Predicting the Lexical Form and Prosody of Feedback for Synthesis

Carol Figueroa[1,2]

[1]Furhat Robotics, Sweden
[2]Aix-Marseille University, France
carol@furhatrobotics.com

Abstract

In human conversations, listeners produce backchannels or feedback such as 'mhm', 'yeah', and 'wow' which convey different communicative functions depending on their prosodic realization and the context in which they are produced. However, many conversational systems or spoken dialogue systems lack the ability to generate feedback with different prosodic realizations or lexical forms. Therefore, the aim of this PhD is to predict the lexical form and prosodic features of feedback in order to synthesize feedback which are appropriate in the conversational context.

Index Terms: feedback, backchannels, prosody, speech synthesis, communicative functions

1. Motivation

There has been a lot of work in trying to incorporate short feedback utterances in conversational systems, however, the focus has been mainly on predicting the timing of when to add them [1, 2, 3, 4, 5, 6, 7, 8], or predicting the function of the feedback [6, 7, 9, 10, 11, 12]. In contrast, work on synthesizing feedback has been limited [12, 13, 14]. Two important dimensions for generating feedback has had less focus: predicting the lexical form [9] and predicting the prosody [15].

Therefore our goal for this PhD is to train a model, see Figure 1, that takes as input the communicative function we would like to convey and the context from the interlocutor and outputs a lexical form and prosodic features of feedback. This model should predict different lexical forms for a specific communicative function. For example, if we would like to generate a feedback that expresses agreement, the model could predict 'yeah', 'mhm', 'exactly', 'absolutely' or 'right'. The prosody should also be appropriate for the conversational context. If the model predicts 'yeah' for agreement, the prosody of the 'yeah' should sound like agreement and not surprise. Previous work has found similarity in pitch between backchannels and the preceding utterance of the interlocutor, this similarity may be why backchannels are unobtrusive [16, 17]. If listener's align (a phenomenon where interlocutors' speaking style converge) their prosodic features in their feedback, then this should also be reflected in our vocal feedback model. By predicting different lexical forms and prosodic features of feedback, we can avoid synthesizing feedback that is monotonous or repetitive. Furthermore, these variations can create a more natural or human-like feedback for conversational systems.

2. Research Questions

1. What are the different communicative functions of feedback?

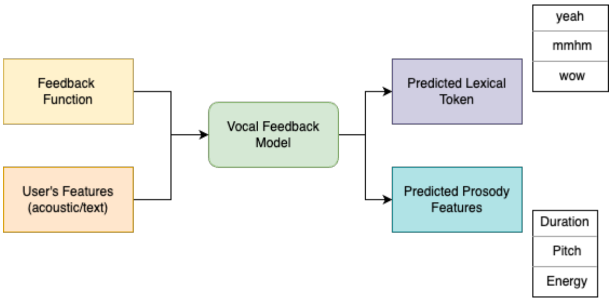

Figure 1: *Proposed vocal feedback model.*

2. Is there prosodic alignment between certain feedback functions and the preceding or following utterance of the interlocutor?
3. Should we represent prosodic features as discrete prosodic representations (e.g. duration, pitch, intensity) or distributed representations (e.g. prosody embeddings, self-supervised speech representations [18])?

3. Thesis Roadmap

3.1. Feedback function annotation scheme

Many existing schemes for annotating communicative functions of feedback are based on the four feedback functions: *Contact*, *Perception*, *Understanding*, and *Attitudinal reactions* proposed by [19]. However, we found that the existing schemes were not nuanced enough. Ultimately, we decided to use function labels from existing schemes but also added labels which we decided on by listening to the feedback in the Switchboard corpus [20]. The Switchboard corpus is an American English corpus which consists of about 2500 telephone conversations. We settled upon 10 functions: *Continue*, *Non-understanding*, *Agree*, *Disagree*, *Yes-response*, *No-response*, *Sympathy*, *Disapproval*, *Mild Surprise*, and *Strong Surprise* [21].

3.2. Automatic annotation of feedback functions

We manually annotated 2179 instances of feedback with their corresponding communicative function. Since this amount of data is not enough to train a deep neural network model for our vocal feedback model and manually annotating each feedback with a function is too time-consuming, we decided to train a classifier to automatically annotate the remaining possible feedback tokens in Switchboard. We experimented with a combination of lexical and prosodic features extracted from the feedback and context features from the $4000\,\mathrm{ms}$ of the preceding utterance of the interlocutor.

1. Feedback features
   - Lexical tokens which are one-hot enconded
   - Prosodic features: duration, mean pitch, pitch slope, pitch range, and mean intensity
2. Context features
   - Part-of-speech (POS) tags of the preceding utterance, we used the top 30 POS bigrams calculated using term frequency-inverse document frequency (TF-IDF)
   - Dialog Act of the preceding utterance
   - SimCSE [22] sentence embedding of the preceding utterance

Given the small amount of training data that we have, we decided to use traditional machine learning models such as Support Vector Machines (SVM) as well as a pre-trained language model (GPT-3) for classification. We also experimented with using the probability distributions of the GPT-3 model as an input feature to the SVM classifier. We achieved good performance with only using SimCSE and lexical features (f1-score 0.72), as well as fine-tuning GPT-3 (f1-score 0.80). These scores were comparable to the inter-annotator agreement (f-score 0.74). Since the prosodic features were not so helpful we concluded that either the context SimCSE embeddings were enough to classify the functions or we need to represent the prosodic features differently. Finally, we used the SVM trained on the SimCSE embeddings and the lexical encodings to classify the remaining possible feedback instances in Switchboard.

3.3. Feedback and alignment

We investigated whether local prosodic alignment exists between the feedback utterances and the 500 ms preceding, as well as the following utterance of the interlocutor for each feedback function in our scheme. We measured local alignment using Pearson's correlation, where we expected correlations between the prosodic features (mean pitch, pitch slope, mean intensity) of the two interlocutors. We found that in terms of intensity, listeners align their *Continue* and *Agree* feedback to the intensity of the preceding utterance of the interlocutor.

3.4. Prosodic Representations

In order to understand what prosodic features to predict and how prosody should be represented, we are planning a first listening test with two methods of synthesis.

1. Using the World vocoder [23]
   - We first get phone level alignments of a feedback in Switchboard (reference) and a target speaker for the same lexical token e.g. 'yeah'.
   - Then we use time-domain pich synchronous overlap (TD-PSOLA) to increase or reduce the duration of the phones in the target voice feedback.
   - We then use the World vocoder to change the energy and the f0 values of the target speaker to be similar to the reference speaker. This is done at the frame level.
2. Using FastPitch [24]
   - FastPitch has prediction models for duration, pitch, and energy which are predicted at the phone level.
   - We can extract the duration, mean pitch, mean energy for each phone in the feedback of the reference speaker and override the FastPitch predictions for the target speaker so that they are similar to the reference.

Participants will be given a clip of the original conversation in Switchboard, where one speaker is talking and the other speaker is giving feedback. They will also be given the same clip except the original feedback will be replaced with either the feedback synthesized by the World vocoder or FastPitch. Participants will then be asked to judge whether the function/intent of the synthesized feedback has changed or remained the same. From this listening test we will be able to conclude whether representing prosody as discrete representations either at the frame or phone level is a good representation.

3.5. Vocal Feedback Model and Synthesizing Feedback

Once we know whether we should predict discrete or distributed representations of prosody we can then begin training our vocal feedback model. If we use discrete prosodic representations such as duration, mean pitch, and mean energy per phone we can replace FastPitch's predictions with our predictions. If we use distributed representations of prosody such as prosody embeddings we can use this to condition the melspectograms in FastPitch. Finally, we will perform a final listening test where we will evaluate the prosodic and lexical form predictions of our vocal feedback model.

4. Key Challenges

- Finding a face-to-face corpus for American English which was not task-oriented was difficult. We decided to use the Switchboard corpus despite that the audio quality is not the best and even though the conversations are not face-to-face they come do come close to face-to-face conversations.
- Since FastPitch is trained on long utterances, it has a hard time synthesizing short feedback tokens. We will need to fine-tune FastPitch with examples of short feedback utterances in order to improve the quality of synthesis.
- Evaluating the synthesized feedback will also be a challenge. Using an objective evaluation such as mean opinion score (MOS) is not suitable for our task since we want to synthesize variations of prosody and lexical form which might differ from the reference sample.

5. Main Contributions

- We have proposed an annotation scheme for different communicative function of feedback which have more attitudinal information.
- The results of our investigation whether certain feedback functions align to the preceding utterance of the interlocutor can be used to inform alignment models.
- Through our synthesis experiments we will be able to determine what type of prosodic representations are needed for synthesizing feedback.
- If our vocal feedback model is incorporated into spoken dialogue systems it will make human-machine interactions more natural.

6. Acknowledgements

I'd like to thank my supervisors Gabriel Skantze and Magalie Ochs for their guidance and support. This work is being funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 859588.

7. References

[1] N. Ward and W. Tsukahara, "Prosodic features which cue backchannel responses in english and japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[2] R. Ruede, M. Müller, S. Stüker, and A. Waibel, "Enhancing Backchannel Prediction Using Word Embeddings," in *Proc. Interspeech 2017*, 2017, pp. 879–883.

[3] R. Ruede, M. Müller, S. Stüker, and A. Waibel, "Yeah, right, uh-huh: a deep learning backchannel predictor," in *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*. Springer, 2019, pp. 247–258.

[4] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous agents and multi-agent systems*, vol. 20, pp. 70–84, 2010.

[5] R. Ishii, X. Ren, M. Muszynski, and L.-P. Morency, "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*, 2021, pp. 131–138.

[6] A. Boudin, R. Bertrand, S. Rauzy, M. Ochs, and P. Blache, "A multimodal model for predicting conversational feedbacks," in *International conference on text, speech, and dialogue*. Springer, 2021, pp. 537–549.

[7] A. I. Adiba, T. Homma, and T. Miyoshi, "Towards immediate backchannel generation using attention-based early prediction model," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7408–7412.

[8] A. Axelsson, H. Buschmeier, and G. Skantze, "Modeling feedback in interaction with conversational agents—a review," *Frontiers in Computer Science*, vol. 4, 2022.

[9] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. Ward, "Prediction and Generation of Backchannel Form for Attentive Listening Systems," in *Proc. Interspeech 2016*, 2016, pp. 2890–2894.

[10] D. Ortega, C.-Y. Li, and N. T. Vu, "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)*. IEEE, 2020, pp. 8064–8068.

[11] J. Y. Jang, S. Kim, M. Jung, S. Shin, and G. Gweon, "Bpm_mt: Enhanced backchannel prediction model using multi-task learning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3447–3452.

[12] D. Lala, K. Inoue, T. Kawahara, and K. Sawada, "Backchannel generation model for a third party listener agent," in *Proceedings of the 10th International Conference on Human-Agent Interaction*, 2022, pp. 114–122.

[13] T. Stocksmeier, S. Kopp, and D. Gibbon, "Synthesis of prosodic attitudinal variants in german backchannel ja." in *INTERSPEECH*, 2007, pp. 1290–1293.

[14] S. C. Pammi, "Synthesis of listener vocalizations: towards interactive speech synthesis," 2011.

[15] A. Nath and N. G. Ward, "On the predictability of the prosody of dialog markers from the prosody of the local context," *Proc. Speech Prosody 2022*, pp. 664–668, 2022.

[16] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proc. Interspeech 2010*, 2010, pp. 3054–3057.

[17] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *2015 AAAI Spring Symposium Series*, 2015.

[18] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1104–1111.

[19] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of semantics*, vol. 9, no. 1, pp. 1–26, 1992.

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

[21] C. Figueroa, A. Adigwe, M. Ochs, and G. Skantze, "Annotation of communicative functions of short feedback tokens in switchboard," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1849–1859.

[22] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021.

[23] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[24] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.