The prosodic expressions of negative micro-sentiment vary across corpora

Dimitri E. Lyon and Nigel G. Ward

Abstract Dialog systems could benefit from the ability to detect instantaneous user dissatisfaction and other negative micro-sentiments. While this ability has been developed for some specific corpora, the question of generality has not previously been broadly examined. In this paper we investigate whether the prosodic expression of negative microsentiment is similar across three diverse corpora: of cooperative gameplay, commercial dialogs, and casual conversations. We find very weak cross-domain performance and major differences in the patterns found, suggesting that no simple model can handle negative microsentiment in general.

1 Motivation

Microsentiments, by which we mean sentiments that appear over short periods of time, such as over an utterance or sub-utterance, are very important in real-time spoken dialog. For dialog systems, the ability to detect the user's negative microsentiments would support monitoring dialog quality and the appropriate adjustment of behavior and policies, either in real time or offline training. Ideally we would like a general-purpose model of negative microsentiment, even if not optimal for any specific corpus, one generally valid across domains.

However, while microsentiment modeling and continuous speech emotion recognition have recently received significant attention [13, 17, 27, 18, 22], and despite much work on cross-domain emotion recognition (e.g. [7]), it is not known whether microsentiment is expressed similarly across domains. In this paper we report a first

Dimitri E. Lyon
University of Texas at El Paso, 500 W University Ave, El Paso, TX 79968,
e-mail: lyondimitri777@gmail.com

Nigel G. Ward
University of Texas at El Paso, 500 W University Ave, El Paso, TX 79968
e-mail: nigelward@acm.org

examination of this question, limiting attention to negative microsentiment and to its prosodic aspects. We focus on prosody because, while lexical information is often informative, vocabularies can be large, and domain adaptation is a challenge [32, 12]. In contrast, the number of prosodic patterns involved in conveying negative sentiment is probably on the order of a dozen, meaning that the prospects for domain-general detection may be greater. Further, there is evidence that people "express positivity more in their word choices, whereas negativity is expressed more through tone of voice" [**?**].

We approach this research question by building models of the prosodic correlates of negative sentiment in three corpora. We then compare within-domain and cross-domain performance and examine what the models learned.

2 Related Work

In this section we relate our question to some major research areas and some relevant recent studies.

An important line of work has considered the potential value of inferring microsentiment for purposes of training dialog systems [21, 16, 22]. Most work on this topic, even when addressing spoken dialog, has focused on text-based features [14, 9], although microsentiment is often conveyed with prosodic, non-lexical behaviors. The most directly relevant work is the tradition overviewed in [22], developing fine-grained models of user satisfaction with dialog systems. The focus has, however, generally been on predicting these from dialog-manager behaviors and speech recognition output and associated measures, rather than exploiting information in the user's behavior. [22] also explored cross-domain generality, but only indirectly, by measuring the value of dialog policies learned from user satisfaction modeling in one information-giving domain when applied to a different information-giving domain.

Another important general line of work has explored the expression of sentiment, including various subtle multimodal indications. However the focus of most of this work has been on feelings regarding non-present people, brands, happenings and so on, rather than here-and-now microsentiments. For many purposes, microsentiment detection may be more useful, as microsentiments are commonly directed to some specific thing that the interlocutor or speaker has just done, said, or mentioned. Moreover, the time scale is of most work on sentiment is tens of seconds or minutes, as for product-review videos [31], where the feeling is constant, rather than varying at the utterance- or frame-level. Furthermore, most of this work has examined monologue data, not dialog.

Looking specifically at modeling microsentiment in spoken dialog, there are two very relevant studies. [17] used pretrained models to predict utterance-by-utterance sentiment in the Switchboard Corpus. The results were fairly good, but generality of the model was not examined. [27] investigated user dissatisfaction at the utterance level in a corpus of simulated commercial dialogs. Their model classified both utterances and individual 10 ms frames, based on prosodic features of the local context. The frame-level classification performance was modestly above baseline, but supported reasonably good utterance-level discriminations. Again, the generality of the model was not examined.

Linguistic studies have shown that many complaints and contrasts (which often contrast a disappointing reality to some ideal situation) involve one or two specific prosodic patterns spanning a phrase or clause [19, 24]. Early attempts to build explainable or concise models of the prosody of sentiment in-the-large [4, 8] have largely been abandoned, in favor of black box modeling. However, by shifting the focus to microsentiment, we hope that the prosodic indications become simpler and easier to model.

3 Data

We investigated three corpora, chosen for their realism, their relevance to dialog-systems applications, and their diversity, representing three genres: cooperative gameplay, task-oriented dialog, and casual conversation.

3.1 Watergirl Corpus

The Watergirl corpus consists of dialogs between people playing the online cooperative two-player game Fireboy and Watergirl [25]. This was collected to investigate the possibility of creating a robot player, ideally as interactive and fun as a human partner. As a first step to building such a player, we wish to explore the possibility of instantaneously tracking the players' near-continuous subtle indications of their level of satisfaction or dissatisfaction. (Eventually we also wish to support construction of a robot player able to effectively and convincingly convey its own current state, dissatisfied or otherwise.) In this game the causes of dissatisfaction were diverse, but included falling in a hazard and having to restart a level, noticing that the other player had made a bad choice, failing to execute a move properly, or realizing that a challenge was harder than it first appeared.

Microsentiment in this corpus has not been previously studied. As a preliminary, we hired a student to provide annotations. Following a simple annotation guide [26], he marked all clear instances of dissatisfaction, even if not particularly strong, basing his decisions on tone, word choice, timing, and context. Labels were generally applied to turns, but if the speaker's tone shifted within a turn, the turn was split into smaller regions. Each turn was given one of seven labels, or left unlabeled, as described in Table 1.

Table 1 Watergirl Corpus Annotation Categories and Counts

Label	Count	Description
ds	555	Dissatisfied with self
do	403	Dissatisfied with other person
dg	95	Dissatisfied with game or game progress
dr	147	Repair/Correction; When someone either indicates that they don't understand, or the other person points it out.
d	78	Miscellaneous Dissatisfaction
p	798	Pleased; Used for very pleased utterances
o	67	Outside speaker/Out of character
No Label		Neutral/Normal

3.2 UTEP Dissatisfaction Corpus

The UTEP Dissatisfaction Corpus (UTEP Calls, for short) is a set of mock commercial telephone calls [3, 2, 27]. It was collected to support research on the automatic detection of conversers with nefarious aims, by examining the ways interlocutors responded to the offers they made. Labeling was done to mark utterances, or occasionally sub-utterances, that indicated dissatisfaction; this variously surfaced as incredulity, annoyance, anger, and in other ways. Negative microsentiment in this corpus has previously been modeled [27], however here we used the final full corpus, using the partitions detailed in "train-dev-test-sets.txt" in [2].

3.3 Switchboard Corpus

Switchboard is a collection of informal telephone conversations between strangers [10]. Negative sentiment is not uncommon and is in part indicated prosodically [28]. We use the SWBD-sentiment annotations of [5], where crowdworkers annotated each utterance as negative, neutral, or positive in sentiment. Annotators were also required to provide justifications for their positive and negative labels. We only used utterances which were annotated by three people, the default, and which were unanimously given the same label. Sentiment in this corpus has previously been modeled by [17].

4 Methods

As noted above, our method centered around training per-corpus models of negative microsentiment, each with the task of categorizing each frame as being within a region labeled negative or non-negative. Each model used only prosodic features of the local context.

4.1 Data Preparation

We split each corpus into training, dev, and test sets, such that each test set contained only unseen dialogs, giving the frame counts seen in Table 2.

Table 2 Number of frames of each time for each corpus and subset

Corpus	Class	Train	Dev	Test	Total
Watergirl, explicit labels only	Neu	51950	19302	20127	91379
	Neg	51950	45289	51371	148610
Watergirl, with augmented labels	Neu	120222	285583	274147	679952
	Neg	120222	45389	51371	216882
UTEP Calls	Neu	7899	14084	54543	76526
	Neg	7899	7581	20893	36373
Switchboard	Neu	19224	31454	32411	83089
	Neg	19371	6689	6081	32141

For the UTEP Calls and Watergirl corpora, if the annotation label is “n”, “nn”, or “p”, the frame is considered neutral and assigned a *y* value of 0. If the annotation label is “d”, “dd”, “ds”, “do”, “dg”, or “dr”, the frame is considered dissatisfied, and assigned a *y* value of 1. If the annotation label is “o”, the frame is excluded. For the Switchboard data, both neutral and positive frames were mapped to 0 and negative to 1. By combining classes we lost information, but enabled the three corpora to be treated similarly.

We also did an alternate data preparation for the Watergirl corpus. Since its annotation focused on the negative labels, with explicit positive labels given only if the speaker was judged very pleased, there was an imbalance in the training data. Accordingly we tried expanding the amount of training data by exploiting the fact that unlabeled utterances were implicitly neutral in sentiment. Thus the augmented data included all (implicitly or explicitly) neutral speech frames together with the positive frames in the non-negative category. This alternate corrected for the scarcity of positive labels, but at the cost of an imbalance in the opposite direction.

4.2 Features

In order to characterize the local prosodic context of each frame to be classified, we used features from approximately 3 seconds on either side of that frame. These features measure or proxy for intensity, pitch height, pitch range, speaking range, creakiness and Cepstral Peak Prominence (CPPS) [23, 27, 1]. (CPPS has been found to effectively measure breathiness in clinical applications [11], and reasonably well also for dialog [29].) These features are well-normalized and designed to be robust. We used a total of 125 features, including 124 for these prosodic features, computed over various windows and offsets, and 1 for the time into dialog. These are extracted for each frame to classify, that is to say, computed with a stride of 10 milliseconds.

4.3 Metrics

Our primary metric for prediction performance is $F_{.25}$. We chose this because, for any practical application, precision will probably be more important than recall: it would be more important for a model to have a higher certainty for the frames that it determines to be negative than for it to find every instance of negativity. We also report precision, and recall. Our baseline for comparison is a trivial model that just guesses that every frame is dissatisfied.

4.4 Models

Since our aim here was to explore rather than to optimize performance, since two of the data sets were quite small, and because we wanted interpretable models, we used mostly simple linear regression. We built models separately for each of the three corpora. In each case, the threshold was chosen to maximize the F-score on the devset data (in this respect differing from our previous study on the UTEP Calls Corpus [27]).

For the Switchboard corpus, we also tried another model: k Nearest Neighbors. This was because we noted a greater diversity in the types of negative micro-sentiment in this corpus. Since the corpus is so large, to reduce the time for each experiment, we randomly culled it to 1 in 100 frames. We used Euclidean distance across the 125 features for determining the nearest neighbor. We used $k = 1$, as increasing k did not appear to significantly increase performance.

5 Results

Before approaching our main question, regarding the generality of negative sentiment models across corpora, we explored how well these modeling methods worked for each individual corpus.

First, for the Watergirl corpus, As seen in the first two lines of Table 3, this model comfortably outperforms the baseline. Further, with the alternate data preparation, with the augmented labels, the result was much better, as seen in Table 4. (Because the data sets are different, the baseline here is different from that in Table 3.)

Second, for the UTEP Calls corpus, as seen in the first two lines of Table 5, the model outperformed the baseline. While direct comparison to previous work [27] is not possible, the performance is in the same ballpark.

Table 3 Detection quality for the Watergirl corpus (for explicit labels only).

Classifier	$F_{.25}$	Precision	Recall	MSE
baseline	0.73	0.72	1.00	0.28
same-corpus training	0.84	0.90	0.42	0.45
training on UTEP Calls	0.76	0.82	0.34	0.52

Table 4 Detection quality for the Watergirl corpus, with label augmentation.

Classifier	$F_{.25}$	Precision	Recall	MSE
baseline	0.17	0.16	1.00	0.84
same-corpus training	0.39	0.39	0.46	0.20

Table 5 Detection quality for the UTEP Call corpus.

Classifier	$F_{.25}$	Precision	Recall	MSE
baseline	0.29	0.28	1.00	0.72
same-corpus training	0.30	0.30	0.33	0.40
training on Watergirl	0.27	0.27	0.34	0.44

Table 6 Detection quality for the Switchboard corpus.

Classifier	$F_{.25}$	Precision	Recall	MSE
Baseline	0.17	0.16	1.00	0.84
same-corpus training, linear regression	0.18	0.17	0.50	0.46
same-corpus training, kNN	0.17	0.16	0.54	0.51
trained on Watergirl, linear regression	0.16	0.16	0.34	0.39
trained on UTEP Calls, linear regression	0.18	0.17	0.35	0.37

Third, for the Switchboard corpus, as seen in Table 6, performance of the linear regression model was only slightly better than baseline. The k-nearest neighbors model was slightly below baseline. The performance was much lower than that obtained by [17], doubtless due to the simplicity of our models and our use of only prosodic features.

Additionally, we did testing on each individual label in the Watergirl corpus. We trained each model on data that was labeled either with that label, or with a neutral or positive label. The results are shown in Table 7. The regressor outperformed the baseline at $p < 0.05$ when trained on the labels “ds”, “do”, “dg”, and “dr”. The regressor had the highest error reduction when trained and tested on repairs/corrections in speech, which had the annotation “dr” 1. The regressor had the highest F-score when trained and tested on dissatisfaction expressed at the other person, which had the label, “do” 1.

As a way to gauge the reliability of these results, we used chi-square tests to judge the performance of each model, in terms of correct versus incorrect frame identifi-

Table 7 Performance metrics of Watergirl Linear Regression on each individual label

Label	Model	$F_{.25}$	Precision	Recall	MSE
d	Baseline	0.12	0.12	1.00	0.88
	Regressor	0.12	0.12	0.19	0.25
ds	Baseline	0.50	0.49	1.00	0.51
	Regressor	0.69	0.75	0.31	0.39
do	Baseline	0.50	0.49	1.00	0.51
	Regressor	0.84	0.87	0.53	0.27
dg	Baseline	0.14	0.14	1.00	0.86
	Regressor	0.21	0.20	0.49	0.34
dr	Baseline	0.27	0.26	1.00	0.74
	Regressor	0.68	0.68	0.66	0.17

cations, compared to the best prosody/content-ignorant baseline we could imagine: predicting at random according to the actual frequencies of positive and negative frames in the test set for each corpus. At $p < 0.05$, we found statistically significantly better performance for all same-corpus Watergirl models, the UTEP Calls-trained Watergirl model, the same-corpus linear regression Switchboard model, and the UTEP Calls trained Switchboard model. However, we did not find statistical significance in the difference for same-corpus UTEP Calls model.

We turn now to our main research question, that of cross-domain performance. First, for the UTEP Calls corpus, performance of a Watergirl-trained model was below baseline, as seen in the last line of Table 5. Second, for Watergirl, the model trained on Calls was modestly above baseline, as seen in the last line of Table 3, although, unsurprisingly, not as good as the model trained on the same corpus. Third, for Switchboard, as seen in the last two lines of Table 6, while performance was poor for the Watergirl model, the Calls-trained model did relatively well, both in $F_{.25}$ and in MSE. Given the weak performance of the Switchboard model, we did not even try applying it to the other corpora. Overall, we see that cross-domain performance was weak at best.

6 Analysis and Discussion

Figures 1 – 4 show the coefficients of the various linear regression models for the prosodic features. There are possibly two weak tendencies towards commonalities: three out of the four models had a positive correlation between dissatisfaction and intensity 400ms after the frame being predicted, as seen in Figures 1, 2, and 3, and two of the four models had a negative correlation between dissatisfaction and pitch narrowness in the -1600ms to -800ms window, as seen in Figures 2 and 3. However, overall from the figures it is clear that the negative micro-sentiment patterns learned differ across the corpora. This also matches the impression given by post-hoc casual observation of examples of negative sentiment in the various corpora.

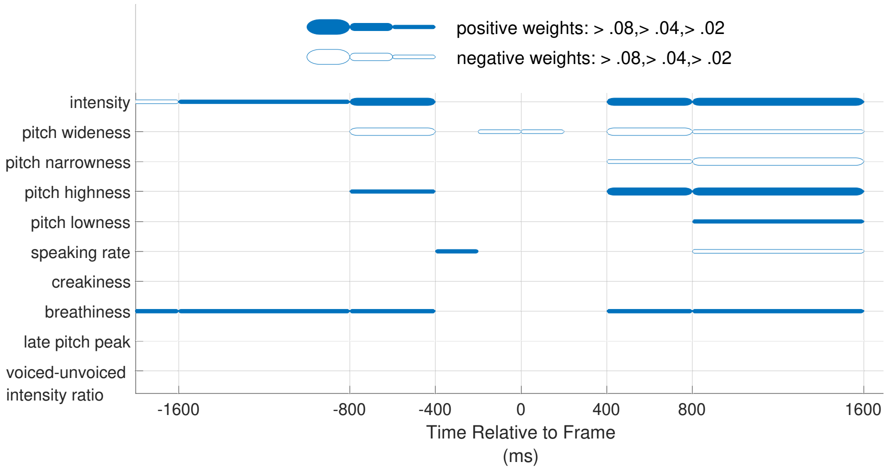

Fig. 1 Visualization of the linear regression model coefficients for with the Watergirl corpus using only the explicitly labeled frames.

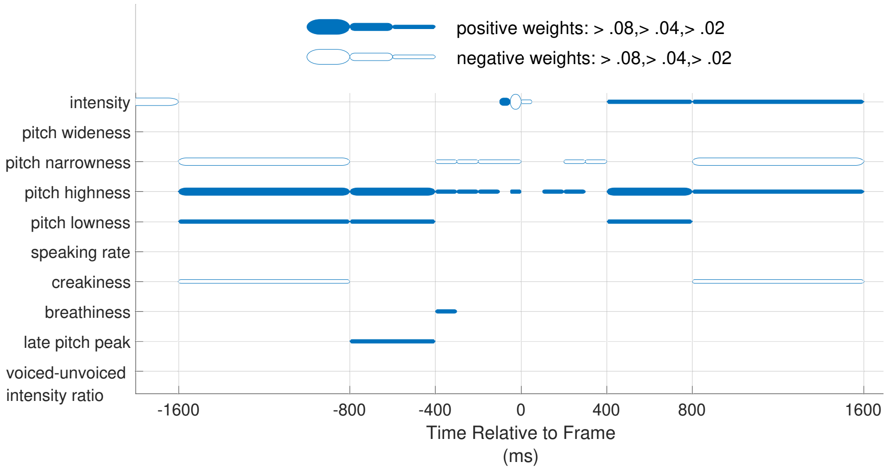

Fig. 2 Visualization of the linear regression model coefficients for the Watergirl corpus using the augmented labels.

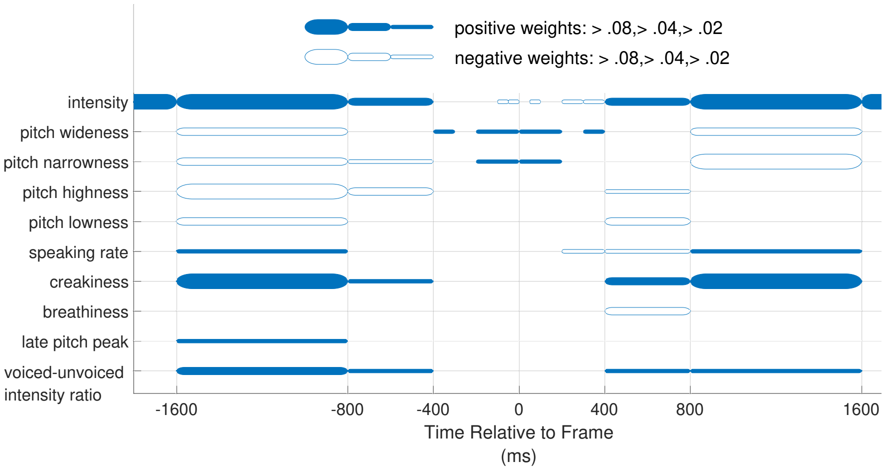

Fig. 3 Visualization of the linear regression model coefficients for the UTEP Dissatisfaction Call corpus

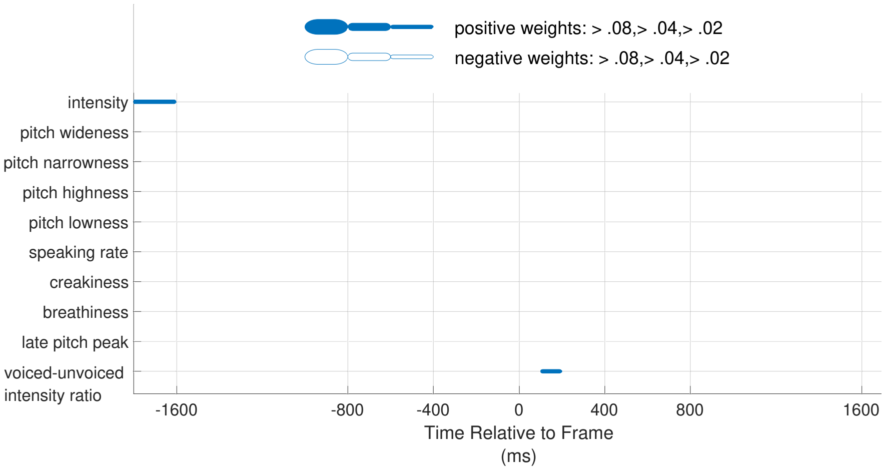

Fig. 4 Visualization of the linear regression model coefficients for the Switchboard corpus.

For the Watergirl corpus, the good performance suggests that a prosody-based model would be useful for real-time monitoring of the user's satisfaction with the partner's gameplay and other factors. The most salient and consistent correlation is with regions rich in high pitch; this is easy to understand as connecting with the fact that in this game disasters happen mostly near times of high excitement and engagement, and that affect bursts at the moment of disaster are often in high pitch [25, 20].

For the Calls corpus, the performance was only modest at the frame level, but would be able to support better performance for utterance-level classifications, as seen in previous work [27]

The Switchboard model had only has two coefficients with an absolute value above the display threshold, .02, as seen in Figure 4. One possible interpretation is that negative micro-sentiment in Switchboard is not consistently well expressed by any single prosodic pattern. However there is also another explanation, suggested by examining a few of the annotations in SWBD-sentiment: these annotations do not seem to be consistently about sentiment. For example, a discussion of the fact that hunting dogs should not be kept in the house, to avoid spoiling them, was annotated as negative in sentiment by all three annotators, with the evidence being the presence of the word *spoil*, although when we read or listened to this utterance we detected no negativity at all. In general, it may be that the SWBD-sentiment annotations relate more to lexical valence than to sentiment in the normal sense of the word. Thus it seems that the question of the utility of prosody for detecting negative microsentiment in casual conversation remains open.

7 Summary, Implications, and Future Work

We found that simple prosodic models were effective for identifying moments of negative microsentiment in a corpus of cooperative gameplay, but only marginally useful for free conversation.

We also found, contrary to expectation, that our prosody-based models for detecting negative micro-sentiment did not generalize well across corpora. This suggests that, at least over the short term, models will need to be trained for each corpus or corpus type. More generally, despite the tendency to sometimes consider “sentiment” as if it were a unitary construct, this suggests the existence of significant diversity within negative sentiment.

While better performance is very likely obtainable using better methods — such as the inclusion of features representing more aspects of prosody, wider contexts, and additional types of information (lexical, visual, turn timing, etc.) [6], and the use of pretrained models instead of hand-crafted features [30, 15] — we think that our conclusion about the diversity of the expressions of negative microsentiment will still hold.

While there still may be, over the long term, the potential to create a unifying model of negative micro-sentiment, we think that would need to take into account at least differences in genre, domain, and interaction style [28], and also variation in the intensity, in the triggers of the sentiment, and in the discourse goals served by expressing the negative sentiment. The result could have general value but would be far from simple.

Acknowledgments

We thank Adrian Avendano for annotating the Watergirl corpus, Jonathan Avila for help creating the annotation guidelines and for Matlab advice, and the anonymous reviewers for incisive comments.

References

1. Jonathan Avila and Dimitri Lyon. Models for estimating dissatisfaction in spoken dialog. https://github.com/DimitriLyon/dissatisfaction-models.
2. Jonathan Avila, Nigel Ward, and Aaron Alarcon. The UTEP corpus of dissatisfaction in spoken dialog, March 2021. https://github.com/joneavila/utep-dissatisfaction-corpus.
3. Jonathan E. Avila, Nigel G. Ward, and Aaron M. Alarcon. The UTEP corpus of dissatisfaction in spoken dialog. Technical Report UTEP-CS-21-23, University of Texas at El Paso.
4. Anton Batliner, Stefan Steidl, Bjorn Schuller, et al. Whodunnit: Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech and Language*, 25:4–28, 2011.
5. Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. A large scale speech sentiment corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6549–6555, Marseille, France, May 2020. European Language Resources Association.
6. Shammur Absar Chowdhury, Evgeny A Stepanov, and Giuseppe Riccardi. Predicting user satisfaction from turn-taking in spoken conversations. In *Interspeech*, pages 2910–2914, 2016.
7. Laurence Devillers, Christophe Vaudable, and Clément Chastagnol. Real-life emotion-related states detection in call centers: a cross-corpora study. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
8. 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. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:190–202, 2016.
9. Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics*, pages 4194–4204, 2022.
10. John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.
11. Yolanda D Heman-Ackah, Deirdre D Michael, Margaret M Baroody, Rosemary Ostrowski, James Hillenbrand, Reinhardt J Heuer, Michelle Horman, and Robert T Sataloff. Cepstral peak prominence: a more reliable measure of dysphonia. *Annals of Otology, Rhinology & Laryngology*, 112(4):324–333, 2003.
12. Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Bjorn Wolfgang Schuller. Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. *IEEE Transactions on Affective Computing*, to appear.
13. Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke. Acoustic and lexical sentiment analysis for customer service calls. In *IEEE ICASSP 2019*, pages 5876–5880, 2019.
14. Weixin Liang, Kai-Hui Liang, and Zhou Yu. HERALD: an annotation efficient method to detect user disengagement in social conversations. In *ACL*, pages 3652–3665, 2021.
15. Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung yi Lee, and Nigel G. Ward. On the utility of self-supervised models for prosody-related tasks. In *IEEE Workshop on Spoken Language Technology (SLT)*, 2022.

16. Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765, 2020.
17. Zhiyun Lu, Liangliang Cao, Yu Zhang, Chung-Cheng Chiu, and James Fan. Speech sentiment analysis via pre-trained features from end-to-end ASR models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7149–7153, 2020.
18. Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau. On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 373–380, 2021.
19. Richard Ogden. Prosodic constructions in making complaints. In Dagmar Barth-Weingarten, Elisabeth Reber, and Margret Selting, editors, *Prosody in Interaction*, pages 81–103. Benjamins, 2010.
20. Marc Schroder. Experimental study of affect bursts. *Speech Communication*, 40:99–116, 2003.
21. Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010, 2021.
22. Stefan Ultes and Wolfgang Maier. User satisfaction reward estimation across domains: Domain-independent dialogue policy learning. *Dialogue & Discourse*, 12(2):81–114, 2021.
23. Nigel Ward. Midlevel prosodic features toolkit, February 2022. https://github.com/nigelgward/midlevel.
24. Nigel G. Ward. *Prosodic Pattterns in English Conversation*. Cambridge University Press, 2019.
25. Nigel G. Ward and Saiful Abu. Action-coordinating prosody. In *Speech Prosody*, pages 629–633, 2016.
26. Nigel G. Ward and Jonathan E. Avila. Dissapointment project annotation guide: Watergirl version. at https://www.cs.utep.edu/nigel/microsentiment/, June 2021.
27. Nigel G. Ward, Jonathan E. Avila, and Aaron M. Alarcon. Towards continuous estimation of dissatisfaction in spoken dialog. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 13–20, 2021.
28. Nigel G. Ward and Jonathan E. Avlia. A dimensional model of interaction style variation in spoken dialog. *Speech Communication*, 2022, submitted.
29. Nigel G. Ward, Ambika Kirkland, Marcin Włodarczak, and Eva Székely. Two pragmatic functions of breathy voice in American English conversation. In *11th International Conference on Speech Prosody*, pages 82–86, 2022.
30. Yangyang Xia, Li-Wei Chen, Alexander Rudnicky, Richard M Stern, et al. Temporal context in speech emotion recognition. In *Interspeech*, pages 3370–3374, 2021.
31. AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.
32. Biqiao Zhang, Emily Mower Provost, and Georg Essl. Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. *IEEE Transactions on Affective Computing*, 10(1):85–99, 2019.