Classification of Feedback Functions in Spoken Dialog Using Large Language Models and Prosodic Features

Carol Figueroa
Furhat Robotics
Aix-Marseille University
carol@furhatrobotics.com

Magalie Ochs
Aix-Marseille University
magalie.ochs@lis-lab.fr

Gabriel Skantze
KTH
Royal Institute of Technology
skantze@kth.se

Abstract

Feedback utterances such as 'yeah', 'mhm', and 'okay', convey different communicative functions depending on their prosodic realizations, as well as the conversational context in which they are produced. In this paper, we investigate the performance of different models and features for classifying the communicative function of short feedback tokens in American English dialog. We experiment with a combination of lexical and prosodic features extracted from the feedback utterance, as well as context features from the preceding utterance of the interlocutor. Given the limited amount of training data, we explore the use of a pre-trained large language model (GPT-3) to encode contextual information, as well as SimCSE sentence embeddings. The results show that good performance can be achieved with only SimCSE and lexical features, while the best performance is achieved by solely fine-tuning GPT-3, even if it does not have access to any prosodic features.

1 Introduction

In human-human conversations, short feedback tokens such as 'mhm', 'yeah', and 'wow' serve different communicative functions. For example, 'yeah' can indicate a response to a question, express agreement to an opinion, convey surprise, or simply signal that the interlocutor should continue speaking, depending on the prosodic realization, as well as the conversational context. The terms *feedback* and *backchannels* are sometimes used interchangeably. However, in this paper we use the term backchannel to denote a specific type of feedback that signals that the speaking partner should continue speaking.

There has been a lot of work on incorporating user-generated or system-generated feedback in dialog systems and human-robot interactions (Axelsson et al., 2022). Most work on incorporating feedback in dialog systems have focused on the timing of backchannels (Ward and Tsukahara, 2000; Ruede et al., 2017, 2019; Morency et al., 2010; Adiba et al., 2021; Boudin et al., 2021; Ishii et al., 2021). There has also been work on predicting which type of backchannel or feedback to produce (i.e., predicting what function the backchannel or feedback should convey) (Kawahara et al., 2016; Ortega et al., 2020; Adiba et al., 2021; Boudin et al., 2021; Jang et al., 2021; Lala et al., 2022). In this paper, we focus on the classification of the communicative function of short feedback tokens (i.e., assign the function of feedback), given their lexical and prosodic form as well as the preceding conversational context.

A model that automatically classifies the communicative function of feedback can be used for different purposes. When used offline, such a model could be used to automatically annotate the functions of feedback in a speech corpus. The annotated feedback can then be used to, for example, gain insights into human conversational behavior, or to learn how to synthesize feedback with appropriate prosody, given the feedback function. When used online in a spoken dialog system, it could be used to classify feedback coming from the user.

In this paper, we investigate the performance of different models and features for classifying the communicative function of short feedback tokens in the Switchboard corpus (Godfrey et al., 1992). We use our previously proposed annotation scheme (Figueroa et al., 2022), consisting of 10 feedback functions: continue, non-understanding, agree, disagree, yes/no response, sympathy, mild/strong surprise, and disapproval. For the classification task, we use lexical and prosodic features from the short feedback token, as well as contextual features from the preceding utterance of the interlocutor. Since the representation of dialog context is non-trivial, especially considering the limited amount of annotated data at our disposal, we also investigate the use of probability distributions from a pre-trained large language model (GPT-3) as input to a Support Vector Machine (SVM) classifier, along with the

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

previously mentioned features. To the best of our knowledge, probability distributions from GPT-3 have not been used as an input to another machine learning algorithm for this problem before.

2 Related Work

The classification of feedback functions is related to the more general problem of Dialog Act classification, where the goal is to identify the communicative function of an utterance in dialog. However, in most Dialog Act classification schemes, backchannels are typically treated as a single dialog act category and no fine-grained distinctions are made (Stolcke et al., 2000; Dielmann and Renals, 2008; Liu et al., 2017).

When it comes to the more specific problem of classifying the communicative function of feedback, the only related work we are aware of are Prévot et al. (2015), Neiberg et al. (2013), and Gravano et al. (2007). Although Gravano et al. (2007) do not specifically classify feedback, they do classify affirmative words, which function as a backchannel or acknowledgment/agreement. They use JRIP, a machine learning algorithm to classify affirmative words using text-based, timing, and acoustic-prosodic features from both the affirmative words and context preceding and following the affirmative words. While Neiberg et al. (2013) do not propose a classifier for feedback functions, they use semi-supervised annotations and prosodic clustering to investigate how different prosodic realizations of feedback affect the function of feedback tokens. In Prévot et al. (2015), feedback functions are classified into two levels: *base* function, and *evaluation* function, which respectively correspond to *generic* and *specific* listener responses (Bavelas et al., 2000). A Random forest classifier is first used to classify feedback in the *base* level into the following functions: contact, acknowledgement, evaluation-base, answer, elicit or other. If the feedback is classified into the evaluation-base function, another Random forest classifier is used to classify the feedback into the following functions: approval, expectation, amusement, or confirmation/doubt. Lexical, acoustic, and position information is used of the feedback. Bigrams and the function of the context (the previous utterance) are also used for the classification task.

Feedback Function	**Count**	**GPT-3 prompt label**
(C) Continue	1024	Continuer
(U) Non-understanding	63	Misunderstand
(A) Agree	435	Agree
(D) Disagree	46	Disagree
(Y) Yes-response	56	Yes-answer
(N) No-response	114	No-answer
(S) Sympathy	82	Sympathy
(MS) Mild Surprise	103	Interest
(SS) Strong Surprise	191	Surprise
(Ds) Disapproval	65	Reproach
(O) Other	77	Other

Table 1: Feedback functions, count of manually annotated data, and corresponding labels in GPT-3 prompt.

3 Communicative Functions of Feedback

A number of annotation schemes have been proposed for annotating the communicative functions of feedback (Allwood et al., 1992, 2007; Bunt, 2009; Buschmeier et al., 2011; Neiberg et al., 2013; Prévot et al., 2015, 2016; Malisz et al., 2016; Figueroa et al., 2022). As mentioned, feedback can be categorized as having two communicative functions: *generic* and *specific* (Bavelas et al., 2000; Prévot et al., 2015, 2016; Ortega et al., 2020; Boudin et al., 2021). Generic feedback can be thought of as *continuers*; they encourage the interlocutor to continue speaking (Schegloff, 1982). Specific feedback can be thought of as *assessments*; they are listener responses that depend on the context of the interlocutor (Goodwin, 1986). The DIT++ taxomony of dialogue acts also categorizes feedback by two functions, *allo-feedback* and *auto-feedback* which carry information about attention, perception, interpretation, evaluation, and execution of the feedback.

Allwood et al. (1992) introduced four communicative functions of feedback:

- *Contact*: whether the interlocutor is willing and able to continue the interaction
- *Perception*: whether the interlocutor is willing and able to perceive the message
- *Understanding*: whether the interlocutor is willing and able to understand the message
- *Attitudinal reactions*: whether the interlocutor is willing and able to react and (adequately)

respond to the message, specifically whether he/she accepts or rejects it.

These four feedback functions are related to the four levels of joint actions of an addressee proposed by Clark (1994) which are important for establishing common ground. The four feedback functions introduced by Allwood et al. (1992) have inspired many annotations schemes for annotating functions of feedback (Allwood et al., 2007; Buschmeier et al., 2011; Malisz et al., 2016; Neiberg et al., 2013).

In this work, we use our previously proposed annotation scheme (Figueroa et al., 2022), consisting of 10 feedback functions: continue, non-understanding, agree, disagree, yes/no response, sympathy, mild/strong surprise, and disapproval. The scheme also includes an Other category that is used to capture lexical tokens that are not feedback but share the same lexical form as feedback, for example, discourse markers ('okay, let's begin') or literal uses ('he was standing on the right'). The feedback functions continue and understanding can be thought of being in the contact, perception, or understanding grounding level, whereas the other feedback functions are on the attitudinal grounding level.

4 Method

4.1 Corpus and feedback functions

We extracted short feedback tokens from the Switchboard corpus (Godfrey et al., 1992), according to the definition and selection criteria given by Figueroa et al. (2022). Switchboard consists of about 2,500 dyadic telephone calls between 500 native speakers of American English, recorded in two separate channels and lasting about 3-10 minutes. The corpus also contains transcriptions and word level time-alignments.

In total, Switchboard contains 85,956 instances of potential feedback tokens, according to the working definition in Figueroa et al. (2022). Note that this definition is based on the lexical form of the token, and thus may include instances which are not in fact feedback, such as discourse markers. Thus, we train our classifier to also classify such instances as Other. From the full set, we compiled a set of 2256 instances, which were manually annotated with one of the 10 communicative functions (plus Other), as identified in Figueroa et al. (2022), by listening to them in context. Table 1 lists these functions and their counts in our data set.

4.2 Feedback features

For the short feedback utterance, we use the lexical token as well as its prosodic realization as features.

Lexical tokens (e.g. 'yeah', 'wow') and non-lexical tokens (e.g. 'mhm', 'hm') were encoded as one-hot encodings using the scikit-learn Python library (Pedregosa et al., 2011).

Prosodic features – duration, mean pitch, pitch slope, pitch range, and mean intensity – were extracted from the feedback instances. We used Parselmouth (Jadoul et al., 2018) to extract pitch (F0 Hz) and intensity (dB) values. The pitch values were first transformed to log scale and then z-score normalized, intensity values were also z-score normalized. The normalization was done per speaker, where the mean and standard deviation for each speaker were computed from their entire conversation. Pitch slope was calculated by subtracting the mean of the z-score normalized pitch values of the second half of the feedback from the mean of the z-score normalized pitch values of the first half of the feedback.

4.3 Context features

We also added contextual features from 4000 ms of the interlocutor's utterance preceding the feedback. Previous work in feedback modeling have extracted features from the context by either setting an arbitrary window length or number of words. We experimented with a window length of between 1500 - 4000 ms and found that 4000 ms often captured full sentences. We decided to only use features from the preceding utterance of the interlocutor (and not any future context) in order to make the model applicable for online classification.

Part-of-speech (POS) tags of the preceding utterance were extracted using the spaCy Python library. From these, POS bigrams were created and sorted by their term frequency-inverse document frequency (TF-IDF), treating the 10 feedback functions as documents and the POS bigrams as terms. From this list, the top 30 bigrams were selected and used as one-hot features.

Dialog Acts were automatically assigned to the interlocutor's utterance using DialogTag (Malik, 2020), a Python libary. We collapsed the following dialog tags into a single 'Question' tag: 'Yes-No-Question', 'Declarative Yes-No-Question', 'Rhetorical-Question', 'Wh-Question', and 'Tag-Question'. The dialog tags were then one-hot encoded.

Sentence Embedding of the previous utterance was obtained using SimCSE (Gao et al., 2021), which is an auto-encoding embedding technique based on contrastive learning. During training, SimCSE uses BERT encodings of the input and then fine-tunes the parameters using the contrastive learning objective which pushes together semantically similar pairs and pushes apart semantically dissimilar pairs. We used the sup-simcse-bert-base-uncased pre-trained model of SimCSE which is readily available on Github (Gao et al., 2021).

4.4 GPT-3

As an alternative to the context features listed above, we also explored the use of GPT-3 from OpenAI (Brown et al., 2020) to encode the previous utterance, as well as the lexical form of the feedback token. We tested three different approaches: zero-short, few-shot, and fine-tuning.

For *zero-shot* classification, we provided GPT-3 with a prompt similar to the one shown in Table 2. The prompt ends with the opening bracket at the end, and GPT-3 is asked to predict the next token (marked in bold). This is done using the davinci-003 model. For *few-shot* classification, we provided an example of each function, with both the dialog and the corresponding label, in addition to the instructions.

The third approach is to *fine-tune* GPT-3. We fine-tuned the davinci base model, since davinci-003 is not available for fine-tuning. For fine-tuning, there are no instructions or examples in the prompt; the model is only given training examples, which consist of input text (the preceding utterance and the feedback token) and its associated output (the function label).

Note that, in Table 2, the feedback function labels in the GPT-3 prompt have been changed from the ones listed in the first column of Table 1. Since GPT-3 generates word pieces, we changed the feedback function labels in the prompt so that they would not start with the same first letters. This way, we can simply inspect the first generated word piece from GPT-3 and map it to one of the functions.

For zero-shot and few-shot classification, we also explored if the prediction could be used as an input feature to the feedback function classifier, rather than using it directly. For this, we use the probability distribution that GPT-3 outputs over potential function labels (or rather their prefix).

GPT-3 Prompt
The following is a list of dialog acts and their description in parentheses: - Continuer (Backchannel) - Misunderstand (Expressing non-understanding) - Agree (Agreeing with a statement) - Disagree (Disagreeing with a statement) - Yes-answer (A positive answer to a yes/no question) - No-answer (A negative answer to a yes/no question) - Sympathy (Expressing empathy) - Reproach (Expressing disapproval or disgust or disappointment) - Interest (Expressing interest) - Surprise (Expressing surprise) - Other (thinking or interrupting conversation) The following is a dialog between two persons. The dialog acts are written in brackets. A: i was mowing the lawn yesterday B: mhm **[continuer]**

Table 2: Prompt given to GPT-3.

From GPT-3, we can get the top five labels that would have been generated by the language model and their corresponding probabilities. For example, given the feedback 'yeah', GPT-3 could predict the following word pieces: 'Ag' 74%, 'Contin' 1.7%, 'Yes' 21%, 'agree' 3%, and 'yes' 0.3%. From these probabilities, we can create a vector where feedback function (A)gree is assigned 77%, (C)ontinue 1.7%, (Y)es-response 21.3%, and all other functions plus the Other category are assigned 0%. These probability distributions can then be used as input features to the main function classifier.

For all GPT-3 models we use the following settings: temperature=0, max_tokens=1, frequency_penalty=0, presence_penalty=0.6, and logprobs=5.

4.5 Function classifier

The task of the main classifier is to classify the feedback function, given the features listed above. As explained, GPT-3 can be used both as a main classifier and as a method for encoding lexical and contextual information, which can then be used as input to another classifier. Since we did not have a large data set to train a deep learning model, we explored three machine learning models which can handle small data sets (Forman and Cohen,

2004): Support Vector Machine (SVM), Logistic Regression, and Random Forest, using the classifiers implemented in scikit-learn a Python library (Pedregosa et al., 2011).

For the three classifiers, we set the parameter class_weight as balanced. For the SVM classifier we used a linear kernel, we also experimented with radial basis function kernel but the linear kernel gave the best results. For the Logistic Regression classifier we set the max_iter to 200. For all other parameters we used the default settings. In general, we found that the SVM classifier performed the best, and thus only report our results from the SVM classifiers.

We experiment with different combinations of input features and evaluate our SVM classifiers using 10-fold cross validation. In order to evaluate the model performance, we use the F1-weighted score. In the cases where we do not fine-tune GPT-3 and use it directly as a classifier, we do not use cross validation, but instead use our entire annotated data.

5 Results

5.1 Classifier performance

Table 3 summarizes the F1-weighted scores of the different models with different combinations of input features. For comparison, we also report the majority-class baseline, as well as the inter-annotator agreement annotations from our previous work (Figueroa et al., 2022). Note that only 1124 feedback utterances were annotated for the inter-annotator agreements. In cases where the annotators could not decide on a single function (e.g. 'A/C'), we chose one of the functions randomly while calculating the F1-weighted score. This procedure was averaged over 10 times.

When only lexical features are used (Model 1), we get a fairly high F1-weighted score (0.63) which outperforms the baseline. We used the majority-class baseline which returns the frequent class label. The prosodic features are not very informative, and adding them to the lexical features do not improve the performance further (Model 16).

Among the contextual features, SimCSE is clearly the most informative (Model 7). Just using GPT-3 as a zero-shot or few-shot classifier or as input features does not appear to be very useful (Models 10,11,13,14), considering that it also encodes lexical information about the feedback utterance; the performance is on par with Model 1 which only uses lexical features. There is also no significant difference between using zero-shot or few-shot (Model 13,14) ($t(18) = 1.585$; $p = 0.13$). Lexical features in combination with SimCSE, on the other hand, do give a better performance: Model 19 performs as well as the inter-annotator agreement score. Figure 1 shows the confusion matrix of Model 19 trained on 1804 examples and evaluated on a 452 test set. We can see that it performs poorly on (D)isagree, (Ds) Disapproval, and (Y)es-response. This poor performance could be due to the few training examples.

To improve this score further, we need to add prosodic features, the dialog act, and the GPT-3 distributions to the lexical and SimCSE features. While Model 20 performs significantly better than Model 19 ($t(18) = 2.509$; $p = 0.02$), the difference is not very big, considering the much larger feature set. Figure 2 shows the confusion matrix of Model 20 trained on the same 1804 examples and

Model #: Features	**F-score**
1: Lexical	0.63
Prosody	
2: Duration	0.10
3: Mean pitch	0.16
4: Pitch slope	0.24
5: Pitch range	0.18
6: Mean intensity	0.15
Context	
7: SimCSE	0.32
8: Dialog act (DA)	0.14
9: Part-of-speech (POS)	0.09
GPT-3	
10: Zero-shot majority*	0.61
11: Few-shot majority*	0.65
12: Fine-tuned*	**0.80**
13: Zero-shot as features (ZS)	0.61
14: Few-shot as features (FS)	0.63
Combinations	
15: Prosody (all)	0.37
16: Lexical + Prosody (LexPro)	0.63
17: Lexical + GPT-3 (ZS)	0.68
18: Lexical + GPT-3 (FS)	0.69
19: Lexical + SimCSE	**0.72**
20: LexPro + SimCSE + DA + GPT-3 (FS)	**0.76**
Majority-class baseline	0.28
Inter-annotator agreement	**0.74**

Table 3: F1 weighted scores for different feature sets. *Uses GPT-3 (and not SVM) as the main classifier.

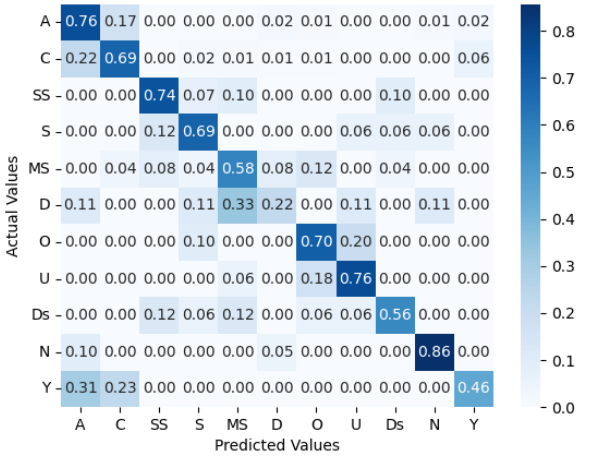

Figure 1: Confusion matrix for Model 19.

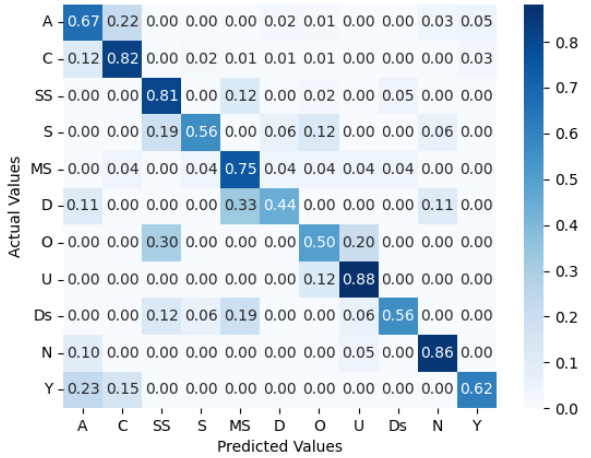

Figure 2: Confusion matrix for Model 20.

evaluated on the same 452 test set. By adding the prosodic features and the dialog act, we can see that it improves the classification for (C)ontinue, (SS) Strong Surprise, (MS) Mild Surprise, (D)isagree, (U) Non-Understanding, and (Y)es-response.

The best performing model is the GPT-3 fine-tuned classifier (Model 12), which performs significantly better than Model 20 (t(18) = -2.803; p = 0.01). Figure 3 shows the confusion matrix of Model 12 trained on the 1804 examples and evaluated on the 452 test set. The GPT-3 fine-tuned classifier improves the classification for (C) Continue, (SS) Strong Surprise, (O) Other, (U) Non-understanding, (N)o-response, and (Y)es-response.

To conclude, models 12, 19, and 20 are all viable classifiers for feedback functions, and they all seem to perform on par with the inter-annotator agreement. The choice of classifier depends on specific requirements, for example whether it should be used offline or online, and whether access to GPT-3 is available.

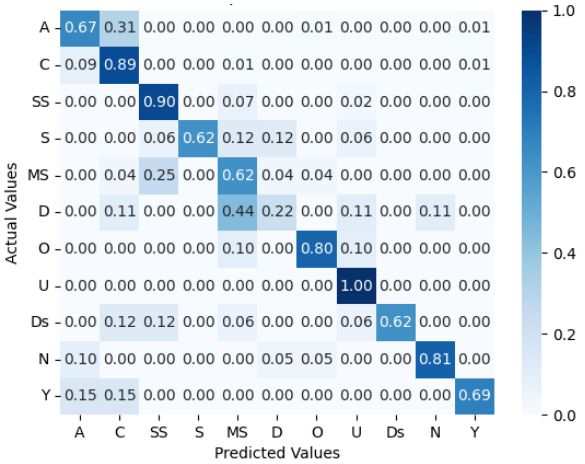

Figure 3: Confusion matrix for Model 12.

5.2 Labeling the remaining Switchboard corpus

Given that we now have working classifiers of feedback functions for Switchboard, we finally experimented with applying one of them to the remaining set of 83,700 potential feedback instances in the Switchboard corpus, in order to study the general distribution of the communicative functions. For this, we used Model 19, as it has a low cost while the performance is relatively good. The distributions are shown in Table 4 which include the distributions of the 2,256 manually annotated lexical tokens and the 83,700 automatically annotated lexical tokens. In total, there were 74,106 instances of actual feedback (not Other), according to the classifier. As can be seen, (C)ontinue and (A)gree are the most frequent feedback functions.

5.3 Investigating sex differences

To illustrate how this classification can be used for further analysis, we also broke down these numbers based on the sex of the listener (i.e., the interlocutor producing the feedback), as provided in the Switchboard corpus. This is shown in Table 4. Note that in the metadata of Switchboard there are only two options for sex, female and male. A chi-square test revealed that sex influences the type of feedback ($\chi^2(9) = 1165.71, p < .001$). Analysis of the standardized residuals ($\alpha = 0.05$) revealed that there were significant differences in most feedback types, as indicated in Table 4. Perhaps most notably, the use of (S)ympathy, and (SS) Strong surprise is much more frequent for females than males. To further investigate whether these effects are also affected by the sex of the interlocutor re-

Function	Total	Tot. %	F %	M %	FF %	FM %	MM %	MF %	Ov %
(C) Continue	39499	51.8	**54.2**	**48.8**	54.7	53.6	48.4	49.3	45.3
(U) Non-understanding	342	0.45	**0.34**	**0.58**	0.31	0.39	0.56	0.60	26.9
(A) Agree	22809	29.9	**26.3**	**34.3**	26.3	26.3	**35.3**	**33.2**	45.3
(D) Disagree	986	1.29	**1.12**	**1.51**	**0.96**	**1.32**	1.56	1.44	36.0
(Y) Yes-response	4101	5.38	**4.93**	**5.93**	5.05	4.77	5.90	5.97	45.1
(N) No-response	787	1.03	1.02	1.05	0.96	1.08	**0.90**	**1.24**	34.6
(S) Sympathy	1775	2.33	**3.19**	**1.25**	3.29	3.06	**1.01**	**1.55**	46.7
(MS) Mild surprise	2325	3.05	3.03	3.07	2.82	**3.31**	3.01	3.13	37.7
(SS) Strong surprise	3023	3.96	**5.03**	**2.64**	4.81	5.31	2.62	2.66	41.8
(Ds) Disapproval	638	0.84	0.81	0.86	0.78	0.85	0.80	0.94	39.8
(O) Other	9671								

Table 4: Distribution of manually annotated and automatically annotated tokens in the Switchboard corpus. Distribution percentages are calculated excluding the Other category. Automatic annotations used Model 19. F=Female, M=Male. FM = Female-to-Male feedback, etc. Bold numbers denote significant deviations from the expected distribution ($\alpha = 0.05$). Ov=Overlap.

ceiving the feedback, we also split these numbers based on the sex of both interlocutors, as can be seen in Table 4. Chi-square tests revealed that there was indeed such an effect, both when the feedback was produced by males ($\chi^2(9) = 46.2, p < .001$) and females ($\chi^2(9) = 34.7, p < .05$). For example, in male-male conversations, there is less use of (S)ympathy, compared to in male-female conversations.

Our analysis also shows that in general, females produce 2.73 feedback tokens per minute, whereas males produce 2.23 feedback tokens per minute. Our findings only reflect observations in the Switchboard corpus and therefore these findings may not be generalizable to other corpora.

5.4 Analysis of overlap

Another example in which this classification can be used is in analyzing whether certain feedback functions overlap more or less with the speech of the interlocutor. In order to determine whether a feedback was overlapping or not, we took the start time of the feedback and searched for that timestamp in the speech of the interlocutor, if that timestamp occurred during or the start of the interlocutor's speech we assigned the feedback as overlapping. If the start time of the feedback occurred during the interlocutor's silence or laughter we assigned the feedback as not overlapping. The percentage of overlap for each feedback type is shown in Table 4.

Using this method, we find that (U) Non-understanding, (D)isagree, (N)o-response, (MS) Mild Surprise, and (Ds) Disapproval tend to not overlap as much with the interlocutor's speech. Listeners may wait to produce a feedback function (U) Non-understanding until the end of the interlocutor's turn in order to first see if they can repair their comprehension of what was said or being said. Listener's may also wait to produce feedback functions with negative connotations such as (D)isagree, (N)o-response, and (Ds) Disapproval, in order to decide whether they should take the turn, or to further respond to what the interlocutor has said.

We had expected feedback functions (C)ontinue, (A)gree, (S)ympathy, (MS) Mild Surprise, and (SS) Strong Surprise to overlap with the speech of the interlocutor. However, we find that for (C)ontinue, (A)gree, (S)ympathy, (SS) Stong Surprise, these feedback functions almost equally overlap and not overlap. Further analysis should be done to see if the silences of the interlocutor's are short breaths or longer pauses. It would be interesting to do an analysis similar to the one done by Goodwin (1986) where they compared assessments and continuers. They found that although assessments and continuers share similar contexts (they are said during the speech of the interlocutor), continuers bridge turn-constructional units of the interlocutor, whereas assessments do not interrupt the subsequent unit of the interlocutor. This type of analysis which takes into consideration conversational units of the interlocutor may give more insight into where exactly these feedback functions occur within the interlocutor's turn, as well as give information whether the feedback functions which occur during the interlocutor's silences are between or within turn-constructional units.

6 Discussion

Although fine-tuning GPT-3 (Model 12) performs the best, it may not be suitable for an online setting or for annotating large corpora. This model is dependent on OpenAI's API which can have downtime and using it can be costly. Model 19 (SimCSE + Lexical), which can fairly accurately predict feedback functions on par with human annotators, can be an option for online settings or for annotating large corpora.

The prosodic features performed poorly, and in the end they did not contribute much to the best-performing models. The best model, using a fine-tuned GPT-3, did not use any prosodic features at all. This is perhaps a bit surprising, since prosody should help to disambiguate feedback tokens which are not easily classifiable given only textual information, such as 'no' when it is used as negative agreement. On the other hand, it might be the case that the preceding context contains redundant information, and could for example help to disambiguate a question (preceding a No-response) vs. a statement (preceding a Disapproval). In any case, future work should explore better prosodic features, using distributed, self-supervised speech representations (Lin et al., 2023). It is also interesting to note that discrete representations of dialog context, such as Dialog Acts and Part-of-speech, performed much more poorly than the distributed representations (SimCSE).

One reason that the Dialog Acts may not have performed well (Model 8) could be due to the preceding context being misclassified with the incorrect dialog act. Therefore, this is an error that was propagated into the model. This propagation of errors can also be said for the probability distributions by GPT-3 (Models 13, 14). For future work, further analysis should be done on how these errors affect the model.

There is also more experimentation that could be done with GPT-3. Without fine-tuning, the probability distributions from zero-shot and few-shot classifications did not perform better than the lexical one-hot encodings. Experimentation with different prompts could improve the GPT-3 features. In future work, we would like to use a separate training set to fine-tune the GPT-3 model so that we can evaluate the probability distributions of the fine-tuned model, and potentially combine them with other features. One potential route could also be to add prosodic information to the prompts by discretizing them. For example, pitch slope could be disctretized by describing it as flat, rising, or falling.

Our classification models have only been trained and evaluated with the Switchboard corpus, it would be interesting to see how our best models perform with other corpora, such as corpora where the interlocutors are speaking face-to-face.

7 Conclusion

In this paper, we proposed different models which can automatically classify 10 communicative feedback functions: continue, non-understanding, agree, disagree, yes/no response, sympathy, mild/strong surprise, and disapproval. We experimented with different combinations of lexical and prosodic features from the feedback utterances, as well as context features from the preceding utterance of the interlocutor as input to a SVM classifier. For contextual features, we investigated the use of probability distributions from the predicted function labels from a zero-shot or few-shot GPT-3 classifier, as well as SimCSE sentence embeddings. Finally, we also compared with a fine-tuned GPT-3 classifier.

Our experiments show that just using lexical features and SimCSE gives a fairly good performance, on par with inter-annotator agreement. While using GPT-3 in a zero-shot or few-shot fashion does not contribute much, a fine-tuned GPT-3 model outperforms all other models, even though no prosodic information is used.

The automatic annotations of the communicative functions of feedback in the Switchboard corpus by Model 19 can be found in this repository: https://github.com/carolfigPhD/FeedbackAnnotationScheme.

Limitations

We are aware that one limitation in terms of reproducibility is that GPT-3 may not return the same labels if the experiments were to be run again. Moreover, GPT-3 is like a black-box, when we fine-tune the model we do not know what exactly is being fine-tuned. Another limitation is accessibility, not everyone will have access to GPT-3 which can be costly and is dependent on the services of OpenAI.

We have also not examined if there are differences in feedback in face-to-face conversations compared to telephone conversations. As mentioned, we have only trained and evaluated our

classifiers with the Switchboard corpus but have not evaluated with a face-to-face corpus.

Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 859588.

References

Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7408–7412.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling feedback in interaction with conversational agents—a review. *Frontiers in Computer Science*, 4.

Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A multimodal model for predicting conversational feedbacks. In *International conference on text, speech, and dialogue*, pages 537–549. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Harry Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Hendrik Buschmeier, Zofia Malisz, Marcin Włodarczak, Stefan Kopp, and Petra Wagner. 2011. 'are you sure you're paying attention?' - 'uh-huh' communicating understanding as a marker of attentiveness. In *Proc. Interspeech 2011*, pages 2057–2060.

Herbert H Clark. 1994. Managing problems in speaking. *Speech communication*, 15(3-4):243–250.

Alfred Dielmann and Steve Renals. 2008. Recognition of dialogue acts in multiparty meetings using a switching dbn. *IEEE transactions on audio, speech, and language processing*, 16(7):1303–1314.

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, and Gabriel Skantze. 2022. Annotation of communicative functions of short feedback tokens in switchboard. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1849–1859.

George Forman and Ira Cohen. 2004. Learning from little: Comparison of classifiers given little training. In *Knowledge Discovery in Databases: PKDD 2004: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, 2004. Proceedings 8*, pages 161–172. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, 9(2):205–217.

Agustín Gravano, Stefan Benus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Proc. Interspeech 2007*, pages 1613–1616.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 131–138.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. Bpm_mt: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3452.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *Proc. Interspeech 2016*, pages 2890–2894.

Divesh Lala, Koji Inoue, Tatsuya Kawahara, and Kei Sawada. 2022. Backchannel generation model for a third party listener agent. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, pages 114–122.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G Ward. 2023. On the utility of self-supervised models for prosody-related tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111. IEEE.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Bhavitvya Malik. 2020. Dialogtag.

Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, Stefan Kopp, and Petra Wagner. 2016. The ALICO corpus: Analysing the active listener. *Language resources and evaluation*, 50(2):411–442.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems*, 20:70–84.

Daniel Neiberg, Giampiero Salvi, and Joakim Gustafson. 2013. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55(3):451–469.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Laurent Prévot, Jan Gorisch, and Roxane Bertrand. 2016. A cup of cofee: A large collection of feedback utterances provided with communicative function annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3180–3185.

Laurent Prévot, Jan Gorisch, and Sankar Mukherjee. 2015. Annotation and classification of french feedback communicative functions. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 298–306.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*, pages 247–258. Springer.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings. In *Proc. Interspeech 2017*, pages 879–883.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:93.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.