples, as there are fewer dissatisfied dialogs and even in those many utterances are not dissatisfied. There are many more neutral utterances, since not all utterances in the dissatisfied dialogs are dissatisfied. The number of n and d utterances in the training, dev, and test sets are 46 and 24, 52 and 23, and 256 and 82. The average labeled utterance being about 2 seconds long, for the test set the frame counts were 54543 neutral and 20893 disappointed.

As our primary goal is detecting dissatisfaction, the baseline is to always predict dissatisfaction, and high precision is our primary goal. However recall also has some importance, so we also report $F_{.25}$ results.

5 Initial Feature Set

Most research in this area uses utterance-aligned features, but we wanted to avoid the travails of defining or performing segmentation, so we simply computed prosodic features everywhere. Specifically, we compute features for timepoints sampled every 10 milliseconds (a 10 ms stride), using features that span about 3 seconds on either side of the point being classified. Much research on paralinguistic prosody assumes that affective states directly affect the prosody in stable ways for a second or more, and accordingly uses global averages or simple functionals, but work on the prosodic correlates of stance and dialog acts suggests that here we need the ability to represent temporal configurations of prosodic features (Ward, 2019; Ward and Jodoin, 2019). Accordingly, we used a feature set that includes time-offset features which together tile a local span. Specifically we based this on a feature inventory included in the Midlevel Prosodic Features Toolkit (Ward, 2021), mono.fss. This includes measures of intensity, of pitch height (high or low), of pitch range (narrow or wide), of speaking rate (using energy flux as a proxy), and of creakiness, as this set worked well for detecting various stances (Ward et al., 2018). To this we added features for the Cepstral Peak Prominence (Smoothed) (CPPS) across two windows, based on our observation that breathy voice was saliently present in many dissatisfied utterances. CPPS is an effective measure for breathiness in clinical applications (Heman-Ackah et al., 2003), although seldom yet used in studies of dialog.

6 Analysis

To understand how each feature was contributing, we looked at correlations and also histograms, since the relationships were seldom simply linear. Dissatisfied utterances tended to include more silent or very quiet frames, with neutral utterances richer in relatively loud frames.

A clearer picture emerges when we examine the coefficients in the model for the features at specific temporal offsets, as seen in Figure 1. (The actual values are available at the companion website: http://www.cs.utep.edu/nigel/disappointment/.) Low intensity features over about 3 seconds around

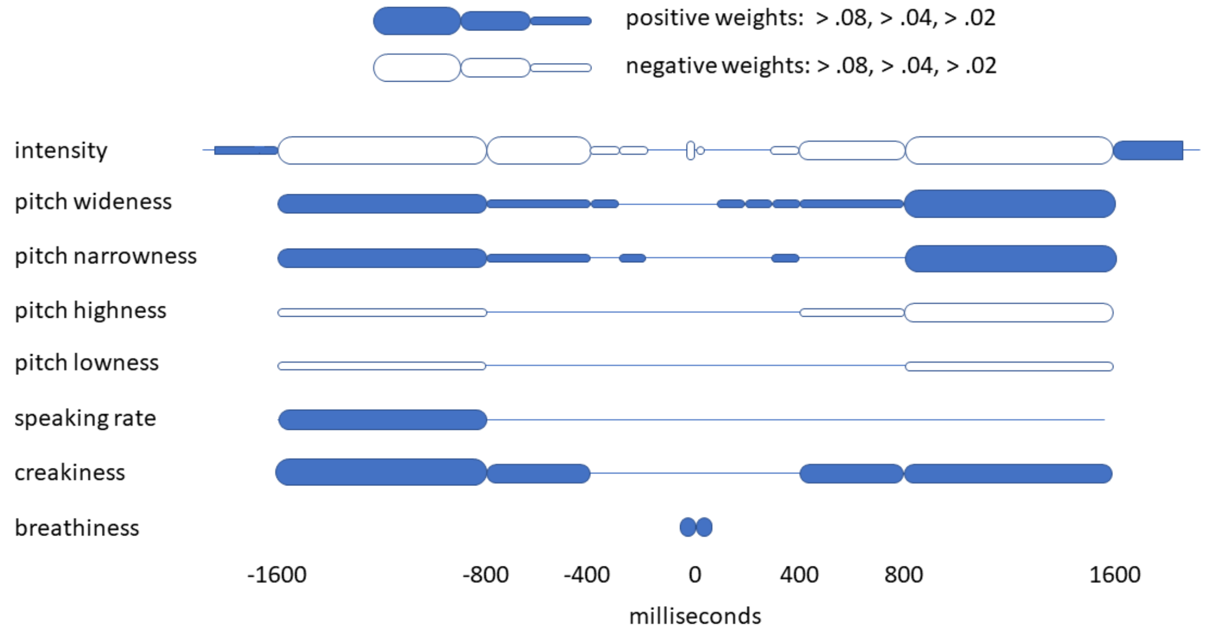

Figure 1: Features with relatively strong weights in the linear model for predicting the label dissatisfied per frame, where 0 ms is the start of the frame.

the frame being predicted had positive weights, with the more distant intensity features having negative weights; thus intensity that is low relative to the local context is the informative pattern. Both the wide pitch and narrow pitch features were indicative of disappointment, marking departures from a normal moderate pitch range. This fact aligns with the literature about the prosodic constructions used in complaining (Ogden, 2010; Ward, 2019). Creaky voice was also indicative of disappointment, which may relate to its reported role in marking disengagement (Ward, 2019). So did a couple hundred milliseconds of high CPPS, contrary to expectation. Low creakiness and high volume also correlated with a lack of dissatisfaction, which may reflect a general tendency for people when pleased to use clear and "pleasant" voices, with strong periodicity and harmonicity. In general the prosodic indications are not local to single syllables or words, but are present distributed across wider spans.

Seeking further understanding, we listened to a sampling of successes. Although our simplistic model could only learn one pattern, that pattern matched diverse ways of expressing dissatisfaction. This included a complaint, *I think this is still too much*, with narrow pitch on the first words and stress with high CPPS on the word *still*, and a quiet, annoyed *no thank you* (audio for these examples are at http://www.cs.utep.edu/nigel/disappointment/). Inversely, an example of a successful non-dissatisfied prediction was for a warm, fairly loud, slightly harmonic, moderately high-pitched, closing *thank you*.

We also listened to a sampling of failures. Misses included many frames from one dialog where excessive record gain had caused constant clipping, and some frames near a loud beep in the background. Our feature computations are not robust to such noise. We also examined false alarms. Many were in frames near regions of silence, such as at the start of an utterances or in the vicinity of a disfluent pause, even for pauses that, to our ears, did not seem perplexed or emphatic. Some false alarms occurred during the customer's explanation of their need, for example in the word *flat* in *my front left tire that is flat because of a nail*. While these did not express dissatisfaction with the merchant's behaviors, and so were not annotated as dissatisfaction, they certainly did express a negative assessment. While this could suggest tweaking the annotation guidelines, the more important lesson is that accurately predicting dissatisfaction requires modeling the stage of the dialog, not just the local context.

This analysis suggested that our model has explanatory value and validity, and thus may be likely to generalize well.

	precision	recall	$F_{.25}$
baseline	.43	1.00	.45
model	.57	.81	.58

Table 1: Frame-level Predictions of Dissatisfaction

	precision	recall	$F_{.25}$
baseline	.38	1.00	.39
model	.62	.73	.62

Table 2: Utterance-level Predictions of Dissatisfaction.

7 Revised Feature Set and Models

Based on the above analysis, we augmented the prosodic feature set with a time-into-dialog feature, for a total of 91 features. (We also did some small experiments with alternative feature sets based on OpenSmile's eGeMaps configuration (Eyben et al., 2016), but obtained no benefit.) We continued to use the simple linear regression model for our basic task, of predicting dissatisfaction at the frame-level. (Small experiments with logistic regression and k-nearest neighbors provided no benefit.) For utterance-level predictions we simply averaged the predictions for every frame within the utterance.

8 Results

Tables 1 and 2 show the performance of our frame-level and utterance-level models, on the test data. While the choice of threshold ultimately depends on the use scenario, here for each model we report performance at the value which maximizes $F_{.25}$.

For the frame-level detections, the performance was modest. As an indication of the scope for improvement, our model's agreement with the annotator, in terms of Cohen's Kappa, was .32, far below that of our secondary human annotators, whose agreements ranged from .57 to .71. Nevertheless, the frame-level model was good enough to support reasonable performance for the utterance-level discriminations.

9 Discussion and Future Work

Much previous work seems to assume that modeling dialog quality requires sophisticated methods to infer elusive hidden states. However here, thanks to a broad set of prosodic features and modeling in terms of temporal configurations, we obtain promising results without sophisticated modeling. This may open the way to a strong, incremental training signal useful for rapidly tuning spoken language chatbots and other dialog systems to better satisfy their users, after significant future work.

Future work should address the weaknesses noted above, perhaps in part by adding features to capture cross-participant behaviors (Gorisch et al., 2012) and timings. Better models are another priority topic. To consider the stage of the dialog and other factors, models that represent wider context should be tried (Ultes et al., 2017b). To support such advances, code for our existing, simple models is freely available (Avila, 2021a).

We also should try these methods on dialogs from different genres and exhibiting quality issues of other kinds. We also need to do ablation studies to better identify the sources of performance and to evaluate our model in comparison to others. Such comparisons have been rare in this research area, due to a lack of shared datasets, but our new corpus will enable other researchers to report directly comparable results.

Finally, since we see some level of performance across speakers with different native languages, we should investigate the possibility of universal, language-independent detection of dissatisfaction.

Acknowledgments

We thank Google for support, Marcin Wlodarczak for the the CPPS code, and Adrian Avendano for the annotation.

References

Kumar Abhinav, Alpana Dubey, Sakshi Jain, Veenu Arora, Asha Puttaveerana, and Susan Miller. 2019. Aqua: automatic quality analysis of conversational scripts in real-time. In *International Conference on Artificial Intelligence and Soft Computing*, pages 489–500.

Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, Yushi Aono, and Tomoki Toda. 2020. Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:715–728.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *ICSLP*.

Jonathan E. Avila. 2021a. Models for detecting dissatisfaction in spoken dialog. Https://github.com/joneavila/utep-dissatisfaction-models.

Jonathan E. Avila. 2021b. UTEP dissatisfaction corpus data. Https://github.com/joneavila/utep-dissatisfaction-corpus.

Jonathan E. Avila, Nigel G. Ward, and Aaron Alarcon. 2021. The UTEP corpus of dissatisfaction in spoken dialog. Technical Report UTEP-CS-21-23, University of Texas at El Paso.

Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. 2019. Domain-independent turn-level dialogue quality estimation via user satisfaction estimation. In *Implications of Deep Learning for Dialog Modeling, special session at Sigdial 2019*.

Vera Cabarrão, Mariana Julião, Rubén Solera-Ureña, Helena Moniz, Fernando Batista, Isabel Trancoso, and Ana Isabel Mata. 2019. Affective analysis of customer service calls. *ExLing 2019*, 25:37.

Shammur Absar Chowdhury, Evgeny A Stepanov, and Giuseppe Riccardi. 2016. Predicting user satisfaction from turn-taking in spoken conversations. In *Interspeech*, pages 2910–2914.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Olga Egorow, Ingo Siegert, and Andreas Wendemuth. 2017. Prediction of user satisfaction in naturalistic human-computer interaction. *Kognitive Systeme*, (1).

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:190–202.

Jan Gorisch, Bill Wells, and Guy J Brown. 2012. Pitch contour matching and interactional alignment across turns: An acoustic investigation. *Language and Speech*, 55(1):57–76.

Yolanda D. Heman-Ackah, Deirdre D. Michael, Margaret M. Baroody, Rosemary Ostrowski, James Hillenbrand, Reinhardt J. Heuer, Michelle Horman, and Robert T. Sataloff. 2003. Cepstral peak prominence: a more reliable measure of dysphonia. *Annals of Otology, Rhinology & Laryngology*, 112(4):324–333.

Jon Irastorza and M. Ines Torres. 2018. Tracking the expression of annoyance in call centers. In Ryszard Klempous, Jan Nikodem, and Peter Zoltan Baranyi, editors, *Cognitive Infocommunications, Theory and Applications*, pages 131–151. Springer.

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 315–323.

Woosung Kim. 2008. Using prosody for automatically monitoring human-computer call dialogues. *Proceedings of Speech Prosody 2008*, pages 79–82.

Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke. 2019. Acoustic and lexical sentiment analysis for customer service calls. In *IEEE ICASSP 2019*, pages 5876–5880.

Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765.

Jordi Luque, Carlos Segura, Ariadna Sánchez, Marti 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*, pages 2346–2350.

Athanasios Lykartsis, Margarita Kotti, Alexandros Papangelis, and Yannis Stylianou. 2018. Prediction of dialogue success with spectral and rhythm acoustic features using dnns and svms. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 838–845.

Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. 2017. Dial One for scam: A large-scale analysis of technical support scams. In *NDSS Symposium*, pages 1–15.

Sebastian Möller, Klaus-Peter Engelbrecht, and Robert Schleicher. 2008. Predicting the quality and usability of spoken dialog services. *Speech Communication*, 50:730–744.

Sebastian Möller and Nigel Ward. 2008. A framework for model-based evaluation of spoken dialog systems. In *Sigdial*.

Donn Morrison, Ruili Wang, and Liyanage C De Silva. 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98–112.

Richard Ogden. 2010. Prosodic constructions in making complaints. In Dagmar Barth-Weingarten, Elisabeth Reber, and Margret Selting, editors, *Prosody in Interaction*, pages 81–103. Benjamins.

Meghna Abhishek Pandharipande and Sunil Kumar Kopparapu. 2013. A language independent approach to identify problematic conversations in call centers. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 7(2):146–155.

Pragaash Ponnusamy, Alireza Roshan Ghias, Chenlei Guo, and Ruhi Sarikaya. 2020. Feedback-based self-learning in large-scale conversational AI agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13180–13187.

Vikram Ramanarayanan, Matthew Mulholland, and Yao Qian. 2019. Scoring interactional aspects of human-machine dialog for language learning and assessment using text features. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–109.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.

Margret Selting. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6:371–388.

Anastasiia Spirina, Maxim Sidorov, Roman B Sergienko, and Alexander Schmitt. 2016. First experiments on interaction quality modelling for human-human conversation. In *ICINCO, vol. 2*, pages 374–380.

Svetlana Stoyanchev, Soumi Maiti, and Srinivas Bangalore. 2019. Predicting interaction quality in customer service dialogs. In Maxine Eskenazi, Laurence Devillers, and Joseph Mariani, editors, *Advanced Social Interaction with Agents: Proceedings of the 8th International Workshop on Spoken Dialog Systems*, pages 149–159. Springer.

Stefan Ultes, Paweł Budzianowski, Inigo Casanueva, Nikola Mrkšic, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašic, and Steve Young. 2017a. Domain-independent user satisfaction reward estimation for dialogue policy learning. In *Proceedings of Interspeech*, pages 1721–1725.

Stefan Ultes and Wolfgang Minker. 2014. Interaction quality estimation in spoken dialogue systems using hybrid-HMMs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 208–217.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2017b. Analysis of temporal features for interaction quality estimation. In *Dialogues with Social Robots*, pages 367–379. Springer.

Christophe Vaudable and Laurence Devillers. 2012. Negative emotions detection as an indicator of dialogs quality in call centers. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5109–5112.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with Paradise. *Natural Language Engineering*, 6:363–377.

Nigel G. Ward. 2019. *Prosodic Pattterns in English Conversation*. Cambridge University Press.

Nigel G. Ward. 2021. Midlevel prosodic features toolkit (2016-2021). https://github.com/nigelgward/midlevel.

Nigel G. Ward, Jason C. Carlson, and Olac Fuentes. 2018. Inferring stance in news broadcasts from prosodic feature configurations. *Computer Speech and Language*, 50:85–104.

Nigel G. Ward and James A. Jodoin. 2019. A prosodic configuration that conveys positive assessment in American English. In *International Congress of the Phonetic Sciences*.

Geoffrey Zweig, Olivier Siohan, George Saon, Bhuvana Ramabhadran, Daniel Povey, Lidia Mangu, and Brian Kingsbury. 2006. Automated quality monitoring for call centers using speech and NLP technologies. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 292–295.

Appendix: Supplementary Materials

Transcript of a doomed dialog. Post-utterance tags indicate how many annotators marked each for disappointment. The audio is available at the paper website: http://www.cs.utep.edu/nigel/disappointment.

2:10 M How can I help you today?

2:12 C Well, I have a Honda Civic and I need to repair a tire that is flat.

2:22 M Alright, you got a flat? So right now our shop's pretty busy and so if you wanted it repaired today we're gonna have to add a forty dollars just for convenience because we're really booked today and then it would be a ten dollar tire repair. But, I could help you out with a deal. I can give you a bundle and I can waive that convenience fee. So let me tell you some bundles we have.

2:45 C Alright. **d**(1)

2:46 M So the first one we have is the Dream Car bundle. It comes with a car detail, a tire rotation, a full tire inspection, and the tire repair for only two hundred ten dollars.

2:57 C Alright, it's off my budget. **d**(1)

3:01 M Little bit off your budget? How about the Premium bundle then? It comes with a car wash, a tire rotation, and tire repair for a hundred fifty.

3:12 C Alright, it's very off my budget. **d**(3) I only have ten dollars to spend and I only need that tire fixed. **d**(2)

3:23 M Okay, well, how 'bout, I could, let me introduce you to our lowest bundle then. I know you only have ten and this one's sixty, but it's the Ease of Mind bundle because when you fix the tire you want to make sure everything else is fine so we'll fix the flat and we'll do a complete tire inspection and make sure there aren't any holes in any of your tires. And you know, I think it's the best option really because you get to look at everything and make sure everything is okay with your car. It gives you the ease of mind.

3:50 C And it cost, how much?

3:55 M Sixty dollars.

3:56 C Sixty dollars? **d**(2)

3:58 M Yes.

3:59 C Oh. **d**(3) I can't, I really can't. **d**(3) Can you, you can't fix it for ten dollars? **d**(1) Can you, I need the tire ready tomorrow at 6 PM. **d**(1)

4:13 M Oh okay, well the best I can do then without a bundle would just be the fifty dollars with the tire repair for ten dollars and the convenience fee since there's not gonna be a bundle. Is that okay?

4:29 C Can you repeat please?

4:31 M So the only option I can give you then would be the standard tire repair, but since we weren't able to come to an agreement on the bundle it would still have that forty dollar convenience fee so it would come out to fifty dollars. Is that okay?

4:45 C So it's forty dollars? You're saying?

4:50 M Yes.

4:51 C Yeah, I can't. **d**(4) I really can't, I'm sorry. **d**(4)

4:54 M Okay, well I'm sorry we weren't able to help you sir.

4:57 C Yeah, no problem.

4:59 M Alright, well have a good day.

5:02 C You too. Thank you, good bye.