Aix*Marseille université
Socialement engagée

NNT : 2023AIXM0001

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 16 décembre 2024 par

Carol Figueroa

Not the “when” and “where” but the “how” : modeling and generating spoken feedback for conversational agents

Discipline Informatique

École doctorale
ED 184 Mathématiques et Informatique

Laboratoire/Partenaires de recherche
Laboratoire d’Informatique & des Systèmes (LIS)
Furhat Robotics AB

Composition du jury

Chloé Clavel Directrice de recherche, INRIA Paris	Rapporteure
Nigel Ward Professeur, The University of Texas at El Paso	Rapporteur
Catherine Pelachaud Directrice de recherche, CNRS	Examinatrice
Joakim Gustafsson Professeur, KTH Royal Institute of Technology	Examinateur
Frédéric Béchet Professeur, Aix-Marseille Université	Président du jury
Magalie Ochs Associate Professeur, Aix-Marseille Université	Directrice de thèse
Gabriel Skantze Professeur, KTH Royal Institute of Technology	Co-Directeur de thèse

LIS

Affidavit

I, undersigned, Carol Figueroa, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Magalie Ochs and Gabriel Skantze, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the french national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Marseille, France, October 21, 2024



This work is licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Public License

List of publications and participation in conferences

List of publications produced in the framework of the thesis project :

1. Figueroa, Carol, Adaeze Adigwe, Magalie Ochs, and Gabriel Skantze. “Annotation of Communicative Functions of Short Feedback Tokens in Switchboard." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022
2. Figueroa, Carol, Štefan Beňuš, and Gabriel Skantze. “Prosodic Alignment in Different Conversational Feedback Functions." In *20th International Congress of Phonetic Sciences*. 2023
3. Figueroa, Carol, Magalie Ochs, and Gabriel Skantze. “Classification of Feedback Functions in Spoken Dialog Using Large Language Models and Prosodic Features." In *27th Workshop on the Semantics and Pragmatics of Dialogue*. 2023
4. Figueroa, Carol, Marcel de Korte, Magalie Ochs, Gabriel Skantze. “Mhm... Yeah? Okay! Evaluating the Naturalness and Communicative Function of Synthesized Feedback Responses in Spoken Dialogue.” In *25th Meeting of the Special Interest Group on Discourse and Dialogue*. 2024

Participation in conferences and summer schools during the thesis period :

1. Speech Processing Courses in Crete : Inclusive Neural Speech Synthesis 2021
2. The 12th Lisbon Machine Learning School 2022
3. The 13th Language Resources and Evaluation Conference 2022
4. The 20th International Congress of Phonetics Sciences 2023
5. The 27th Workshop on the Semantics and Pragmatics of Dialogue 2023
6. The 25th Meeting of the Special Interest Group on Discourse and Dialogue 2024

Résumé et mots clés

Durant une conversation, l'interlocuteur produit souvent de nombreux feedbacks, i.e. courtes réponses telles que *mhm*, *oui*, et *wow*. Suivant leur forme lexicale et leur prosodie, ces feedbacks peuvent véhiculer des fonctions communicatives différentes comme l'*accord*, le *désaccord* ou encore la *surprise*. Les agents conversationnels comme Alexa, Siri, et GPT-4o, de plus en plus répandus, n'ont souvent pas la capacité de générer des feedbacks nuancés permettant d'exprimer différentes fonctions communicatives. La plupart des recherches sur la modélisation des feedbacks se sont concentrées sur la prédiction de quand un agent peut produire un feedback ou sur la fonction communicative qu'elle doit transmettre, tandis que peu d'attention est accordée à la prédiction de la forme lexicale ou de la réalisation prosodique du feedback.

Cette thèse se concentre sur la modélisation et la génération de feedbacks vocaux pour les agents conversationnels. Plus précisément, l'objectif est la synthétise des feedbacks avec une prosodie appropriée au contexte. Nous proposons un modèle de feedbacks vocaux qui prédit à la fois la forme lexicale et les caractéristiques prosodiques des feedbacks, en fonction de la fonction communicative que l'agent conversationnel doit transmettre et du contexte conversationnel.

Pour ce faire, nous commençons par identifier les fonctions communicatives des feedbacks dans les conversations entre humains, ce qui aboutit à un schéma de codage. Nous analysons ensuite la distribution des formes lexicales pour chaque fonction et réalisons une analyse prosodique afin de déterminer si différentes fonctions présentent des caractéristiques prosodiques distinctes. Pour accélérer le processus d'annotation, nous entraînons un classificateur en utilisant nos annotations manuelles. Nous expérimentons diverses combinaisons de caractéristiques d'entrée, incluant des distributions de probabilité provenant de GPT-3, des caractéristiques lexicales et prosodiques des feedbacks, ainsi que des caractéristiques contextuelles de l'interlocuteur.

Les feedbacks n'interrompent pas la conversation. Certaines recherches suggèrent que ces réponses sont discrètes en raison de l'*alignement prosodique*, un phénomène où la prosodie de l'interlocuteur devient similaire à celle du locuteur. Dans cette thèse, nous étudions si les interlocuteurs alignent leurs caractéristiques prosodiques avec celles des énoncés précédents de l'interlocuteur, pour toutes les fonctions de notre schéma de codage. Enfin, nous transposons la prosodie des feedbacks des conversations entre humains vers un locuteur cible en explorant différentes méthodes de synthèse. Nous réalisons un test d'écoute pour évaluer la crédibilité, l'adéquation de la prosodie, ainsi que la préservation de la fonction communicative des feedbacks synthétisés.

Mots clés : feedbacks, backchannels, fonctions communicatives, alignement prosodique, modélisation de la prosodie, synthèse vocale, agents conversationnels, interaction humain-machine

Abstract and keywords

In conversations, listeners often produce short feedback responses such as *mhm*, *yeah*, and *wow*. Through their lexical form and prosody, these responses convey important communicative functions like *agreement*, *disagreement*, and *surprise*. While conversational agents like Alexa, Siri, and GPT-4o are becoming more prevalent, they often lack the ability to generate these nuanced feedback responses. Most research on modeling feedback responses has focused on predicting whether the agent should produce a feedback response or what communicative function it should convey, with little attention given to predicting the lexical form or prosodic realization.

This thesis focuses on modeling and generating spoken feedback responses for conversational agents. Specifically, it investigates what is required to synthesize feedback responses with appropriate prosody for their context. This includes examining how these short feedback responses should sound—their prosodic realization and representation. Moreover, we propose components for a vocal feedback model that predicts both the lexical form and the prosodic features of feedback responses, based on the communicative function the conversational agent should convey and the context of the interlocutor.

We begin by identifying the communicative functions of feedback responses in human-human conversations, resulting in a coding scheme. We then analyze the distribution of lexical forms for each function and perform prosodic analysis to determine whether different functions exhibit distinct prosodic features. To expedite the annotation process, we train a classifier using our manual annotations. We experiment with various combinations of input features, including probability distributions from GPT-3, lexical and prosodic features of the feedback responses, and contextual features of the interlocutor.

A key characteristic of feedback responses is that they do not interrupt the conversation. Some evidence suggests that feedback responses are unobtrusive due to *prosodic alignment*, a phenomenon where a listener's prosody becomes similar to that of their interlocutor. We investigate whether listeners align their prosodic features with the prosody of their interlocutor's preceding and following utterances across all functions in our coding scheme. Lastly, we transplant the prosody of feedback responses from human-human conversations to a target speaker using two synthesis methods: one transplants the prosody at the frame-level, and the other at the phone-level. We conduct a listening test to evaluate the naturalness, appropriateness of the prosody, and preservation of the communicative function of the synthesized feedback responses.

Keywords: feedback responses, backchannels, communicative functions, prosodic alignment, prosody modeling, speech synthesis, conversational agents, human-computer

interaction

Dedication

Esta tesis está dedicada a mi mamá Laura Adame y mi abuelita Teodora Adame, las mujeres más fuertes que conozco. And to my mischievous dog, Panda Adame, whom I was so lucky to have in my life. En eveneens aan opa Piet de Korte, die ik elke dag nog mis.

Acknowledgements

I would like to begin by thanking Esther Klabbers-Judd. If Esther had not posted about this position on LinkedIn and encouraged me to apply, I'm not sure if this journey would have happened. Esther and I met at the Workshop for Young Female Researchers in Speech Science and Technology in 2018, where she was my assigned mentor. That was six years ago, yet she has continued to mentor me throughout the years. I am immensely grateful for her feedback on this thesis. Esther, thank you for taking me under your wing, for advocating for me, and for your help throughout these past years during my PhD.

This journey would also have not been possible without the references from Pierre Lanchantin and Simon King. Pierre, thank you for being such an awesome boss and for inspiring me to pursue a PhD. Simon, thank you for your continued mentorship, you've always given your former SLP students the feeling that we could reach out to you for advice, even after all these years. In addition, I'm grateful for your advice to keep an eye out for a Marie Curie PhD, which gave me the confidence to apply for this PhD.

I would also like to extend my gratitude to my academic advisor, Magalie Ochs, for all the feedback she has given to me throughout these three and a half years. Navigating a PhD from a different country and in a different language has been challenging, and I'm grateful for your help with managing the administrative hurdles in French. Although external circumstances didn't allow me to stay in Marseille for a longer time, I appreciated feeling welcomed during the time that I was there. Last but not least, I would especially like to thank you for helping to secure the necessary finances for the last publication and the expenses of the defense.

This thesis would not have been possible without the guidance from my industrial advisor Gabriel Skantze. The COBRA topic for this PhD was initially quite broad, meaning it needed to be narrowed down. I'm very appreciative of how in one of our first meetings to discuss what I could potentially focus on, Gabriel took into account my interests and opinions. I think this example accurately reflects the type of advisor that he is. Gabriel treats his PhD students as fellow collaborators and for this I am very grateful. For the past three years, Gabriel has always gone out of his way to make time to discuss my progress, even when time was limited, such as when he joined Adaeze and I on a video call a few hours before our first paper deadline at LREC. I'm deeply thankful to Gabriel for always having a conference in mind for where to submit our work, which is just one example of him providing the structure that I desperately needed. Gabriel, thank you for your guidance throughout, and for your patience when dealing with my tortoise-like speed of work progress.

I would also like to extend my gratitude to Catherine Pelachaud and Joakim Gustafs-

son, who are the members of my comité du suivi. Presenting my work to you every year to discuss my progress and receiving your feedback has been incredibly helpful to me in navigating this PhD.

As part of being in this COBRA consortium, I have had the opportunity to be a part of Furhat Robotics. I would like to thank all the wonderful people who I've met during my time at Furhat: Andy JOhnston, Jonas Beskow, Samer Al Moubayed, Dennis Korneev, Nicholas Pugin, Louis Simeonidis, Charlie Caper, Mathieu Rey-Herme, Subramanian Alagappan, Niclas Drevinger, Filip de Figueiredo, Jonathan Ahlbom, Chris Wood, Quinten Stekelenburg, Fabienne Cap, Susanna Dillenbeck, Roger Nimbratt, Pål Gylling Linder, Nils Hagberg, Edvin Eminovic, Kaspar Kööp, Johan Ehrenfors, Susanne Leiger, Richard Hede, Anne Jansen, Ken Vu, Arnaud Henneville-Wedholm, Morgan Bell, Jacob Korniak, Myriam Tay, Christina Egerstrom, Linn Mann, and Princess Pandili. I would especially like to express my gratitude to a couple of people. Susanne, thank you so much for onboarding me and for helping me settle into both Furhat and Sweden, and for all your help with all the administrative matters. Similarly, Princess, thank you for all your help in navigating the administrative hurdles between COBRA and Furhat. Kaspar, thank you for being such a great manager. I really appreciated that you checked in on how Marcel and I were settling in, your many suggestions of what to do in this great city, and your golden nuggets of knowledge on what to do and what not do in the PhD. Morgan, you were also such a great manager. Thank you for checking in on how I was doing, not only in my PhD progress but also in life. I really appreciated that you always asked if you could help in any way. Also, thank you for participating in my pre-pilot for my last study, it helped me a great deal. Mathieu, Linn, and Christina, I am so thankful for your friendship and I can't wait to spend more time with you. Johan, you were always curious about what I was working on and that meant a lot to me because it made me feel less alone. You were my first friend in Sweden, and I truly appreciate how you made Marcel and I feel welcomed.

Although I was at KTH for a limited amount of time, the people there made me feel welcome. Thank you Anna Klezovich, Harm Lameris, Jim O'regan, Shivam Mehta, Siyang Wang, Bram Willemsen, Jens Edlund, Christina Tånnander, Gustav Henter, Alireza Kamelabad, Katya Torubarova, Bahar Irfan, Haotian Qi, Livia Qian, and Éva Székely. Thank you Alireza and Katya for always being friendly and for inviting Marcel and I into your home. Bahar, thank you for all your advice throughout these past few months, I feel very lucky to have gotten to know you. Haotian and Livia, thank you both for making me feel welcome in your room; it was great experiencing what it would have been like to work with lab mates. Livia, I wish I could have met you earlier, I look forward to our future foodie adventures. I would like to give my sincerest thanks to Éva Székely. Éva, thank you for making me feel welcomed at KTH and at your home, but also for always offering help with anything related to TTS and with the feedback on this thesis. Thank you for encouraging me to submit the last paper to SigDial, which helped a great deal in making sure that it could still be published.

As part of the COBRA consortium, I was also lucky to have my secondment at the Slovak Academy of Sciences for three months. Thank you Štefan Beňuš, Róbert Sabo, Sakhia Darjaa, Marián Trnka, Milan Rusko, Joanna Kruyt, and Jay Kejriwal for making

me feel welcomed. Joanna and Jay, it was a great pleasure to be your lab mate for the three months I was there, and experiencing for the first time what it is like to have lab mates. It was a great experience to be able to discuss our work, share ideas, chit-chat but also to silently work on our own thing, but feel that we had each other for support. I would especially like to thank my co-supervisor Štefan Beňuš. Štefan went above and beyond to make Byron and I welcomed in Slovakia. I feel very lucky to have been supervised by you and wished there had been more opportunities to collaborate. Štefan, thank you for your guidance in the alignment paper, but also for the life advice that you have given me.

I would like to give special thanks to some of our senior COBRA members: Esther Klabbers-Judd, Christine Mooshammer, Susanne Fuchs, Elisabeth Degand, and Štefan Beňuš, for always checking in on the ESRs and asking about our well-being throughout the PhD. Thank you for making us feel supported.

The COBRA consortium was made up of fifteen ESRs: Lena-Marie Huttner, Dorina de Jong, Emilia Kerr, Salomé Antoine, Tom Offrede, Greta Gandolfi, Mercedes Villalobos Cardozo, Junfei Hu, Johannah O'Mahony, Zheng (Byron) Yuan, Joanna Kruyt, Jay Kejriwal, Adaeze Adigwe and Chinmaya Mishra. I want to express many thanks to you all for making the consortium a big support network. Although we were in different countries, every time we met for the COBRA events, it felt like seeing old friends. Thank you Chinmaya, Tom, Byron, Lena and Adaeze for helping me by chatting with me for an hour so I could collect data for my conversational corpus. Johannah O'Mahony, thank you for being a great friend, for checking in on me and for giving me your support. Mercedes, thank you for cheering me on, making sure that I was alive and for sending me photos of your cat Maki, which have really cheered up my day. Lena, thank you so much for your cat photos of Nermal, and for your feedback on this thesis, especially for the alignment chapter and the background section, and for helping me navigate the ADUM system. I would like to give my deepest thanks to Adaeze. For the past three years, Adaeze and I have made time to meet online on a weekly basis to discuss our work, progress, share ideas and to check in on each other. I may not have had a lab mate most of the time but you were definitely a wonderful virtual lab mate. Adaeze, you are one of my closest friends, my confidant, and your support has meant everything to me.

I would also like to thank my whole family, my tíos and tías, my primos and primas, my brothers and sisters—Roxana, Noe, Daniel, Iris and James: thank you for cheering me on from the other side of the world. I would also like to thank my husband's family: Joke de Korte, Remco de Korte, Lena de Korte, Edith de Korte, Erno de Korte, Rémon de Korte, and Fatimah Augsburger for cheering me on from this side of the world.

I would like to thank my mom Laura Adame and my grandma Teodora Adame for their support and for their love. Although my grandma was sad that I moved away from the US again, she has always made it known how much she loves me and how proud she is of me. I would like to thank my mom for always supporting me in my dreams, and for giving me the freedom to pursue them. Thank you for checking in on me every week and for making me feel part of the family even when I was far away.

Lastly, I would like to thank my husband Marcel de Korte. Words cannot express how

thankful I am to have had your support throughout this PhD. I would first like to thank you for encouraging me to apply for this PhD position. I truly could not have survived this PhD journey without your support and continuous encouragement, especially during times when I doubted myself. During busy times and when deadlines were approaching, you have always made sure that there was something nice for me to eat, and that everything in the house was taken care of so that I could focus on my work. You have read my papers, gave comments and suggestions, and helped me by reading this thesis from start to finish. You have been so selfless in order to help me succeed. In the end, we were even able to have a joint publication, thank you for helping me set up FastPitch. Marcel, thank you for unconditional love and for bringing so much joy throughout this PhD journey.

Contents

Affidavit 2

List of publications and participation in conferences 3

Résumé et mots clés 4

Abstract and keywords 6

Dedication 8

Acknowledgements 9

Contents 13

List of Figures 17

List of Tables 20

Introduction 24
- 0.1. Why are feedback responses important for human-human communication? . . . 25
- 0.2. Differences between human-human communication and human-conversational agent communication . . . 27
- 0.3. Motivation . . . 29
- 0.4. U.S. English human-human speech corpora . . . 31
- 0.5. Research questions . . . 33
- 0.6. Thesis Contributions and Outline . . . 38

1. Background 40
- 1.1. Feedback, backchannels, continuers, or listener responses? . . . 40
- 1.2. Communicative functions of feedback responses . . . 44
  - 1.2.1. Schemes based on Allwood, Nivre, et al. (1992)'s communicative functions . . . 45
  - 1.2.2. Schemes based on generic/specific distinction . . . 50
- 1.3. Relationship between prosody and communicative functions of feedback responses . . . 54
  - 1.3.1. Prosodic description or prosodic analysis studies . . . 54
  - 1.3.2. Perception studies . . . 56
  - 1.3.3. Classification studies . . . 61

1.4. Alignment and feedback responses . . . 68
1.4.1. What is alignment? . . . 69
1.4.2. Measuring alignment in speech . . . 71
1.4.3. Previous studies . . . 75
1.5. Modeling feedback responses in conversational agents . . . 83
1.5.1. Modeling vocal and visual feedback responses . . . 84
1.5.2. Predicting the timing of feedback responses . . . 85
1.5.3. Predicting the lexical form of feedback responses . . . 87
1.5.4. Predicting the communicative function of feedback responses . . . 88
1.5.5. Predicting the prosody of feedback responses . . . 93
1.5.6. Synthesis of feedback responses . . . 94

2. Annotation scheme for communicative functions of feedback 100
2.1. Introduction and Motivation . . . 100
2.2. The Switchboard corpus . . . 102
2.3. Previous annotations of backchannels in the Switchboard corpus . . . 102
2.4. Finding feedback responses and clip creation . . . 103
2.5. Proposed annotation scheme . . . 104
2.6. Results . . . 109
2.6.1. Inter-annotator agreement . . . 109
2.6.2. Lexical distribution . . . 111
2.6.3. Prosodic Analysis . . . 113
2.7. Discussion . . . 116
2.8. Summary . . . 117

3. Automatic classification of feedback functions 118
3.1. Introduction and Motivation . . . 118
3.2. Related Work . . . 120
3.3. Manual annotations of feedback responses . . . 122
3.4. Input features . . . 122
3.4.1. Feedback features . . . 123
3.4.2. Context features . . . 124
3.5. GPT-3 . . . 124
3.6. Classifiers . . . 128
3.7. Results . . . 128
3.7.1. Performance of classifiers . . . 129
3.7.2. Labeling remaining potential feedback responses . . . 131
3.7.3. Investigating sex differences . . . 133
3.7.4. Analysis of Overlap . . . 134
3.8. Discussion . . . 135
3.9. Summary . . . 136

4. Prosodic alignment in different feedback functions 138
4.1. Introduction and Motivation . . . 138

4.2. Annotations used for measuring alignment . . . 140
4.3. Extraction of prosodic features . . . 141
4.4. Measuring local alignment . . . 142
4.5. Results . . . 143
4.5.1. Mean Pitch . . . 144
4.5.2. Mean Intensity . . . 146
4.5.3. Pitch Slope . . . 148
4.6. Discussion . . . 148
4.7. Summary . . . 149

5. Prosodic Representation of feedback responses for synthesis 151
5.1. Introduction and Motivation . . . 151
5.2. Synthesis methods . . . 154
5.2.1. Signal processing approach . . . 154
5.2.2. Text-to-speech approach . . . 155
5.3. Experimental design . . . 157
5.3.1. Participants . . . 157
5.3.2. Stimuli . . . 157
5.3.3. Participants' tasks . . . 158
5.3.4. Statistical analysis . . . 158
5.4. Results . . . 158
5.4.1. Naturalness ratings . . . 158
5.4.2. Appropriateness ratings . . . 159
5.4.3. Perception of communicative functions . . . 160
5.5. Discussion . . . 161
5.6. Summary . . . 162

Conclusion 163

Bibliography 171

APPENDICES 194

A. Compilation lists of lexical forms of feedback responses in literature 195

B. Language and Corpus type in modeling feedback responses 199

C. Summary of evaluation metrics used for modeling feedback responses 202

D. Communicative functions and definitions in predicting the functions of feedback responses 204

E. Tukey pair-wise tests 208

F. Top 30 part-of-speech bigrams **210**

G. DialogTag examples **211**

H. Confusion matrices for feedback function classification **213**
H.1. Lexical feature 213
H.2. Prosodic features 214
H.3. Context features 216
H.4. GPT-3 218
H.5. Combinations 220

I. Pearson's correlation coefficients **222**
I.1. Mean Pitch Results 222
I.2. Mean Intensity 224
I.3. Pitch Slope 225

J. Switchboard clip transcriptions **227**
J.1. Clip transcriptions 227
J.1.1. Set 1 227
J.1.2. Set 2 228

K. Screenshots of listening test **231**

L. Confusion Matrices of intra-annotator agreement **235**

List of Figures

1. Spectrograms of various *yeah* feedback responses conveying different communicative functions. . . . 26
2. Examples of conversational agents. . . . 28
3. Proposed components for a vocal feedback model. . . . 33
4. Proposed two step process for modeling and generating vocal feedback responses. . . . 34
5. Component from the vocal feedback model that are addressed by R1. 35
6. Components from the vocal feedback model that addressed by R2. . . 35
7. Components from the vocal feedback model that addressed by by R3. 36
8. Components from the vocal feedback model that are addressed by R4. 37
9. Components from the vocal feedback model that are addressed by R5. 37

1.1. Summary of annotation schemes for annotating feedback responses. . 53
1.2. Affect bursts spectrum by Schröder (2003). . . . 57
1.3. Neiberg and Gustafson (2010)'s illustration of three timing features used in classifying Swedish grunts. . . . 66
1.4. Pickering et al. (2004)'s diagram of the coupled production and comprehension processes according to the interactive-alignment model. . . . 70
1.5. Edlund, Heldner, and Hirschberg (2009)'s illustration of convergence and synchrony. . . . 71
1.6. R. Levitan and Hirschberg (2011)'s illustration of proximity, convergence, and synchrony. . . . 72
1.7. Wynn et al. (2022)'s proposed framework for classifying alignment types. 73
1.8. Wynn et al. (2022)'s illustration of high/low proximity/synchrony combinations. . . . 74
1.9. Wynn et al. (2022)'s illustration of all eight types of alignment. . . . 75
1.10. Heldner, Edlund, and Hirschberg (2010)'s visualization of comparison of three transition types BC, S+P1, X2. . . . 77
1.11. Heldner, Edlund, and Hirschberg (2010)'s visualization of comparison of backchannels and affirmative words. . . . 78
1.12. R. Levitan, Gravano, et al. (2011)'s visualization of comparison of conversational pairs and non-conversational pairs. . . . 79
1.13. R. Levitan, Gravano, et al. (2011)'s visualization of comparison between mean differences in the first half of the conversation vs. those in the second half of the conversation. . . . 79
1.14. Åsa Wallers et al. (2006)'s stimuli for Swedish feedback responses *m* and *a*. 96

1.15. Stocksmeier et al. (2007)'s pitch curves for the German feedback response *ja*. 97

2.1. Components from the vocal feedback model that are addressed by R1 and R2. 101
2.2. Labels in Valence-Arousal Space. 108
2.3. Feedback token labels distribution across annotators. 109
2.4. Normalized Confusion matrix between the annotators. 111
2.5. Distribution of lexical forms per communicative function. 113

3.1. Components from the vocal feedback model addressed by R3. 119
3.2. Visualization of the classifier that classifies the communicative function of feedback responses. 120
3.3. Confusion matrix for Model 19. 131
3.4. Confusion matrix for Model 20. 132
3.5. Confusion matrix for Model 12. 132

4.1. Components from the vocal feedback model addressed by R4. 139
4.2. The three correlations computed. 143
4.3. Pearson's correlation coefficients between the mean pitch of feedback functions and preceding/following 500 ms of the interlocutor. 144
4.4. Pearson's correlation coefficients between the mean intensity of feedback functions and preceding/following 500 ms of the interlocutor. . . 146

5.1. Components from the vocal feedback model addressed by R5. 152
5.2. Visualization of the lexical form and prosodic features of the feedback response serving as input to a text-to-speech system such as FastPitch. 154
5.3. Naturalness rating distribution per condition. 159
5.4. Appropriateness rating distribution per condition. 160

H.1. Confusion matrix for Model 1. Features = lexical. 213
H.2. Confusion matrix for Model 2. Features = duration. 214
H.3. Confusion matrix for Model 3. Features = mean pitch. 214
H.4. Confusion matrix for Model 4. Features = pitch slope. 215
H.5. Confusion matrix for Model 5. Features = pitch range. 215
H.6. Confusion matrix for Model 6. Features = mean intensity. 216
H.7. Confusion matrix for Model 7. Features = SimCSE embeddings. 216
H.8. Confusion matrix for Model 8. Features = dialog act. 217
H.9. Confusion matrix for Model 9. Features = part-of-speech tags. 217
H.10. Confusion matrix for Model 10. GPT-3 used as the main classifier with zero-shot approach. 218
H.11. Confusion matrix for Model 11. GPT-3 used as the main classifier with few-shot approach. 218
H.12. Confusion matrix for Model 13. Features = probability distributions from GPT-3 using zero-shot approach. 219

H.13 Confusion matrix for Model 14. Features = probability distributions from GPT-3 using few-shot approach. 219
H.14 Confusion matrix for Model 15. Features = duration, mean pitch, pitch slope, pitch range, mean intensity. 220
H.15 Confusion matrix for Model 16. Features = lexical, duration, mean pitch, pitch slope, pitch range, mean intensity. 220
H.16 Confusion matrix for Model 17. Features = lexical, probability distributions from GPT-3 using zero-shot approach. 221
H.17 Confusion matrix for Model 18. Features = lexical, probability distributions from GPT-3 using few-shot approach. 221

K.1. Screenshot of the clip presentation and the list of communicative functions with their corresponding descriptions. 231
K.2. Screenshot of the question asking the participant to select the communicative function. 232
K.3. Screenshot of the question asking the participant to rate the naturalness of the feedback response on a Likert scale. 233
K.4. Screenshot of the question asking the participant to rate the appropriateness of the prosody of the feedback response on a Likert scale. . . . 234

L.1. Participants' perception of the original Switchboard feedback (Original) compared to feedback synthesized by signal processing approach (SignalP). 235
L.2. Participants' perception of the original Switchboard feedback (Original) compared to feedback synthesized by TTS approach (TTS). 236
L.3. Participants' perception of the original Switchboard feedback (Original) compared to the monotone (Mon) feedback. 236

List of Tables

1. The levels of joint actions by the speaker and addressee. 24

1.1. List of terms used to refer to feedback responses. 42
1.2. List of communicative functions proposed in the literature of feedback responses. 46
1.3. The four basic communicative functions proposed by Allwood, Nivre, et al. (1992). 47
1.4. The four levels of joint action by H. H. Clark (1994). 47
1.5. MUMIN scheme by Allwood, Cerrato, et al. (2007). 48
1.6. Buschmeier et al. (2011)'s annotation scheme. 49
1.7. Scheme for annotating the discourse functions of affirmative words by Gravano, Beňuš, Hirschberg, et al. (2007). 50
1.8. Scheme for annotating the discourse functions of Slovak affirmative words by Beňuš (2016). 51
1.9. Prévot, Gorish, et al. (2015)'s annotation scheme. 52
1.10. Boudin et al. (2021)'s annotation scheme. 53
1.11. Ward (2004) and Ward (2006)'s mapping of prosodic features and their meaning. 54
1.12. Lai (2008)'s example of *really* functioning as a question. 62
1.13. Lai (2008)'s example of *really* functioning as a backchannel. 62
1.14. Lai (2008)'s example of *really* functioning as a question and backchannel. 62
1.15. Heylen et al. (2007)'s features for classifying *yeah*. 63
1.16. Truong and Heylen (2010)'s prosodic features for classifying *yeah*. . . . 64
1.17. Gravano, Beňuš, Hirschberg, et al. (2007)'s features for classifying affirmative words. 65
1.18. Prévot, Gorish, et al. (2015)'s features for classifying communicative functions of French feedback responses. 68
1.19. Example of alignment. 71
1.20. Example of misalignment. 71
1.21. List of previous studies investigating prosodic alignment between feedback responses and the preceding utterance of the interlocutor and corresponding alignment type they measured. 76
1.22. List of acoustic-prosodic features used in R. Levitan, Gravano, et al. (2011)'s alignment study. 78
1.23. List of prosodic features used in Ochi et al. (2024)'s alignment study. . . 83
1.24. List of publications predicting the timing of vocal feedback responses. 86
1.25. List of publications predicting the lexical form of feedback responses. 89

1.26. List of publications from 2020-2022 predicting the communicative functions of feedback responses. . . . 90
1.27. List of publications from 2023 predicting the communicative functions of feedback responses. . . . 91
1.28. List of publications from 2024 predicting the communicative functions of feedback responses. . . . 92
1.29. List of publications predicting the prosody of feedback responses. . . . 95
1.30. Åsa Wallers et al. (2006)'s perceived meanings of Swedish feedback responses *m* and *a*. . . . 96
1.31. Pammi, Schröder, Charfuelan, et al. (2010)'s MaryXML values for feedback responses. . . . 98

2.1. Labels and function descriptions in our proposed annotation scheme. 104
2.2. Example of incipient speakership lexical token by Jurafsky et al. (1998). 105
2.3. Example of the two tracks of actions by H. H. Clark (1996). . . . 106
2.4. Joint action of the addressee at each grounding level by H. H. Clark 1994 and corresponding feedback communicative function by Allwood, Nivre, et al. 1992. . . . 106
2.5. Labels in Grounding Framework. . . . 107
2.6. Inter-Annotator Percentage Agreement Score. . . . 110
2.7. Distribution of lexical forms and corresponding count per communicative function. . . . 112
2.8. Summary of duration, mean pitch, pitch slope, pitch range, and the count of annotations per communicative function. . . . 114

3.1. Feedback communicative functions, count of manually annotated data, and corresponding labels in GPT-3 prompt. . . . 123
3.2. Zero-shot approach: Prompt given to GPT-3. . . . 125
3.3. Few shot approach: Prompt given to GPT-3. . . . 126
3.4. Example of the columns of a csv for fine-tuning GPT-3. . . . 127
3.5. F1 weighted scores for different feature sets in our communicative function classification task. . . . 130
3.6. Distribution of manually annotated and automatically annotated feedback responses in the Switchboard corpus. . . . 133

4.1. Feedback communicative functions and count of manually annotated data used in alignment study. . . . 141

5.1. Cohen's kappa coefficient scores per comparison of intra-annotator agreement. . . . 160

5.1. List of prosodic features and level of representation used throughout this thesis. . . . 166

A.1. Alphabetized list of feedback lexical forms found in literature ranging from A-N. . . . 196

A.2. Alphabetized list of feedback lexical forms found in literature ranging from O-U. . . . 197
A.3. Alphabetized list of feedback lexical forms found in literature ranging from W-Z. . . . 198

B.1. List of language and corpus type used in publications predicting the timing of feedback responses. . . . 199
B.2. List of language and corpus type used in publications predicting the lexical form of feedback responses. . . . 200
B.3. List of language and corpus type used in publications predicting the communicative function of feedback responses. . . . 200
B.4. List of language and corpus type used in publications predicting the prosody of feedback responses. . . . 201

C.1. List of publications predicting the timing of feedback responses and the type of evaluation metrics reported. . . . 202
C.2. List of publications predicting the lexical form of feedback responses and the type of evaluation metrics reported. . . . 202
C.3. List of publications predicting the communicative function of feedback responses and the type of evaluation metrics reported. . . . 203
C.4. List of publications predicting the prosody of feedback responses and the type of evaluation metrics reported. . . . 203

D.1. List of categories predicted and their definitions in publications from 2020-2022 predicting the communicative function of feedback responses. 205
D.2. List of categories predicted and their definitions in publications from 2023 predicting the communicative function of feedback responses. . . . 206
D.3. List of categories predicted and their definitions in publications from 2024 predicting the communicative function of feedback responses. . . . 207

E.1. Tukey pair-wise test for duration. . . . 208
E.2. Tukey pair-wise test for mean pitch. . . . 208
E.3. Tukey pair-wise test for pitch slope. . . . 209
E.4. Tukey pair-wise test for pitch range. . . . 209

F.1. Top 30 part-of-speech bigrams used in Figueroa, Ochs, et al. (2023) as context features. . . . 210

G.1. Examples of the 39 dialog acts available from DialogTag. . . . 212

I.1. Pearson's correlation coefficients between the mean pitch of feedback responses and preceding 500 ms utterance of the interlocutor. . . . 222
I.2. Pearson's correlation coefficients between the mean pitch of feedback responses and following 500 ms utterance of the interlocutor. . . . 223
I.3. Pearson's correlation coefficients between the mean pitch of the interlocutor's preceding and following 500 ms utterance. . . . 223

I.4. Pearson's correlation coefficients between the mean intensity of feedback responses and preceding 500 ms utterance of the interlocutor. . . 224
I.5. Pearson's correlation coefficients between the mean intensity of feedback responses and following 500 ms utterance of the interlocutor. . . 224
I.6. Pearson's correlation coefficients between the mean intensity of the interlocutor's preceding and following 500 ms utterance. 225
I.7. Pearson's correlation coefficients between the pitch slope of feedback responses and preceding 500 ms utterance of the interlocutor. 225
I.8. Pearson's correlation coefficients between the pitch slope of feedback responses and following 500 ms utterance of the interlocutor. 226
I.9. Pearson's correlation coefficients between the pitch slope of the interlocutor's preceding and following 500 ms utterance. 226

Introduction

If someone were to ask you to describe a face-to-face conversation between two people, what would you say? You might answer that a conversation involves one person speaking while the other listens. Or perhaps you would describe a conversation as people taking turns speaking and listening. A more unlikely answer might be that conversations are a setting in which people use language. In his book *Using Language*, H. H. Clark 1996 argues that language use is a form of joint action and that conversations are a setting in which language is used. To fully understand this statement, we must first unpack the definitions of language and joint action.

Language is defined as a system of signals that are both linguistic and non-linguistic (ibid., p. 392), and a signal is "any action by which one person means something for another person" (ibid., p. 13). Linguistic signals can be utterances, and non-linguistic signals can be, for example, eye-gaze, laughter, or gestures such as head nods and head shakes. When we think about conversations, we might focus on linguistic signals, but non-linguistic signals are also equally important in face-to-face conversations.

A joint action is an action "that is carried out by an ensemble of people acting in coordination with each other" (ibid., p. 3). Language use is a joint action because both the speaker and the listener coordinate their individual actions. What are these individual actions that the speaker and listener coordinate? According to ibid., at every moment of a conversation, both the speaker and the listener perform corresponding actions, see Table 1, which operate in a hierarchical structure.

Level	**Speaker A's actions**	**Addressee B's actions**
4	A is *proposing* joint project w to B	B is *considering* A's proposal of w
3	A is *signaling* that p for B	B is *recognizing* that p from A
2	A is *presenting* signal s to B	B is *identifying* signal s from A
1	A is *executing* behavior t for B	B is *attending* to behavior t from A

Table 1. – The levels of joint actions by the speaker and the addressee (H. H. Clark 1996, p. 152).

For example, if speaker A says, "Would you like a coffee?" to addressee B, at level 1, speaker A is executing, or vocalizing, the utterance, and addressee B has to attend to the sounds being vocalized by speaker A. At level 2, speaker A presents the utterance to addressee B, and addressee B has to identify the words that are being said. At level 3, addressee B has to recognize, i.e., understand, the meaning of the signal (utterance) being signaled by speaker A. At level 4, addressee B considers answering the question.

Addressee B may take up the proposal by nodding their head or answering "yes"; here the proposed project is getting addressee B to answer speaker A's question.

While it may seem obvious that humans coordinate their actions in conversation, this coordination extends well beyond simple question-answer pairs. This becomes evident when we observe conversations closely. Throughout a conversation, while the speaker is speaking, the listener does not remain silent or motionless. In fact, the listener often provides continuous signals to the speaker and the speaker may adjust their speech in response. These listener signals are referred to as *feedback responses*. Feedback responses can be vocal (e.g., uh-huh, yeah, wow) or visual (e.g., head nods, smiles, laughter) and they may occur together or separately (Heldner, Hjalmarsson, et al. 2013; Prévot, Gorisch, et al. 2016; Jain et al. 2021).

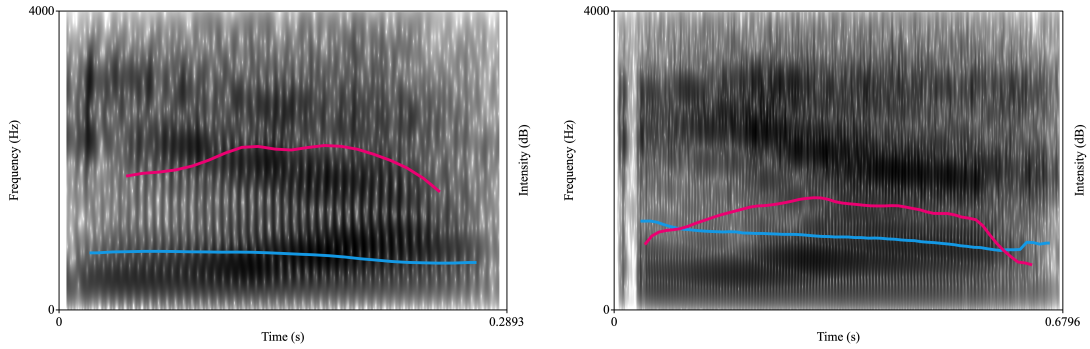
The focus of this thesis is on modeling and generating vocal feedback for conversational agents. While both vocal and visual feedback are equally important in face-to-face conversations, this thesis focuses on vocal feedback because not all conversational agents are embodied, and there has been less work on modeling and generating vocal feedback specifically. To effectively model and generate vocal feedback, we must first understand why feedback responses are important for human-human communication.

0.1. Why are feedback responses important for human-human communication?

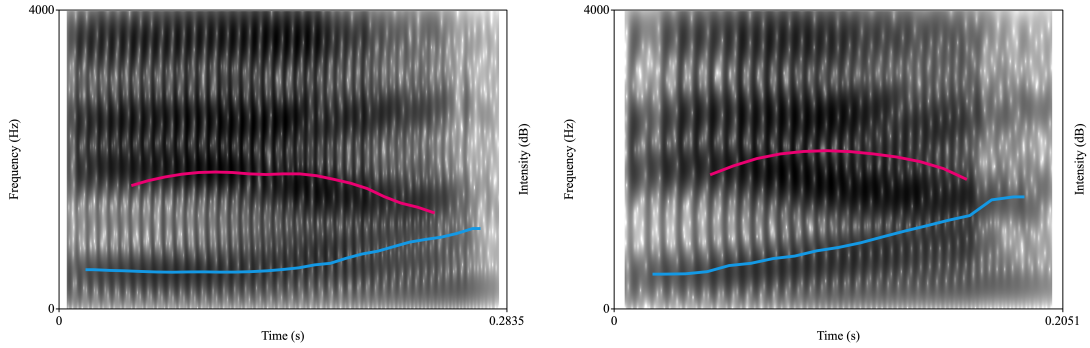
In conversations, feedback responses are essential for establishing common ground, that is, people's "mutual, common, or joint knowledge, beliefs, and suppositions" (H. H. Clark 1996, p. 93). For example, imagine interlocutor A and interlocutor B are on a first date. At the beginning of their conversation, their common ground may consist of knowing each other's name, knowing that they both speak the same language, and knowing that they are both open to meeting someone. If interlocutor A begins by talking about their day and interlocutor B is listening, nodding their head, and maybe responding with "mhm" as well, then interlocutor A can assume that interlocutor B is attending to what they are saying and that the information is being added to their common ground. If interlocutor B is momentarily distracted during the telling of A's story, interlocutor B may signal this by tilting their head and saying "pardon". These feedback responses will signal to interlocutor A that interlocutor B has not understood what was said. In response, interlocutor A may repeat themselves until the information is added to their common ground. These feedback responses help the speaker monitor if there is a breakdown in communication, if so, the speaker adjusts their speech until communication is restored.

In essence, feedback responses provide information about the state of mind of the interlocutor who is listening: are they following, are they understanding, what is their opinion or emotion. The listener can signal their state of mind through feedback responses and the speaker can interpret or recognize the communicative functions of the feedback responses. The listener can also convey various communicative

functions (Allwood, Nivre, et al. 1992; Neiberg, Salvi, et al. 2013; Prévot, Gorish, et al. 2015) to the speaker by varying the prosodic realization and/or lexical form of the vocal feedback response. For example, the word "yeah" can express agreement, disagreement, doubt, surprise, or it can simply signal that the listener is listening and encourage the interlocutor to continue speaking. As demonstrated by the "yeah" realizations from the Switchboard corpus displayed in Figure 1, prosodic correlates such as pitch, intensity, and duration are typically realized in different ways depending on the intended communicative function.



(a) continue (b) agree



(c) disagree (d) surprise

Figure 1. – Spectrograms of various *yeah* feedback responses conveying different communicative functions: (a) continue speaking, (b) agreeing, (c) disagreeing, (d) surprise. Pitch contour in Hz is shown in cyan, intensity contour in dB is shown in magenta.

Feedback responses are also important for the smooth flow of turn-taking in conversations (Yngve 1970). As an analogy presented by Duncan (1972), just as people avoid bumping into each other on the street, people also avoid metaphorically bumping into each other in conversation. Interlocutors avoid "bumping" into each other by coordinating who should take on the role of the speaker and who should be the listener at every moment of the conversation using feedback responses. For instance, with eye gaze, when the interlocutor who is speaking is ready to exchange roles with

the interlocutor who is listening, the speaker will direct their eye gaze towards the listener until there is mutual eye gaze, and then the listener will look away to signal that they will now take on the role of speaker (Kendon 1967). In contrast, listeners can encourage the speaker to continue speaking by responding with "mhm" or "uh-huh" and/or nodding their heads.

One question that may arise is how do speakers know that feedback responses such as "mhm" or an "uh-huh" convey that the speaker should continue speaking? Why are these feedback responses unobtrusive to the flow of the conversation? One explanation may lie in *prosodic alignment*, a phenomenon where interlocutors become similar in their prosodic features (Wynn et al. 2022). For example, if the speaker raises their pitch or energy, the listener may also raise their pitch and energy. R. Levitan, Beňuš, et al. (2015b) found that the prosodic features of *backchannels*, which are a type of feedback responses–signals such as "mhm" or "uh-huh" that signal to the speaker that they should continue speaking while not interrupting the speaker–align to the prosodic features of the preceding utterance of the interlocutor. Heldner, Edlund, and Hirschberg (2010) also found pitch similarity between backchannels and the preceding utterance of the interlocutor; they posited that this may be a reason why backchannels are unobtrusive.

In summary, feedback responses are essential for human-human communication as they help establish common ground, coordinate turn-taking, and convey information about the listener's state-of-mind. In the case of vocal feedback, both the lexical form and prosodic realization play key roles. They signal when there is a breakdown in the conversation and convey the communicative function of the feedback response. Specifically, prosody is important because it ensures that feedback responses remain unobtrusive to the flow of conversation.

0.2. Differences between human-human communication and human-conversational agent communication

Conversational agents are becoming prevalent in our daily lives; they may be in the form of chatbots, voice assistants, virtual agents, or social robots. However, many of these conversational agents fall short of being truly conversational; they are unable to carry conversations in the same way as humans do (Luger et al. 2016; Porcheron et al. 2018; L. Clark et al. 2019).

One aspect that impedes human-like communication is that some of these conversational agents such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, Google Nest and OpenAi's ChatGPT are not embodied. Although virtual agents are embodied, they are not physically present; having a conversation where both interlocutors are in the same physical space is different from having one interlocutor on a screen. In their work, Jamy Li (2015) surveyed previous research comparing virtual agents and physically present robots and found that there is a benefit of having physical presence

compared to only having embodied agents. Social robots such as the NAO robot (Pot et al. 2009) and the Furhat robot (Al Moubayed et al. 2012) are both embodied and physically present. However, both robots are still not able to fully communicate the way humans do. For example, the NAO robot has limbs and can gesture but has a static face; conversely, the Furhat robot has a back-projected face capable of making many face expressions but lacks limbs for gesturing. Examples of these conversational agents are illustrated in Figure 2. To model and generate visual feedback that is as close to human feedback, these conversational agents need to produce all kinds of visual feedback.



(a) Amazon's Alexa (b) DAVI's virtual agent

(c) NAO Robot (d) Furhat Robot

Figure 2. – Example of conversational agents. (a) Amazon's Alexa is not physically embodied but is physically present. (b) DAVI's virtual agent is physically embodied but not physically present. (c) and (d) both social robots are embodied and physically present.

Another aspect that limits human-like communication with these conversational agents is their limited ability in the kinds of dialogue they can carry out with humans. Many of these conversational agents can only handle task-oriented conversations (Porcheron et al. 2018; L. Clark et al. 2019). For instance, users can ask these conversational agents to add items to a virtual shopping cart, get directions, set alarms or

timers, or obtain information about the weather or historical figures. Users typically make these requests either as questions, e.g., "Alexa, what is the weather like today?" or as instructions, e.g., "Alexa, tell me the weather" (Porcheron et al. 2018). However, to be understood, users often modify their speech by "dropping words other than 'keywords', removing colloquial or complex words, reducing the number of words used, using more specific terms, altering enunciation, speaking more specific terms, altering enunciation, speaking more slowly/clearly and changing accent" (Luger et al. 2016). Task-oriented conversations are just one type of interaction; humans also engage in social conversations such as small talk, recounting their day to a friend, or gossiping (L. Clark et al. 2019). The inability to participate in social conversations makes the term "conversational agents" a misnomer.

Currently, OpenAI's ChatGPT and GPT-4o can be considered the most prominent "human-like" conversational agents. However, this is likely to change soon, as the field of conversational AI is rapidly evolving. Both ChatGPT and GPT-4o are based on large language models (LLMs) (Brown et al. 2020; Achiam et al. 2023). LLMs are computational models trained on vast amounts of data to generate human language by predicting the next word in a sequence given the context (Zhao et al. 2023; Chang et al. 2024). These agents are termed "conversational" because they can maintain a dialogue with users, remembering previous interactions and following up on questions. With GPT-4o users can even input text, audio, and visuals. However, the way we converse with these conversational agents is in stark contrast to the way humans converse. Typically, interactions with conversational agents progress turn-by-turn. For instance, when using ChatGPT, if you type while it is responding, you must press a button next to the input box to interrupt it, rather than pressing the usual Enter key. Although ChatGPT continues writing its response as you interrupt it, it only acknowledges the interruption once it has finished its initial response. When talking with GPT-4o, interruptions are possible but can be jarring since the agent is immediately cut off. Even when conversational agents can engage in social conversations, they usually remain silent while the user speaks or writes, responding only after the user has finished. Many conversational agents do not demonstrate active listening by providing vocal or written feedback responses during the user's input.

To sum up, embodiment, physical presence, and the modalities through which humans interact with conversational agents are crucial for making these agents more akin to human-human communication. While feedback responses play a vital role in human-human communication, they are often lacking in conversational agents. Addressing these gaps could significantly improve the effectiveness and naturalness of human-conversational agent communication.

0.3. Motivation

As we have established, both visual and vocal feedback are important for face-to-face conversations. However, visual and vocal feedback are often lacking in conversational agents, despite being essential for making human-conversational agent communica-

tion more human-like. As mentioned, we chose to focus on vocal feedback, as not all conversational agents are embodied, and there has been significant work on modeling and generating visual feedback (Kopp, Allwood, et al. 2008; Bevacqua et al. 2010; L. Huang et al. 2011; Oertel, Lopes, et al. 2016).

In order for these conversational agents to be able to generate vocal feedback similar to humans, the vocal feedback of the conversational agents must meet the following criteria:

1. **Appropriate timing:** The agent should generate vocal feedback at moments (when) and positions (where) that are appropriate for the conversation. For example, the agent should give feedback when the user is telling a long story. Low pitch regions have been shown to indicate appropriate places where feedback responses may be generated (Ward and Tsukahara 2000).
2. **Lexical and prosodic variability:** When talking on the phone, if the other person continuously says "mhm" with a flat pitch without changing their feedback responses, you might interpret it as the them being distracted or bored. Therefore, the agent should generate feedback responses that are variable in lexical form and in prosody so as not to sound monotonous or robotic.
3. **Convey the appropriate communicative function:** The agent should convey the appropriate communicative function for the given context. For example, if the user says, "My dad died today.", the agent's response of "aww" should sound sympathetic, not like the "aww" used in response to "That is a cute puppy."
4. **Be unobtrusive:** Just as in human-human conversations, feedback responses signaling "continue speaking" should be unobtrusive to the conversation. This can be done by modeling prosodic alignment; the agent should align their prosodic features to those of the user in order create feedback responses that blend seamlessly into the conversation.

Incorporating or modeling vocal feedback in conversational agents is an active area of research, see a review of modeling feedback responses (Axelsson et al. 2022) and Section 1.5 for a list of references. However, most existing models primarily focus on predicting the timing of feedback responses, i.e., whether the conversational agent should give a feedback response (Ishii et al. 2021; Adiba, Homma, Bertero, et al. 2021; Jinhan Wang et al. 2024). Others have focused on predicting the type, i.e., what communicative function the conversational agent should convey (Ortega, C.-Y. Li, et al. 2020; Jang et al. 2021; Onishi et al. 2023). More recently, there has been work predicting both timing and type (Lala, Inoue, et al. 2022; Boudin et al. 2024; Choi et al. 2024). Less focus has been on models predicting the prosodic features or lexical form of feedback responses that are appropriate for its context. Currently, the primary approach for these conversational agents to generate vocal feedback with prosodic or lexical form variability is to have a dataset of feedback responses and randomly select from the dataset. Prosodic variability is typically introduced through signal processing techniques (Åsa Wallers et al. 2006; Stocksmeier et al. 2007; Pammi, Schröder, Charfuelan, et al. 2010).

One key characteristic of conversational speech that is missing in these conversational agents is the generation of *short feedback responses*[1] (e.g., "wow", "uh-huh", "right") with contextually appropriate prosody. The prosodic features of vocal feedback are crucial for creating non-monotonous feedback, conveying the intended communicative function, and ensuring that the feedback responses are unobtrusive.

The main research question this thesis investigates is what is needed for synthesizing feedback responses with appropriate prosody for its context. Specifically, we focus on the "how", how these short feedback responses should sound—their prosodic realization and representation.

0.4. U.S. English human-human speech corpora

In order to model vocal feedback from human-human conversations, we first reviewed suitable U.S. English corpora. English was selected because it is typically the most researched language which provides more options of available corpora and existing infrastructure for text-to-speech. U.S. English was specifically selected because it is one of my mother tongues and there are differences between the vocal feedback of English variants (Tottie 1991). In other words, there are intricacies that a non-native speaker has to learn to fully grasp the vocal feedback responses of a foreign language (Ward, Escalante, et al. 2007; Wolf 2008; Cutrone 2010; Sbranna et al. 2023).

One U.S. English corpus that has been annotated for and used to research discourse markers, affirmative words, and backchannels is the Columbia Games Corpus (Beňuš et al. 2007; Gravano, Beňuš, Hirschberg, et al. 2007; Gravano, Beňuš, Chávez, et al. 2007; Gravano and Hirschberg 2009). Both affirmative words and backchannel are types of feedback responses. The Columbia Games Corpus consists of 12 dyadic conversations, totaling 9 hours and 45 minutes of dialogue. Six females and seven males were recorded playing two computer games, the Cards game and the Objects game. The participants were seated in the same room, but a curtain separated them to ensure the interaction was limited to speech communication. The participants had to collaborate to complete the games. In the Cards game, participants matched their cards by describing them to each other. In the Objects game, both participants had objects on their screen, but only one object was in a different position. The participants had to collaborate to move the object to the correct position.

Another collaborative task-oriented corpus is the MIT American English Map Task, which is based on the HCRC Map Task Corpus (Anderson et al. 1991). In this task, each participant has a map: one participant has a route and gives instructions, while the other must reproduce the route on their own map, despite the maps being different. The MIT American English Map Task corpora consists of 16 dyadic conversations and only had eight female participants. Unlike the Columbia Games Corpus, the conversations have not been annotated or transcribed. Since we wanted to model

1. We distinguish feedback responses such as "oh wow" from short feedback responses such as "wow".

vocal feedback that occurred in social conversations we did not consider corpora comprised of task-oriented conversations.

There are a several U.S. English corpora that contain social talk. The Santa Barbara Corpus of Spoken American English includes a variety of different types of conversations, such as task-oriented, face-to-face conversations, telephone conversations, lectures, and sermons. The entire corpus amounts to about 20 hours of recordings. A downside of this corpus is that some conversations include more than 2 people, and some recordings are in a single channel. Although we are unaware if there are differences between vocal feedback in dyadic versus multi-party conversations, we focused on vocal feedback in dyadic conversations since they resemble the settings of most human-conversational agent interactions. Additionally, the conversations are not always balanced, as some have one interlocutor speaking most of the time, such as in sermons and lectures. The dyadic conversations in the corpus were often between acquaintances, couples, friends, or family members.

Three other corpora that also contain conversations between friends or family members are CALLHOME, CALLFRIEND American English-Non-Southern Dialect, and CALLFRIEND American English-Southern Dialect. All of these conversations are telephone conversations. The CALLHOME corpus consists of 120 telephone conversations, each about 30 minutes long. The CALLFRIEND American English-Non-Southern Dialect and CALLFRIEND American English-Southern Dialect both contain 60 telephone conversations, each ranging from 5-30 minutes in length. Although we are unaware of differences in vocal feedback between conversations involving strangers versus friends or family, we preferred conversations between unacquainted interlocutors since conversations between human-conversational agents tend to be less intimate.

Two corpora consisting of telephone conversations between unacquainted interlocutors are the Switchboard corpus (Godfrey et al. 1992) and the Fisher corpus (Cieri et al. 2004). Both in the Fisher corpus and in the Switchboard corpus, participants were given topics to discuss. The Fisher corpus contains 5,850 telephone conversations each about 10 minutes long. However, the Fisher corpus did not meet our requirements since it included non-native English speakers and other varieties of non-U.S. English. Ultimately, we decided to use the Switchboard corpus, a corpus consisting of about 2,500 telephone conversations between 500 speakers who are unacquainted.

One of the main advantages of the Switchboard corpus is that is has been manually transcribed and word level start and end times are provided. In addition, each speaker is recorded on separate channels, which is especially useful for separating feedback responses, as they often overlap with the speaker's speech. Another advantage of the Switchboard corpus is that it has been previously annotated for dialog acts and two types of vocal feedback, feedback that signal agreement and "continue speaking" (backchannels) (Jurafsky et al. 1998). Moreover, the Switchboard corpus has been used to study backchannels (Jurafsky et al. 1998; Lai 2008; Lai 2009) and for modeling backchannels (Ruede et al. 2019; Ortega, C.-Y. Li, et al. 2020; Adiba, Homma, Bertero, et al. 2021; Jinhan Wang et al. 2024; Choi et al. 2024). The main disadvantage of this corpus is that the conversations are not face-to-face. However, as mentioned, most

conversational agents are currently voice-based, making the Switchboard corpus well suited for modeling feedback responses.

A recently released corpus of face-to-face video conversations, the Candor Corpus (Reece et al. 2023), consists of 1,656 conversations, totaling approximately 850 hours. This corpus closely resembles spontaneous face-to-face conversations, with interlocutors in the same physical space. However, there are a few drawbacks: some interlocutors are not native English speakers, conversations can be out of sync due to internet quality (delays are also present, though less frequently in the Switchboard corpus), and the corpus has been automatically transcribed using Automatic Speech Recognition (ASR), which results in numerous transcription errors and sometimes fails to recognize certain feedback responses. Despite these limitations, if the Candor Corpus had been available at the start of this PhD, it would have been our corpus of choice.

To the best of our knowledge, there is no U.S. English corpus that consists of a large set of face-to-face conversations exclusively involving social talk, where both unacquainted interlocutors are recorded in separate channels without overlapping recordings, and that are recorded in high quality audio. Furthermore, such a corpus should ideally be manually transcribed, timestamped, and annotated with different communicative functions of vocal feedback. Given the limitations in existing corpora, in this thesis we used the Switchboard corpus, as we believe it is well suited for voice-based conversational agents.

0.5. Research questions

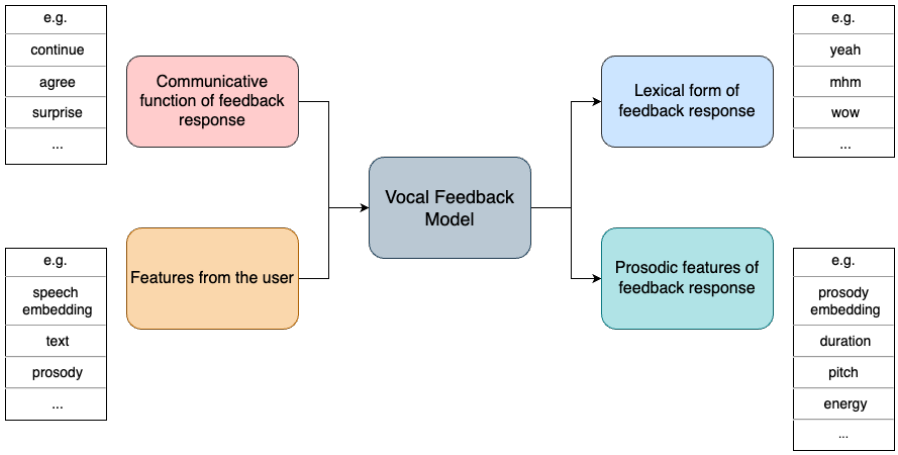


Figure 3. – Proposed components for a vocal feedback model.

As mentioned, the main research question of this thesis is: what is needed for synthesizing feedback with appropriate prosody for its context? More concretely: How should these feedback responses sound (their prosodic realization and representation)?

In this thesis, we will use a Text-to-Speech (TTS) system to synthesize (generate) vocal feedback. A TTS system takes in text and outputs a speech waveform. Specifically, we will use the FastPitch model (Łańcucki 2021), which allows for prosodic manipulation, enabling us to introduce prosodic variability in the generated feedback responses. However, before synthesizing feedback responses, it is essential to model the key aspects of vocal feedback: timing, lexical form, prosodic features, communicative function, and unobtrusiveness.

In order for conversational agents to synthesize vocal feedback that is appropriate for its context, the agent needs to convey the intended communicative function through both the prosodic realization and the lexical form of the vocal feedback, while ensuring that the vocal feedback remain unobtrusive. Therefore, we propose components for a vocal feedback model, see Figure 3.

This proposed vocal feedback model takes as input the communicative function of the vocal feedback that the system should convey. Note that we are not predicting the communicative function; we assume that the agent is given the communicative function either by an interaction designer or another model that predicts the communicative function. The model also takes as input the preceding context, the acoustic features and/or text from the user. The acoustic features are important for the model to align its prosodic features with those of the user. The outputs of the model will be the lexical form and the prosodic features of the vocal feedback. Note that this model does not take into account the timing of the vocal feedback as our main focus is to model "how" these feedback responses should sound.

Thus, given the communicative function of the feedback that the agent should convey and the user's context, the model will predict the lexical form and prosodic features of the vocal feedback. For instance, if the conversational agent is intended to convey surprise after the user says "I just found out I'm having twins.", the model could predict "oh", "wow", or "really", and the prosody should sound like surprise without interrupting the user's talk. Figure 4 shows the two-step process for modeling and generating vocal feedback responses.

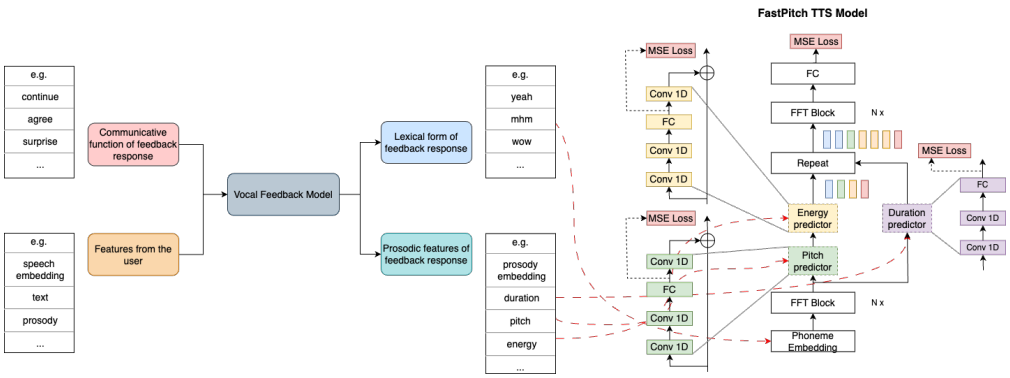


Figure 4. – Proposed two step process for modeling and generating vocal feedback responses.

We further break down each proposed component needed for this vocal feedback model into separate research questions:

R1 *What are the various communicative functions of feedback responses in human-human conversations that a conversational agent may be required to convey?*

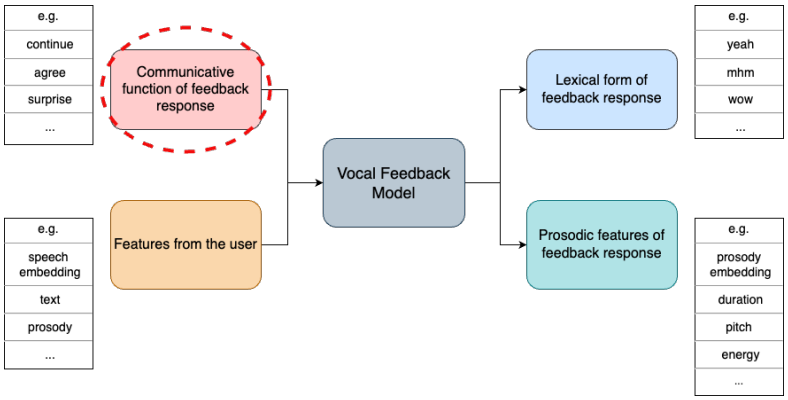


Figure 5. – Component from the vocal feedback model that are addressed by R1.

Figure 5 shows the component of the vocal feedback model that is addressed by R1. Before synthesizing feedback responses, we must first understand the communicative functions that occur in human-human conversations. Given the absence of a standardized annotation scheme for annotating communicative functions of feedback responses and lack of nuance in the labels—particularly for those signaling attitudinal information—we propose a new annotation scheme based on the communicative functions found in the Switchboard corpus.

R2 *Per each communicative function found in human-human conversations what are the different distributions in terms of lexical forms and prosody?*

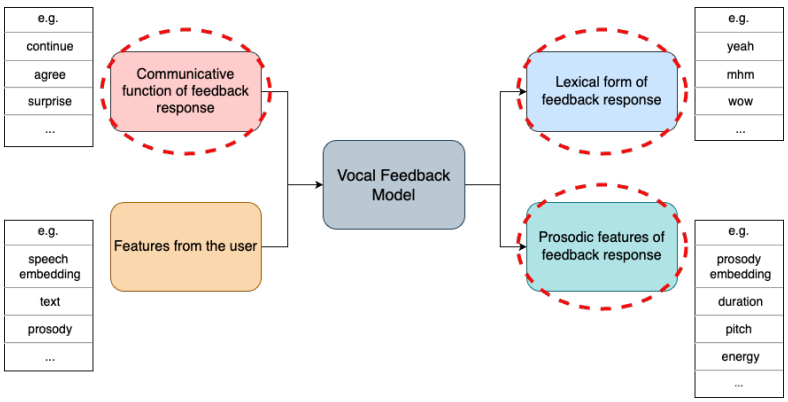


Figure 6. – Components from the vocal feedback model that are addressed by R2.

Figure 6 shows the components of the vocal feedback model that are addressed by R2. As previously mentioned, lexical and prosodic variability are crucial for ensuring that the conversational agent does not sound monotonous. Often, previous research on annotating or investigating vocal feedback responses has omitted a comprehensive list of lexical forms. Therefore, we have compiled a list of lexical forms found in the Switchboard corpus. We also conducted a prosodic analysis to determine whether different communicative functions exhibit distinguishable prosodic features.

R3 *Is it possible to automatically classify the communicative functions of feedback responses based on their lexical form, prosodic features, and preceding context?*

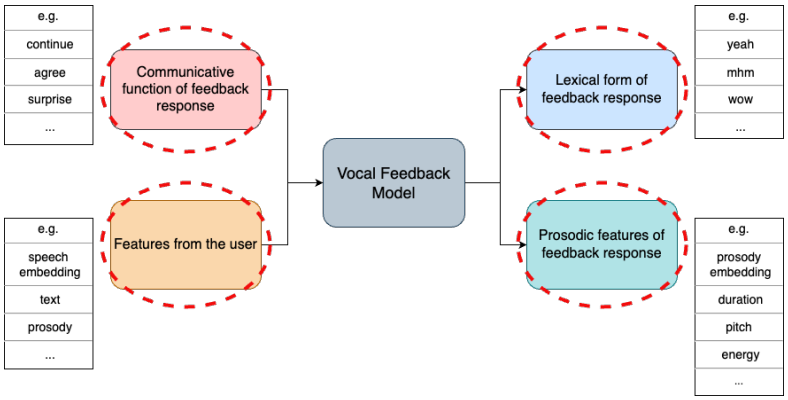


Figure 7. – Components from the vocal feedback model that addressed by R3.

Figure 7 shows the components of the vocal feedback model that are addressed by R3. Previous studies have used prosodic features to disambiguate and classify the communicative functions of feedback responses, but have focused only on certain lexical form (Lai 2008; Truong and Heylen 2010). Others have classified feedback responses into two broad categories: *generic* and *specific* (Prévot, Gorish, et al. 2015). However, these broad categories are not nuanced enough, as such classifiers often lump distinct feedback responses like those expressing surprise and sympathy into a single category. Therefore, we aim to classify the communicative functions found in Switchboard by using the lexical forms, prosodic features, and preceding context.

R4 *Does prosodic alignment occur for different communicative functions of feedback responses?*

Figure 8 shows the components of the vocal feedback model that are addressed by R4. There is some evidence suggesting that feedback responses are unobtrusive to the flow of conversation due to prosodic alignment (R. Levitan, Beňuš, et al. 2015b; Heldner, Edlund, and Hirschberg 2010). However, these studies have focused exclusively on feedback signaling "continue speaking" (i.e., backchannels), which represent only one communicative function. Therefore, we measure prosodic alignment for all communicative functions found in the Switchboard corpus.

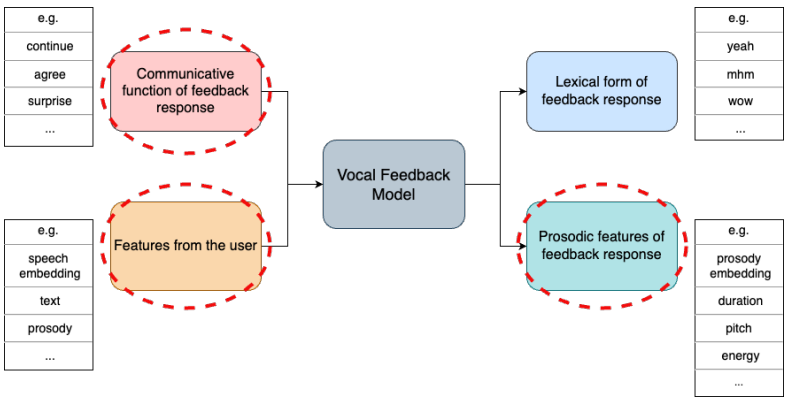


Figure 8. – Components from the vocal feedback model that are addressed by R4.

R5 *Can we synthesize feedback responses that sound natural and appropriate in context while conveying their intended communicative function?*

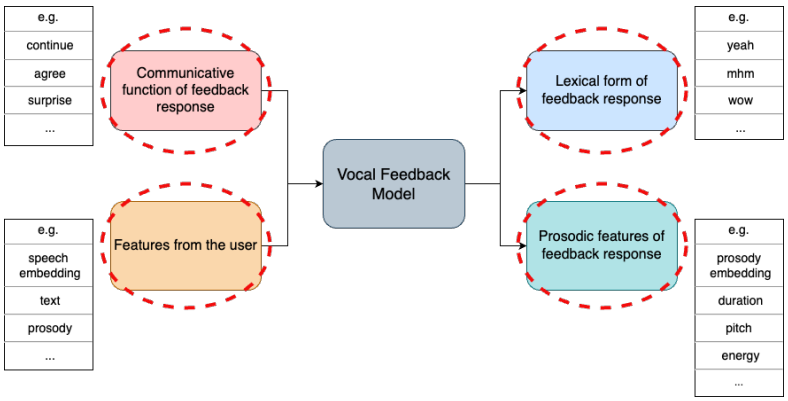


Figure 9. – Components from the vocal feedback model that are addressed R5.

Figure 9 shows the components of the vocal feedback model that are addressed by R5. One method for creating prosodic and lexical variability for feedback responses in conversational agents is to use a database of pre-recorded audios (Campbell 2007; Oertel, Jonell, et al. 2017); however, this approach is only feasible if a sufficient number of recordings are available. Alternatively, some have relied on signal processing techniques (Åsa Wallers 2006; Stocksmeier et al. 2007; Chandler 2023) to synthesize feedback responses with variable prosody. However, signal processing techniques are outdated and are known to degrade the audio quality. In contrast, state-of-the-art neural TTS systems can now generate more natural sounding speech. Therefore, we trained a state-of-the-art neural TTS system that can generate vocal feedback. We conducted a listening test to assess the naturalness, appropriateness of the prosody, as well as the preservation of the intended communicative function of the synthesized feedback responses.

0.6. Thesis Contributions and Outline

The remainder of this thesis is organized as follows:

Chapter 1 – In this chapter, we review the relevant background on vocal feedback responses. We review the various terms used to describe feedback responses, and we examine their different communicative functions and existing annotation schemes. Additionally, we review previous studies investigating the relationship between prosody and the communicative functions of feedback responses. We also discuss the phenomenon of alignment, including the different terms used to describe it, and discuss previous work measuring prosodic alignment in feedback responses. Finally, we review previous research on predicting the timing, lexical form, communicative function, and prosody of feedback responses, as well as work synthesizing feedback responses.

Chapter 2 – This chapter is based on our publication *"Annotation of Communicative Functions of Short Feedback Tokens in Switchboard"* (Figueroa, Adigwe, et al. 2022). In this chapter, we address **R1** by proposing our annotation scheme for communicative functions of vocal feedback responses, as well as presenting the results of our inter-annotator agreement scores. This chapter also addresses **R2** by presenting the distribution of the lexical forms per each communicative function found in the Switchboard corpus. We also present the prosodic analysis of the vocal feedback responses in Switchboard.

Chapter 3 – This chapter is based on our publication *"Classification of Feedback Functions in Spoken Dialog Using Large Language Models and Prosodic Features"* (Figueroa, Ochs, et al. 2023). In this chapter, we address **R3** by presenting the results from our classification task. We experimented with combinations of lexical and prosodic features from feedback responses, as well as context features from the preceding utterance of the interlocutor, in order to classify the communicative functions of the feedback responses in Switchboard. Additionally, we experimented with OpenAI's GPT-3 to classify the communicative function of feedback responses using only lexical and contextual information. We also experimented with GPT-3 by encoding its probability distributions as input features for a main classifier.

Chapter 4 – This chapter is based on our publication *"Prosodic Alignment in Different Conversational Feedback Functions"* (Figueroa, Beňuš, et al. 2023). In this chapter, we address **R4** by presenting the results from our alignment measurements. We measure local alignment between the prosodic features of the feedback responses and the preceding and following utterance of the interlocutor for all the communicative functions found in the Switchboard corpus.

Chapter 5 – This chapter is based on our publication *"Mhm... Yeah? Okay! Evaluating the Naturalness and Communicative Function of Synthesized Feedback Responses*

in Spoken Dialogue" (Figueroa, Korte, et al. 2024). In this chapter, we address **R5** by presenting the results from our listening tests, where participants listened to synthesized feedback responses in the original context of the Switchboard conversation. Participants rated the naturalness and prosodic appropriateness of the synthesized feedback responses and annotated them with their perceived communicative function. We used two synthesis methods: signal processing and TTS. To ascertain if both methods preserved the intended communicative function from the original Switchboard feedback responses, we transplanted the prosody of the feedback response from Switchboard listener at the frame-level using signal processing and at the phone-level using the TTS approach.

Conclusion – In the final chapter, we summarize our results, and discuss their implications for modeling vocal feedback in conversational agents. Lastly, we conclude by outlining directions for future work.

1. Background

Summary

1.1. Feedback, backchannels, continuers, or listener responses? 40
1.2. Communicative functions of feedback responses 44
1.2.1. Schemes based on Allwood, Nivre, et al. (1992)'s communicative functions 45
1.2.2. Schemes based on generic/specific distinction 50
1.3. Relationship between prosody and communicative functions of feedback responses 54
1.3.1. Prosodic description or prosodic analysis studies 54
1.3.2. Perception studies 56
1.3.3. Classification studies 61
1.4. Alignment and feedback responses 68
1.4.1. What is alignment? 69
1.4.2. Measuring alignment in speech 71
1.4.3. Previous studies 75
1.4.3.1. Studies in U.S. English 75
1.4.3.2. Studies in Slovak, Swedish, and Japanese 80
1.5. Modeling feedback responses in conversational agents 83
1.5.1. Modeling vocal and visual feedback responses 84
1.5.2. Predicting the timing of feedback responses 85
1.5.3. Predicting the lexical form of feedback responses 87
1.5.4. Predicting the communicative function of feedback responses . 88
1.5.5. Predicting the prosody of feedback responses 93
1.5.6. Synthesis of feedback responses 94

1.1. Feedback, backchannels, continuers, or listener responses?

There is a plethora of terms used in literature to describe *feedback responses*, as illustrated by the numerous terms listed in Table 1.1, which is partially based on the lists by Fujimoto (2009) and Pammi (2011). Note this is by no means an exhaustive list. Additionally, even for a single term, the spellings are not consistent (eg., backchannel, back channel, back-channel). The abundance of terminology used to describe feedback responses can be likened to the parable of the blind men and an elephant. In

the parable, one blind man feels the trunk and concludes it feels like a snake; another touches an ear and believes an elephant is like a fan; a third feels a leg and thinks it resembles a pillar; a fourth man touches the tail and concludes the elephant is like a rope; a final man touches the tusk and likens the elephant to a spear. Similarly, in the study of feedback responses, researchers often focus on different aspects, leading to varied and sometimes conflicting definitions of what does and does not constitute a feedback response. Just as the blind men each grasp only a part of the elephant, different researchers introduce new terms to fit their specific research questions. In this section we will review the definition of some of these terms and explain why we prefer using the term *feedback responses*.

In his 1952 book *The Structure of English: An introduction to the construction of English sentences*, Fries (1952) discussed various kinds of utterances. He described one class of utterances given by the listener to the speaker as *signals of continued attention*, which do not interrupt the speaker. He also noted that these signals of continued attention occur in face-to-face conversations as head nods. However, it appears that the term *signals of continued attention* did not become widely adopted.

Krauss and Weinheimer (1966) used the term *concurrent feedback* to describe utterances from the listener that occurs simultaneously with the speaker's talk and that "inform the speaker of the effect his message is having." Similarly, while investigating the functions of eye-gaze, Kendon (1967) introduced the term *accompaniment signals*, where "accompaniment" referred to the accompaniment of the speaker's talk and not the eye-gaze. Accompaniment signals were defined as "utterances that the listener produces as an accompaniment to a speaker, when the speaker is speaking at length" (ibid.). Kendon (ibid.) further categorized these accompaniment signals into two classes: *attention signals* (only signaling attention) and *assenting signals* (signaling agreement). Dittmann et al. (1967) investigated where these listener utterances occur and their interaction with head nods (Dittmann et al. 1968), referring to them as *listener responses*. However, all of these terms focused on vocalizations and just as the term *signals of continued attention*, none of these terms were widely adopted.

It was only after Yngve (1970) introduced the term *back channel* that a widely adopted term emerged to refer to vocal and visual listener signals. Yngve (ibid.) coined the term to differentiate between the main channel, where the person currently holding the floor (the right to speak) is speaking, and the *back channel*, where the speaker receives messages (*back-channel messages*) from the listener. Yngve (ibid.) provided examples of short back-channel messages, including short utterances like "yeah" and "uh-huh", sentence completions, comments such as "Oh, I can believe it", and questions like "You've started writing it then – your dissertation?" Despite the initial examples provided by Yngve (ibid.), the definition of backchannels has evolved over time. Duncan (1972), Duncan Jr et al. (1974), and Duncan (1974) adopted the term *back-channel signals* and further expanded the types to include brief requests for clarification, and restatements of what the speaker previously said. Both Yngve (1970) and Duncan (1972)'s work emphasized the role of backchannels in regulating turn-taking.

Others however, such as Schegloff (1982), have rejected the term *backchannel*.

Terms	Publications
signals of continued attention	Fries 1952
concurrent feedback	Krauss and Weinheimer 1966
accompaniment signals	Kendon 1967
listener responses	Dittmann et al. 1967, Dittmann et al. 1968, Fujimoto 2009
back channel	Yngve 1970, Maynard 1986
back-channel messages	Yngve 1970
back-channel signals	Duncan 1972, Duncan 1974
back-channel behavior	Duncan Jr et al. 1974, Duncan 1974, Maynard 1986
minimal responses	Zimmermann et al. 1975, Fishman 1978, Maltz et al. 1982, Young et al. 2004
back-channel responses	Krauss, Garlock, et al. 1977, Maynard 1986
vocalizations	Schegloff 1982
receipt object	Heritage 1984
acknowledgement tokens	Jefferson 1984, Drummond et al. 1993a, Drummond et al. 1993b, Zimmerman 1993
continuers	C. Goodwin 1986
assessments	C. Goodwin 1986, C. Goodwin and M. H. Goodwin 1987
carry-on signal	Stenström 1987
hearer signals	Bublitz 1988
backchannel	White 1989, Tottie 1991
backchannel messages	Tottie 1991
linguistic feedback	Allwood, Nivre, et al. 1992
aizuchi	White 1989, Clancy et al. 1996
marginal word	Du Bois et al. 1993
affirmative responses	Hirschman 1994
reactive tokens	Clancy et al. 1996
generic listener responses	Bavelas et al. 2000
specific listener responses	Bavelas et al. 2000
back-channel feedback	Ward and Tsukahara 2000
feedback response	Heldner, Edlund, and Hirschberg 2010, Inden et al. 2013, Meena et al. 2014

Table 1.1. – List of terms used to refer to feedback responses.

Schegloff (1982) preferred to use the term *continuer* to refer to vocalizations such as "uh-huh", "mhm", and "yeah". He defined these listener vocalizations by their sequential context, arguing that listeners passed the opportunity to take a turn and to initiate a repair. Schegloff (ibid.) contended that, rather than signaling attention, interest, or understanding, these vocalizations claimed those attributes. In a sense, he defined these vocalizations by their action or lack thereof. Additionally, he pointed out a distinction, noting that vocalizations such as "wow" were *assessments*. Similarly, C. Goodwin (1986) distinguished continuers e.g., "mhm" from assessments e.g., "oh wow". Unlike continuers, which simply express that the speaker should continue speaking, assessments are responses to something specific that the interlocutor just said. This continuer/assessment distinctions corresponds to Bavelas et al. (2000)'s *generic/specific response* distinction. Due to this distinction, the definition of backchannels has become synonymous to that of a continuer, rather than the one defined by Yngve (1970).

Likewise, Clancy et al. (1996) makes this distinction in their categorization of *reactive tokens*, which they define as "short utterance produced by an interlocutor who is playing a listener's role during the other interlocutor's speakership." Clancy et al. (ibid.) describe five kinds of reactive tokens:

1. Backchannels: These reactive tokens are based on Schegloff (1982)'s definition and serve as continuers that claim interest or understanding.
2. Reactive expressions: These reactive tokens are the equivalent to assessments.
3. Collaborative finishes: These reactive tokens complete the utterance of the interlocutor who has the speaking role; others have called these sentence completions.
4. Repetitions: These reactive tokens repeat portions of what the interlocutor who has the speaking role has said.
5. Resumptive openers: These reactive tokens are equivalent to *acknowledgement tokens* (Jefferson 1984; Drummond et al. 1993a; Drummond et al. 1993b; Zimmerman 1993). They are reactive tokens occurring in turn initial position and are part of a turn, for example, "**Oh**, well you might have better luck next time."

In contrast, the definition of *back-channel feedback* proposed by Ward and Tsukahara (2000) excludes some types of reactive tokens and other definitions of backchannels. Back-channel feedback is defined from the perspective of the interlocutor producing it, stating that a back-channel feedback "(D1) responds directly to the content of an utterance of the other, (D2) is optional, and (D3) does not require acknowledgment by the other." (D1) excludes sentence completions, (D2) excludes answers to questions, and (D3) excludes clarification requests and resumptive openers. Due to the many definitions of backchannels, Fujimoto (2009) has called for the abandonment of the term and proposed that *listener response*, *listener token*, or *reactive token* be used instead.

Overall, there is a lack of consensus when it comes to defining listener signals. Definitions are often based on vocalized listener signals while ignoring visual ones,

which is why we prefer using the term *feedback response*, as it encompasses both. Even for vocalized listener signals, there is no agreement on whether these should be limited to short vocalizations such as "mhm", "yeah", "uh-huh", or include longer utterances. The length of these utterances affects whether feedback responses are considered a turn or not. However, feedback responses are often described as not claiming the floor (Kendon 1967; White 1989; Drummond et al. 1993a; Clancy et al. 1996). Moreover, researchers seldom provide an exhaustive list of the lexical forms (Fries 1952; White 1989; Tottie 1991; Clancy et al. 1996; Ward and Tsukahara 2000), and often collapse variations like "yah", "yeh", and "yes" into a single category, "yeah". We have compiled lists of lexical forms mentioned in various publications, which can be found in Appendix A, note that there are many different spellings for the same vocalization.

Many terms also define listener signals by their communicative function which can be many. Although Fujimoto (2009) listed more general terms, she did not propose the term *feedback responses*, the term we prefer. The term *feed-back* originates from the field of cybernetics (Wiener 2019), where it is used to describe processes in "which a control unit gets information about the effects and consequences of its actions". Similarly, listeners send vocal and/or visual signals, and the speaker receives this information to monitor how it affects their ongoing talk. For instance, feedback responses from a distracted listener has been shown to negatively affect the quality of narration of the speaker's close-call story (Bavelas et al. 2000).

As previously mentioned, we prefer the term *feedback responses* because it encompasses both vocal and visual responses and because the term does not restrict it to one single communicative function. Throughout this thesis, we will use the term *feedback response* to refer to both short utterances such as "uh-huh" and "okay" as well as longer ones such as "oh wow". However, when we specifically want to highlight that we are referring to single-word utterances, we will use the term *short feedback responses*. It is important to note that all of our annotations and synthesized feedback responses are short feedback responses.

1.2. Communicative functions of feedback responses

Numerous communicative functions have been associated with feedback responses, as exemplified by the list of communicative functions in Table 1.2. However, there is no consensus on which specific communicative functions should be considered when studying feedback responses. Typically, feedback responses are described as, signaling attention, agreement, understanding, and indicating that the speaker should continue speaking. Often, these terms are based on a single communicative function, such as *continuer*. However, some researchers acknowledge that feedback responses are multi-functional. For example, Maynard (1986) observed five functions in their analysis of Japanese and English feedback responses, noting that this is not an exhaustive list: 1) continuer, 2) display of understanding of content, 3) support and empathy

toward the speaker, 4) agreement, and 5) strong emotional response. Additionally, while analyzing the differences between British and American feedback responses, Tottie (1991) proposed that feedback functions have two main functions that may occur together: 1) supportive function that signals understanding and agreement, and 2) a regulative function that encourages the speaker to continue their turn.

Given the numerous communicative functions cited in the literature with regards to feedback responses, this section will review coding schemes that have been proposed and used to annotate the communicative functions of feedback responses. These coding schemes have been applied to various languages, including Danish (Allwood, Cerrato, et al. 2007), English (Bavelas et al. 2000; Gravano, Beňuš, Hirschberg, et al. 2007; Beňuš et al. 2007), Finnish (Allwood, Cerrato, et al. 2007), French (Prévot, Gorish, et al. 2015; Prévot, Gorisch, et al. 2016; Boudin et al. 2021), German (Kopp, Allwood, et al. 2008; Buschmeier et al. 2011; Malisz, Włodarczak, Buschmeier, Skubisz, et al. 2016), Slovak (Beňuš 2016), and Swedish (Allwood, Cerrato, et al. 2007; Neiberg, Salvi, et al. 2013). Some schemes focus solely on vocal feedback responses (Bavelas et al. 2000; Gravano, Beňuš, Hirschberg, et al. 2007; Beňuš et al. 2007; Beňuš 2016; Bunt 2009; Buschmeier et al. 2011; Neiberg, Salvi, et al. 2013; Prévot, Gorish, et al. 2015), others on visual feedback responses (Allwood, Cerrato, et al. 2007), while others have been used to annotate multi-modal feedback responses (Kopp, Allwood, et al. 2008; Malisz, Włodarczak, Buschmeier, Skubisz, et al. 2016; Prévot, Gorisch, et al. 2016; Boudin et al. 2021).

1.2.1. Schemes based on Allwood, Nivre, et al. (1992)'s communicative functions

Many of these coding schemes (Allwood, Cerrato, et al. 2007; Buschmeier et al. 2011; Malisz, Włodarczak, Buschmeier, Skubisz, et al. 2016) are based on the four basic communicative functions of feedback responses introduced by Allwood, Nivre, et al. (1992), which are crucial for face-to-face communication. Table 1.3 lists the communicative functions and their definitions.

Similarly, the four basic communicative functions proposed by Allwood, Nivre, et al. (ibid.) correspond to the four levels of joint action by an addressee proposed by H. H. Clark (1994). Table 1.4 lists the levels of joint action and their description. Note that these joint actions focus on speech and therefore have different names compared to those in H. H. Clark (1996), see Table 1. Although Allwood, Nivre, et al. (1992) and H. H. Clark (1994) use different terms, *Contact* and *Attention*, as well as *Perception* and *Identification*, have similar definitions. However, the terms *Attitudinal reactions* and *Uptake* do not fully correspond, as H. H. Clark (ibid.) does not mention anything about attitudes. Nevertheless, *Uptake* partially corresponds to *Attitudinal reaction* because *Attitudinal reaction* is defined as the acceptance or rejection of the interlocutor, which can be thought of as uptake.

Using the four communicative functions proposed by Allwood, Nivre, et al. (1992), Allwood, Cerrato, et al. (2007) proposed the MUMIN coding scheme which was used to annotate feedback responses with three attributes; Table 1.5 lists the three attributes

Functions	Publications
acceptance	Allwood, Nivre, et al. 1992
acknowledgement	Neiberg and Gustafson 2012a
agreement	Kendon 1967, Yngve 1970, Maltz et al. 1982, Maynard 1986, Bublitz 1988, Tottie 1991, Ward and Tsukahara 2000
approval	Ward and Tsukahara 2000
attention	Fries 1952, Kendon 1967, Yngve 1970, Zimmermann et al. 1975, Fishman 1978, Maltz et al. 1982, Ward and Tsukahara 2000, Bavelas et al. 2000
attitudinal reactions	Allwood, Nivre, et al. 1992
belief	Allwood, Nivre, et al. 1992
boredom	Allwood, Nivre, et al. 1992, Ward and Tsukahara 2000
confirmation	Allwood, Nivre, et al. 1992
contact	Allwood, Nivre, et al. 1992
continue speaking	Zimmermann et al. 1975, Maltz et al. 1982, Schegloff 1982, C. Goodwin 1986, Maynard 1986, Bublitz 1988, Tottie 1991, Drummond et al. 1993b, Neiberg and Gustafson 2012a
disagreement	Neiberg and Gustafson 2012a
disappointment	Allwood, Nivre, et al. 1992
disbelief	Allwood, Nivre, et al. 1992
disinterest	Fishman 1978
empathy	Maynard 1986
enthusiasm	Allwood, Nivre, et al. 1992, Neiberg and Gustafson 2012a
following	Kendon 1967, Maltz et al. 1982, Bavelas et al. 2000
interest	Yngve 1970, Zimmermann et al. 1975, Maltz et al. 1982
judgement	Ward and Tsukahara 2000
listening	Maltz et al. 1982
non-acceptance	Allwood, Nivre, et al. 1992
perception	Allwood, Nivre, et al. 1992
regulative	Tottie 1991
skepticism	Ward and Tsukahara 2000
strong emotional response	Maynard 1986
support	Maynard 1986, Bublitz 1988, Tottie 1991
surprise	Allwood, Nivre, et al. 1992
sympathy	Ward and Tsukahara 2000
uncertainty	Neiberg and Gustafson 2012a
understanding	Zimmermann et al. 1975, Maynard 1986, Bublitz 1988, Tottie 1991, Allwood, Nivre, et al. 1992, Bavelas et al. 2000, Ward and Tsukahara 2000

Table 1.2. – List of communicative functions proposed in the literature of feedback responses.

Function	Definition
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.

Table 1.3. – The four basic communicative functions proposed by Allwood, Nivre, et al. (1992).

Level	Joint Action	Description
1	Vocalization and Attention	The speaker vocalizes an utterance and the addressee attends to the utterance.
2	Presentation and Identification	The speaker presents an utterance and the addressee identifies the utterance.
3	Meaning and Understanding	The addressee understands the meaning of the speaker's utterance.
4	Proposal and Uptake	The speaker proposes a project and the addressee takes up the project.

Table 1.4. – The four levels of joint action by H. H. Clark (1994).

(Basic, Acceptance, and Additional emotion/Attitude) and their descriptions. In the MUMIN scheme there is no separation between *Contact* and *Perception*. However, *Acceptance* which was part of the definition of Allwood, Nivre, et al. (1992)'s *Attitudinal reactions* have been separated.

Attribute	**Description**
Basic	This attribute has two features: (i) Continuation/contact and perception (CP) and (ii) Continuation/contact, perception and understanding (CPU).
Acceptance	This attribute is a boolean feature: which either indicates acceptance, the interlocutor perceived and understood the message and whether they agree with the content, or non-acceptance, they reject the content.
Additional emotion/Attitude	This category lists six emotions: Angry, Disgusted, Frightened, Happy, Sad, Surprised. However, 'other' is also added so as to not limit the list.

Table 1.5. – MUMIN scheme by Allwood, Cerrato, et al. (2007).

The MUMIN scheme has also inspired the function annotation scheme of Neiberg, Salvi, et al. (2013). Neiberg, Salvi, et al. (ibid.), performed two listening tests in which they investigated the relationship between the form and functions of feedback responses. In the first listening test participants were asked to annotate functional similarity between feedback responses, this comparison avoided asking the participants to annotate the functions. However, when participants were asked what they listened for, some identified certain functions such as interest, surprise, certainty, uncertainty, agreement, acceptance, confirmation, and boredom. In the second listening test participants were asked to rate on a 5 point likert scale the functions of the feedback responses:

1. Non-understanding – Understanding
2. Disagreement – Agreement
3. Uninterest – Interest
4. Expectation – Surprise
5. Uncertainty – Certainty

When all of the functions were rated on the center of the scale, the feedback response was assigned a neutral function, which corresponds to *Contact* and *Perception*. Rating the communicative functions of feedback responses on a scale is a unique feature of this annotation scheme. The attitudinal functions interest, surprise, and certainty were based on the first listening test while non-understanding and agreement were

based on the MUMIN annotation scheme. Although this scheme lacks an *Acceptance* function, it does have an *understanding* function and *agreement* function which were part of MUMIN's definition of *Acceptance*.

Allwood, Nivre, et al. (1992) have also inspired the annotation scheme proposed by Buschmeier et al. (2011). This annotation scheme has also been used by Malisz, Włodarczak, Buschmeier, Skubisz, et al. (2016). This annotation scheme contains three positive categories, corresponding negative categories, as well as modifiers; Table 1.6 lists the labels and their descriptions. Unlike the MUMIN scheme, this annotation scheme excludes the *Contact* function. Label A can be used as a modifier to annotate a feedback response such as P3A or it can be used as a label itself. Although this scheme does not have a category for *Contact*, P1 and P2 corresponds to MUMIN's *Basic* function and P3 corresponds to MUMIN's *Acceptance* function.

Label	**Description**
P1	The partner signals perception of the signal. *"I hear you and please continue."*
N1	The partner signals problems with perception. *"What are you saying?"*
P2	The partner signals perception and understanding of the message content. *"I understand what you mean."*
N2	The partner signals perception of the message and problems with understanding the content. *"I do not understand what you mean."*
P3	The partner signals perception, understanding and acceptance of the message or agreement with the message. *"I accept/agree/believe what you" say."*
N3	The partner signals perception and understanding but rejection or disagreement of the message. *"I disagree/do not accept what you say."*
A	The partner expresses an attitude towards the message, e.g., surprise, excitement, admiration, anger, disgust.
C	The partner introduces a new discourse segment or topic.
E	The partner ends the current discourse segment or topic.
?	Unresolved

Table 1.6. – Buschmeier et al. (2011)'s annotation scheme.

The modifiers C, and E that Buschmeier et al. (2011) used were adopted from Gravano, Beňuš, Hirschberg, et al. (2007)'s scheme. Gravano, Beňuš, Hirschberg, et al. (ibid.) developed a scheme for annotating the discourse functions of affirmative words: *alright*, *gotcha*, *huh*, *mm-hm*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yes*, and *yup*. Table 1.7 lists the discourse functions and their descriptions. These affirmative words share the lexical form with feedback responses, and therefore can be regarded as feedback responses. For example, affirmative words labeled as A1 and A2 are feedback responses. Some may also consider affirmative words labeled as P or F as feedback responses, e.g. "**Yeah**, okay I've met him before", depending whether feedback responses can be a full

turn. Similarly, Beňuš (2016) developed a scheme for annotating the discourse function of Slovak affirmative words: *áno*, *dobre*, *hej*, *no*, *mhm*, and *uhhuh*. Table 1.8 lists the discourse functions and their descriptions; note both Gravano, Beňuš, Hirschberg, et al. (2007)'s and Beňuš (2016)'s scheme were developed for task-oriented conversations (Columbia Games Corpus, SK-Games). In this scheme, discourse functions D, PH, R, RP, RZ, and S can be regarded as different types of feedback responses.

Label	Description
A1	**Acknowledgment/agreement.** Indicated *"I believe what you said"*, and/or *"I agree with what you say"*.
A2	**Backchannel.** Indicated only *"I hear you and please continue"*, in response to another speaker's utterance.
C	**Cue beginning discourse segment.** Marks a new segment of a discourse or a new topic.
E	**Cue ending discourse segment.** Marks the end of a current segment of a discourse or a current topic.
P	**Pivot beginning (A1+C).** Functions both to acknowledge/agree and to cue a beginning segment.
F	**Pivot ending (A1 + E).** Functions both to acknowledge/agree and to cue the end of the current segment.
N	**Literal modifier.** Example: **"I think that's okay".**
B	**Back from a task.** Indicates *"I've just finished what I was doing and I'm back."*
K	**Check.** Used with the meaning *"Is that okay?"*
S	**Stall.** Used to stall for time while keeping the floor.
?	Cannot decide.

Table 1.7. – Scheme for annotating the discourse functions of affirmative words by Gravano, Beňuš, Hirschberg, et al. (2007).

Although not explicitly inspired by Allwood, Nivre, et al. (1992)'s four communicative functions or H. H. Clark (1994)'s levels of joint actions, Bunt (2009)'s DIT++ taxonomy of communicative functions for dialogue contain two feedback functions *auto-feedback* and *allo-feedback* which provide information about attention, perception, interpretation, evaluation, and execution. The auto-feedback is used for when an agent provides information about their own processing, while allo-feedback is used for when an agent provides information about the other agent's processing (Bunt 1999).

1.2.2. Schemes based on generic/specific distinction

There are other coding schemes for the annotation of communicative functions of feedback responses that distinguish them into two categories. For example, Prévot,

Label	Description
D	Encourage some action, go on, do something
E	I want to repair/redo something I've just said or did
H	Hesitation, I am stalling for time
J	Softening of what is to follow, a hedge
K	Signal the end of the current topic or discourse segment
L	Literal modifier
N	I want to start a new topic or a new discourse segment
PH	Express assessment of something that just happened, usually on receiving a score
Q	Asking a question, checking
R	I acknowledge that I understand, I got it
RN	I acknowledge that I understand, and I want to start a new discourse segment
RP	I acknowledge that I understand, and please continue (Backchannel)
RZ	I acknowledge that I understand, but I want to add something or express mild disagreement
S	I agree, also as an answer to a question, usually meaning yes
Z	I want to express an idea opposite to implied before, usually meaning but or well
X	None of the labels correspond to the perceived meaning

Table 1.8. – Scheme for annotating the discourse functions of Slovak affirmative words by Beňuš (2016).

Gorish, et al. (2015) first categorized feedback responses into a *base function* (generic) which has other functions in this level. If the feedback response is categorized as *evaluation-base*, then the feedback response is further categorized in the evaluation function (specific). Table 1.9, lists the functions, its subcategories and descriptions. Note that the definition of *acknowledgment* in this scheme differs from the one defined in Gravano, Beňuš, Hirschberg, et al. (2007). In a follow-up work, Prévot, Gorisch, et al. (2016) used the same annotation scheme except, they removed the *Contact* subcategory and the *Evaluation* function was renamed to *Attitude*.

Base Function	**Description**
contact	I am still here listening.
acknowledgment	I have heard/recorded what you said but nothing more.
evaluation-base	I express something more than mere acknowledgment (approval, expression of an attitude,...).
answer	I answer to your question/request.
elicit	Please, provide some feedback.
other	This item is not related to feedback.
Evaluation	**Description**
approval	I approve vs. disapprove/ agree vs. disagree with what you said.
expectation	I expected vs. did not expect what you said.
amusement	I am amused vs. annoyed by what you said.
confirmation/doubt	I confirm what you said vs. I still doubt about what you said.

Table 1.9. – Prévot, Gorish, et al. (2015)'s annotation scheme.

Similarly, Boudin et al. (2021) proposed an annotation scheme, see Table 1.10, which also categorizes feedback responses into two main types: *generic* feedback responses and *specific* feedback responses. Within the specific category, feedback responses are further categorized by their *positive* or *negative* polarity and their *expected* or *unexpected* character. In this scheme, polarity and character are not defined. However, in their follow-up work, in which Boudin et al. (2024) uses a similar scheme, positive and negative are defined as reactions to whether the listener reacted to something positive or negative that the speaker said. Positive or negative refer to the speaker's utterance rather than the polarity of the feedback response. In the updated scheme, *expected* and *unexpected* are replaced with *given* and *new* labels; given and new refer to whether the information that the listener is hearing was talked about before, i.e., earlier in the conversation, or whether it is new information.

To summarize, coding schemes for annotating the communicative functions of feedback responses can be categorized into two main groups: those based on Allwood, Nivre, et al. (1992)'s four basic communicative functions and those that categorize feedback responses into two levels: generic and specific (Bavelas et al. 2000). Figure 1.1

Feedback type	Subcategories
Generic	
Specific	positive-expected positive-unexpected negative-expected negative-unexpected

Table 1.10. – Boudin et al. (2021)'s annotation scheme.

presents a review of the annotation schemes in literature divided by these two groups.

As can be seen, there is no standardized annotation scheme. For example, existing schemes use identical labels with different descriptions or different labels for the same descriptions. Additionally, these proposed schemes lack nuance, particularly when it comes to labels that convey attitudinal information. For instance, the feedback response "ugh", which signals disgust, would either have no category in the current coding schemes, or it would be grouped into a single category with other feedback responses such as "yeah" expressing agreement and "aww" expressing sympathy. Therefore, in Chapter 2, we propose our own annotation scheme which is based on the communicative functions identified in the Switchboard corpus.

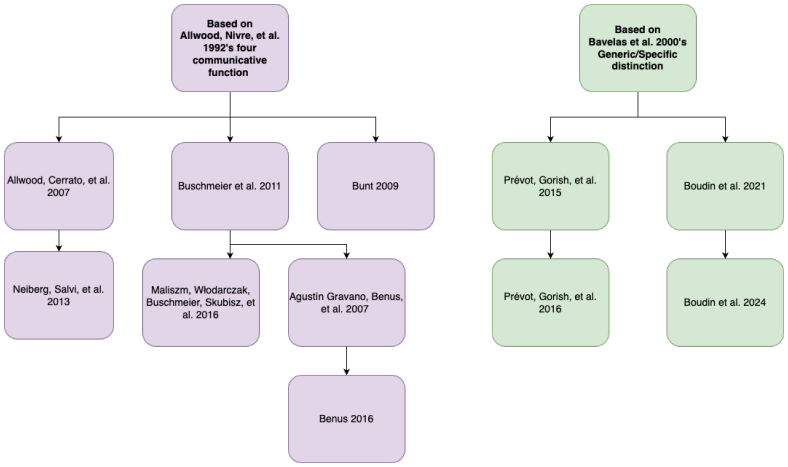


Figure 1.1. – Summary of annotation schemes for annotating feedback responses.

1.3. Relationship between prosody and communicative functions of feedback responses

The communicative function of feedback responses is partly conveyed by the lexical form and partly by the prosodic realization. In this section, we review studies that have investigated the relationship between prosody and the communicative functions of feedback responses. These studies can be categorized into three types: those that describe the prosody of feedback responses found in a corpus or analyzed prosodic features, perception studies in which participants listen to feedback responses and are asked about their communicative functions, and classification studies that use prosodic features to classify feedback responses into their corresponding communicative functions.

1.3.1. Prosodic description or prosodic analysis studies

In a study of non-lexical utterances in U.S. English spontaneous conversations, Ward (2004) and Ward (2006) established a mapping of pragmatic functions to certain prosodic features. Table 1.11 shows the mappings. Note, these non-lexical utterances include feedback responses, fillers and disfluency markers. Other researchers, however, have investigated the prosodic features of specific lexical forms.

sound	**meaning**
syllabification	lack of desire to talk
duration	amount of thought
loudness	confidence, importance
pitch downslope/upslope	degree of understanding / lack thereof
pitch height	degree of interest

Table 1.11. – Ward (2004) and Ward (2006)'s mapping of prosodic features and their meaning.

For example, in a corpus study of U.S. English task-oriented conversations, consisting of three conversations where participants were tasked with recreating a paperclip design, Hockey (1993) analyzed two lexical forms: *okay* and *uh-huh*. Hockey (ibid.) identified three distinct pitch contours for *okay* and two for *uh-huh*. Notably, one specific pitch contour–characterized by a flat or slightly falling first syllable and a rising second syllable–functions as a feedback response. This pitch contour, when used with *okay* and *uh-huh*, signals to the interlocutor that they should continue speaking.

In a larger corpus study of U.S. English spontaneous conversations, the Switchboard corpus, Jurafsky et al. (1998) annotated four different types of backchannels:

continuers, *assessments*, *incipient speakership*, and *agreements*. Jurafsky et al. (1998) observed that "some speakers used lexical form to distinguish agreement from continuers" while others used prosody. Although no statistical analysis was provided, Jurafsky et al. (ibid.) reported that continuers were shorter in duration, had lower pitch, flatter pitch contour, and lower energy than agreements.

Contrary to the findings of Jurafsky et al. (ibid.), Beňuš et al. (2007) observed different prosodic features for *backchannels* and *agreements* in the U.S. English Columbia Games corpus. Eight affirmative words: *alright*, *mhm*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yup* were annotated with the discourse functions listed in Table 1.7. Following the annotations, they performed a prosodic analysis to investigate the characteristics of the different discourse functions. Their analysis showed that backchannels (i.e., continuers) have similar duration to agreements. Additionally, they found that backchannels have higher pitch, higher pitch slope, and higher intensity than agreements.

In a subsequent study, Beňuš (2016) annotated affirmative words in the Slovak SK-Games corpus, a task-oriented corpus modeled after the Columbia Games corpus. Specifically, they annotated the affirmative words: *áno*, *dobre*, *hej*, *no*, *mhm*, *mhm*, *uh-huh* with the discourse functions listed in Table 1.8. However, for their prosodic analysis, they only used *áno*, *no*, *mhm*, and *uh-huh*, since there were only four instances of *dobre* and six of *hej* functioning as *backchannels*. They found that, similar to U.S. English backchannels, Slovak backchannels have a rising pitch. However, they also observed differences. For example, Slovak backchannels are longer than *acknowledgments*, which was not the case in the Columbia Games corpus. Additionally, they noted that Slovak *agreements* had higher intensity (were louder) than *backchannels*, which was the opposite result from that found with the affirmative words in the Columbia Games corpus.

Neiberg and Gustafson (2012b) created a dataset of feedback response tokens for Scottish English, conveying different feedback functions. A professional voice-over artist produced feedback responses given a dialogue context and the intended function. Additional feedback responses were recorded while the voice-over artist played chess and engaged in spontaneous conversations. This study investigated the correlation between six feedback functions– (*acknowledgment*, *continuer*, *disagreement*, *surprise*, *enthusiasm*, and *uncertainty*) – and prosodic features (mean pitch, standard deviation of pitch, pitch slope, mean intensity, and spectral center of gravity). In their prosodic analysis they found that certain feedback functions had distinct prosodic features. Feedback responses signaling *enthusiasm*, *surprise*, *acknowledgment* had the shortest durations, whereas *uncertainty*, *disagreement*, and *continuers* had longer durations. Regarding pitch, *enthusiasm* and *surprise* had a high mean pitch. Unlike other functions, continuers had a rising pitch, which had been previously observed in other studies (Hockey 1993; Beňuš et al. 2007; Beňuš 2016). Moreover, two pairs of functions–*disagreement* and *uncertainty*, as well as *surprise* and *enthusiasm*– shared similar prosodic patterns but could be distinguished by their mean pitch and spectral center of gravity, respectively.

In a German study of three feedback responses: *ja*, *m*, *mhm*, Malisz, Włodarczak, Buschmeier, Kopp, et al. (2012) investigated the prosodic differences between atten-

tive and distracted listeners. Three annotators categorized these feedback responses according to three feedback functions: *P1* (perception), *P2* (perception and understanding), or *P3* (perception, understanding, and acceptance), as listed in Table 1.6. They found that prosodic features–duration, pitch, pitch slope, intensity, intensity slope–were able to distinguish these communicative functions. Specifically, feedback responses from attentive listeners had higher intensity (they were louder), had less intensity variability and had more pitch variability.

Gustafson et al. (2010) analyzed the prosody of *mhm* and *mm* from a Swedish phone-in radio program where the feedback responses where produced by a Swedish radio host. In this study the two feedback responses were annotated for their level of engagement, either passive responses where the radio host signals that they are listening (*acknowledgment*), or active responses where the radio host signals that they are listening as well as signaling interest and encourages the interlocutor to continue speaking. They found that a falling pitch signals acknowledgment, whereas a rising pitch signals interest and encouragement. Interestingly, this study also takes into account the progression of the speech, they find that active responses tend to occur in the beginning of the conversation. As the conversation continues, the mean intensity and pitch of the feedback responses decreases and *mhms* become flatter.

The studies mentioned above investigated the relationship between prosodic features and their pragmatic functions. Others have investigated specific lexical forms and prosodic features, or the interactions between communicative functions and prosodic features. However, studies that examine the relationship between feedback functions and prosody often only compare two types: *continuers* and *assessments/ agreements*. In Chapter 2, we present our prosodic analysis of the feedback responses found in the Switchboard corpus, aiming to understand the differences in prosodic realizations across the different communicative functions in our annotation scheme.

1.3.2. Perception studies

Schröder (2000), Schröder (2003), and Schröder et al. (2006) conducted perception studies on German affect bursts (e.g., “yuck”, “oh”). Figure 1.2 illustrates the spectrum of affect bursts from Schröder (2003)’s study. Although not all affect bursts can be considered feedback responses, some, such as “oh”, “wow” and “yuck” can be categorized as such. In his initial study, Schröder (2000) had actors record affect bursts expressing ten emotions: *admiration*, *threat*, *disgust*, *elation*, *boredom*, *relief*, *startle*, *worry*, *contempt*, and *hot anger*. The study aimed to determine whether participants could recognize the intended emotional meanings of these affect burst when presented in isolation. Participants were asked to identify each affect burst with one of the ten emotions and also rate the affect burst on a seven-point scale across three dimensions: *excited-calm* (arousal), *positive-negative* (valence), and *dominant-subordinate* (control). The overall recognition rate of the intended meanings was 81.1%, although participants found it difficult to identify *threat* and *anger*.

In a continuation of the first study, Schröder (2003) asked half of the participants

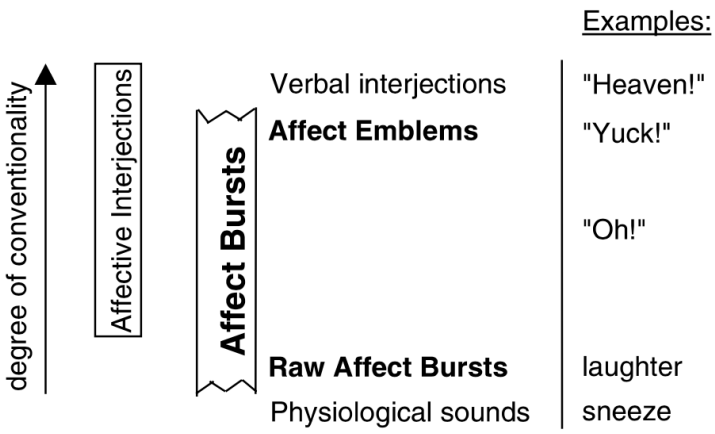


Figure 1.2. – Affect bursts spectrum by Schröder (2003). The illustration is from Schröder (2003).

from the listening test to write orthographic transcriptions of the affect bursts. New participants were then asked to identify the emotion of each written affect burst. The recognition rate in this perception study of written affect bursts was 65.1%, lower than that of the original listening test. Specifically, *threat*, *boredom*, *startle*, and *anger* had lower recognition rates than in the listening test. The results suggest that prosody plays a crucial role in identifying the meaning of affect bursts.

In the past two studies (Schröder 2000; Schröder 2003), affect bursts were not considered feedback responses. However, in their third perception study (Schröder et al. 2006), affect bursts were considered as feedback responses since they were assessed within a context. This study investigated whether affect bursts convey the same emotion in isolation as when presented within a neutral dialogue context. Both German and Dutch participants listened to this neutral sentence in their respective languages: "Yeah, then I told myself, why don't you try it <pause> and then I did it!" (English translation). The affect burst was inserted 150 ms into the pause, the pause itself was 750 ms. As in their previous perception study (Schröder 2000), participants were asked to select one of ten emotions and to rate the affect burst on the dimensions previously mentioned. Participants were presented with both conditions: affect bursts in isolation and affect bursts in context. There were differences in recognition rates between Dutch and German participants. Additionally, they found that context had only a small effect on the perception of the emotional meaning of the affect bursts. In other words, participants' perceptions of the meanings of the affect bursts in isolation compared to in context were similar.

Another study in which participants were asked to evaluate the meaning of vocalizations was conducted by Pammi, Schröder, and Charfuelan (2011). In this study, participants listened to feedback responses in isolation from four British actors, each portraying a distinct character: cheerful (Poppy), neutral (Prudence), gloomy (Obadiah), and aggressive (Spike). The feedback responses were recorded during spon-

taneous conversations, with each actor remaining in character. Participants were then asked to rate the feedback responses on a 1-5 scale for emotions: *anger*, *sadness*, *amusement*, *happiness*, *contempt*, *solidarity* and *antagonism*. Additionally, they rated the feedback responses on a -2 to 2 scale across the following dimensions: *uncertain-certain*, *disagreeing-agreeing*, *disinterested-interested*, and *unaware-anticipated*. Participants also had the option to select "No real impression" for each question. The study found that feedback responses are inherently ambiguous. For instance, the vocalization "aha" could convey five different meanings: *solidarity*, *certain*, *agreeing*, *interested*, and *anticipation*. Their results also highlighted the importance of prosody in conveying the meaning of feedback responses.

Neiberg, Salvi, et al. (2013) conducted two listening tests in a perception study of Swedish feedback responses. The feedback responses were extracted from face-to-face conversations in the Spontal corpus (Edlund, Beskow, et al. 2010). These feedback responses were clustered based on their prosodic distances. Rather than asking participants to annotate the feedback responses with a communicative function, they were asked to judge the similarity of a reference feedback response (a prosodic centroid) compared to four other feedback responses, thereby indirectly assessing their perceived communicative functions. In the first listening test, participants were presented with two scenarios: a "within-category" condition, where the reference feedback response and the four other responses shared the same lexical form, and a "between-category" condition, where they differed in lexical form but shared the same phonological operations (e.g., *m* and *a*, *mm* and *aa*, *mhm* and *aha*). Participants could choose one, many, or none of the feedback responses and were also asked to describe what they were listening for in their decision-making process. Participants listed perceived functions such as *interest*, *surprise*, *certainty*, *uncertainty*, *agreement*, *acceptance*, *confirmation* and *boredom*. Some of these functions were then used in the second listening test. In this test, participants listened to feedback responses in isolation and rated them on a 1-5 point Likert scale based on five communicative functions: *non-understanding - understanding*, *disagreement-agreement*, *uninterest-interest*, *expectation-surprise*, *uncertainty-certainty*. Additionally, Neiberg, Salvi, et al. (2013) visualized the feedback responses and performed a statistical analysis of the prosodic features, finding results consistent with other studies. For example, feedback responses signaling *surprise* had long durations, high mean pitch, and large pitch standard deviation. Feedback responses signaling *interest* also had long durations and high mean pitch, while those signaling *certainty* were short in duration and had a falling pitch slope.

Neiberg and Gustafson (2012a) also conducted two perception studies using Scottish feedback responses. A voice actor was asked to produce feedback responses with specific intended communicative functions, and additional feedback responses were extracted while he played chess and engaged in spontaneous conversations. In the first perception study, participants were presented with feedback responses such as *ah*, *mhm*, *m-m*, *n-hn*, *oh*, *okay*, *u-hu*, *yeah*, and *yes*. They were then asked to assign one of the following communicative functions: *acknowledgment*, *continuer*, *disagreement*, *enthusiasm*, *surprise*, and *uncertainty*. The results showed that participants struggled

to differentiate between *surprise* and *enthusiasm* as well as between *uncertainty* and *disagreement*. Additionally, the results showed that participants were biased by the lexical form of the feedback response. For example, "ah" and "oh" were often categorized as *surprise* even when the intended function was *uncertainty* or *enthusiasm*.

In the second study, Neiberg and Gustafson (2012a) focused exclusively on the perception of the feedback response "okay". Participants were presented with three questions:

1. How typical is the expression for [category] in feedback? (1-10)
2. If someone want to express [category] in feedback, how effective would this vocalization express [category]?
3. Is this expression acted or spontaneous? (acted/spontaneous)

The category mentioned in the questions referred to the specific communicative functions being evaluated. The results showed that participants were unable to reliably distinguish between acted and spontaneous feedback responses. Additionally, the results showed that acted feedback responses were perceived as conveyed the communicate function more effectively than spontaneous feedback responses.

Other perception studies have also investigated how prosodic realization affects the perception of specific lexical forms in feedback responses. For example, Lai (2009) conducted a perception study on the words "really" and "right" to understand how prosody influences their interpretation in expressing either *questioning* or *surprise*. Lai (ibid.) extracted *rights* and *reallys* from the Switchboard corpus. Before the perception study, each *really* was annotated with four labels describing "whether the speaker sounded like they had just heard something": (a0) unsurprising, (a1) new but not particularly unlikely, (a2) new and undesirable, (a3) highly improbable but not contradictory, and (r0) contradictory to their beliefs. The analysis revealed two distinct pitch contours for *really*: one with a peak in the first syllable followed by a fall or leveling of the pitch, and another with a continuous rise that sometimes plateaued at the end. Additionally, the variation in pitch range was found to correspond to different levels of surprise.

In the perception study, participants were presented with the *reallys* and *rights* in isolation and were asked two questions:

1. How surprised does the speaker sound?
2. How much like a real question does this sound like?

Participants rated their responses on a 7-point sliding scale, where 1=not at all and 7=extremely. The results showed that participants' perceptions did not strongly correlate with the annotations from the Switchboard corpus, where the items had been previously labeled as *backchannels* or *questions*. Additionally, the results showed that the lexical form influenced the participants' perception regardless of prosody; for instance, *right* and *really* were interpreted differently on the *question* and *surprise* dimensions even when they shared the same prosody. Specifically, the study also highlighted that pitch range is a key prosodic feature that cues *surprise*.

Unlike most of the perception studies mentioned so far, Gravano, Beňuš, Chávez, et al. (2007) investigated whether context influences the perception of the discourse functions of "okay". In this study, *okays* were extracted from the Objects Games conversations in the Columbia Games corpus. These *okays* were categorized by annotators according to their discourse functions: *acknowledgment/agreement*, *backchannel*, and *cue beginning discourse segment*. Participants were presented with two conditions: *okays* in isolation and *okays* in context. For both conditions, participants had to chose from the three discourse functions. The results showed that participants agreed more in the contextualized condition than in the isolated condition, suggesting that context is important for interpreting the word "okay". Specifically, they found that *acknowledgment/agreement* is the default function in both the isolated and contextualized conditions.

Gravano, Beňuš, Chávez, et al. (ibid.) also investigate which acoustic, prosodic, phonetic, and contextual features participants use to identify the discourse function of *okays* by conducting correlation measurements. In the isolated condition, participants were more likely to identify "okay" as a *backchannel* when it had lower intensity, longer duration, and a longer realization of /eɪ/. On the other hand, participants identified "okay" as a *acknowledgment/agreement* when it had a longer realization of /k/. In the contextualized condition, participants relied on different features, such as the latency of the interlocutor and the number of words before and after the "okay". However, in both conditions, the pitch slope at the end of "okay" was important. When the pitch slope was rising, participants identified the "okay" as a *backchannel*; when it had a falling pitch slope, it was identified as a *cue beginning discourse segment*. Gravano, Beňuš, Chávez, et al. (ibid.) posit that contextual cues may override the other cues relied on when hearing "okay" in isolation.

Recently, a new study examined the perception of *uh-huhs* and *yeahs*. Chandler (2023) crowdsourced both types of feedback responses from U.S. English native speakers using Amazon's Mechanical Turk. The voice talents were asked to produce the feedback responses with an intended communicative function: *agreement*, *belief*, *concern*, *interest*, or *understanding*, and to vary the production on a three-point Likert scale representing degrees of inflection. For example, if the voice talent was asked to say "uh-huh" in a neutral tone, their instruction would say "After the beep say uh-huh like you are neutral towards what the speaker says" and the scale would display: neutral (highlighted), believe a little, and believe very much. In the perception study, new participants were asked to listen to the *uh-huhs* and *yeahs* in isolation. They were informed that they would hear someone responding to an unheard statement and were tasked with rating the intended communicative meaning. For instance, if the feedback response was intended to communicate *interest*, participants would see the following ratings on the screen: very uninterested, slightly uninterested, neutral, somewhat interested, and very interested. Additionally, participants rated their confidence in their responses, where the options were: no idea, not very sure, somewhat sure, fairly confident, and very confident. They found moderate correlation between the intended communicative function of the crowdsourced feedback response and the participants' perceived communicative function.

Chandler (2023) also analyzed 60 acoustic-prosodic features from pitch, energy, and duration. Their analysis showed that feedback responses conveying *agreement* and *belief* were indistinguishable. Overall, they found that duration and pitch were the primary prosodic features that conveyed the communicative function, with duration, pitch peak location, and pitch height being the most crucial features for distinguishing between functions.

Many of these perception studies have presented feedback responses in isolation, with only a few examining them in context. However, there is no consensus on whether feedback responses have inherent or prototypical communicative functions. For some feedback responses, it appears that the lexical form may carry inherent or prototypical communicative functions. It remains unclear whether people rely solely on the lexical form, the prosodic realization, the context, or a combination of all these factors. In our annotation of the feedback responses in the Switchboard corpus, we considered all three factors when assigning communicative functions (see Chapter 2).

1.3.3. Classification studies

In a study of the word "really", Lai (2008) investigated whether prosodic features can disambiguate between instances of "really" that function as short *questions* and those that function as *backchannels*. However, the distinction between "really" functioning as a backchannel or a question is not always clear-cut. Table 1.12 shows a dialogue where "really" functions as a question, while Table 1.13 provides an example of "really" functioning as a backchannel. Additionally, Table 1.14 illustrates a case where "really" functions both as a question and a backchannel. Lai (ibid.) extracted instances of "really" from the Switchboard corpus, which had previously been labeled as either backchannels (450 instances) or questions (130 instances). Pitch, intensity, and duration features were extracted to classify these instances of "really" as either *backchannels* or *questions*. Two classifiers were used: Linear Discriminant Analysis (LDA) and a Support Vector Machine (SVM). Both classifiers performed worse than the baseline, which classified all instances as *backchannels*. The results suggest that the prosodic features of duration, pitch, and intensity are not sufficient to distinguish between "really" functioning as *backchannels* and "really" functioning as *questions*.

Similarly to Lai (ibid.), Heylen et al. (2007) and Truong and Heylen (2010) focused on a single lexical form: "yeah". Heylen et al. (2007) extracted *yeahs*[1] from the AMI Meeting corpus (Kraaij et al. 2005), which consists of multiparty conversations involving both native and non-native English speakers. In their annotations *yeahs* were tagged with the following dialog acts: *assessment* (407 instances), *backchannel* (i.e., continuer) (348), *stall* (224 instances), *inform* (95 instances) and *other* (48 instances). Note that *assessment* is a catch-all category for communicate functions conveying attitudinal information such as *agreements*, *disagreements*, *approval*, *disapproval*. Given that the conversations are multiparty, Heylen et al. (2007) used *conversational states*, represented as a 4-tuple <a, b, c, d>, where each letter represents one of the

1. Yeahs could occur in isolation or as part of a longer utterance.

Speaker	Dialogue
B	You like Lubbock better than Dallas
A	Yeah
B	Why?
A	Uh, because people are so much nicer
B	**Really**?
A	Yes
B	Well people are nice here in Dallas

Table 1.12. – Lai (2008)'s example of *really* functioning as a question.

Speaker	Dialogue
B	Oh I've got some Chinese Hollies that are just outrageous
B	They they are very sharp
A	Oh **really**
B	Do you do your own uh lawn maintenance?
A	Yeah

Table 1.13. – Lai (2008)'s example of *really* functioning as a backchannel.

Speaker	Dialogue
B	I kind of enjoyed that boat I looked at today
B	It's nice and clean
A	**Really**?
B	It wasn't [interrupted]
B	Yeah
A	Did it have a cabin?

Table 1.14. – Lai (2008)'s example of *really* functioning as a question and backchannel.

four participants. Five features were input to the classifier, they are listed in Table 1.15. They utilized a J48 decision tree from the WEKA toolkit to classify "yeah" into one of the categories: *assessment*, *backchannel*, *stall*, *inform* and *other*. Using the five features, the classifier achieved an accuracy of 57%, struggling to distinguish *backchannel* from *stall* and *assessment* from *inform*.

Feature	**Description**
lex	This feature has value 0 if the utterance consists of the word Yeah only. Otherwise 1.
continue	Has value 1 when the producer of the utterance also speaks in the next conversational state. Otherwise 0. This feature models incipient behavior of the backchanneler.
samespeaker	Has value 1 if the conversational state in which this utterance happens has the same speaker, but different from the backchanneler, as the next state. Otherwise 0. This feature indicates that there is another speaker that continues speaking.
overlap	There is speaker overlap in the state where the utterance started.
source	This involves the relation labeling of the annotation scheme. *source* refers to the dialogue act type of the source of the relation of the dialogue act that is realized by the Yeah utterance. If the yeah dialogue acts is not related to some other act the value of this feature is null. The possible values for this feature are: null, assess, inform, suggest, elicitation (which covers all elicitations), and other.

Table 1.15. – Heylen et al. (2007)'s features for classifying *yeah*.

Similar to Heylen et al. (2007), Truong and Heylen (2010) used instances of "yeah" from the AMI corpus and conducted two classification tasks. The first task aimed to classify instances "yeah" into one of three categories: *backchannel*, *assessment*, or *other*. The second classification task aimed to classify instances of "yeah" into one of two categories: *backchannel* or *assessment*. In contrast to Heylen et al. (2007), Truong and Heylen (2010) used prosodic features to disambiguate instances of "yeah". Table 1.16 lists the prosodic features used as inputs to the classifier. Truong and Heylen (ibid.) also utilized the same features from Table 1.15. They employed a C4.5 decision tree from the WEKA toolkit. The results showed that prosodic features improved the performance of the classifiers in both tasks. They also found that *assessments* generally had longer durations than *backchannels*, although no significant differences were observed among the other prosodic features.

Gravano, Beňuš, Hirschberg, et al. (2007) conducted several classification tasks using the affirmative words from the Columbia Games corpus: *alright*, *gotcha*, *huh*, *mm-hm*, *okay*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yes*, *yup*. Annotators labeled each affirmative word with a corresponding discourse function, descriptions of the discourse functions are listed in Table 1.7. For all classification tasks, they utilized the JRIP machine learning

Feature	Description
dur_z	duration over whole 'yeah' (z-score)
pitch_z	mean pitch over whole 'yeah' (z-score)
intens_z	mean intensity over whole 'yeah' (z-score)
absslopepitch_z	absolute slope of pitch over whole 'yeah' (z-score)
coef1, coef2	signs of 1st and 2nd degree polynomials, fitted on pitch values measured over whole 'yeah'
coef1_half, coef2_half	signs of 1st and 2nd degree polynomials, fitted on pitch values measured over 2nd half 'yeah'
pitch_half_z, absslopepitch_half_z	mean pitch and absolute slope of pitch measured over 2nd half of 'yeah' (z-score)

Table 1.16. – Truong and Heylen (2010)'s prosodic features for classifying *yeah*.

algorithm, WEKA's implementation of RIPPER. The inputs to the classifier included a combination of text-based, timing, word acoustic-prosodic, as well as previous-turn acoustic-prosodic features, all of which are detailed in Table 1.17.

Gravano, Beňuš, Hirschberg, et al. (2007) conducted three tasks to categorize all the affirmative words:

1. N vs all other discourse functions
2. (C, P, E, F) vs all other discourse functions
3. (A1, A2, P, F) vs all other discourse functions

Additionally, they performed classification tasks for each specific lexical form that had at least 50 instances, based on the discourse functions assigned by the annotators:

1. *alright*: A1, C, other
2. *mm-hm*: A1, A2
3. *okay*: A1, A2, C, F, P, other
4. *right*: A1, K, N
5. *yeah*: A1, A2

For all classification tasks, the feature combinations perform better than the baseline, which simply selected the majority class. Specifically, the text-only and contextual features yielded the best performances. Specifically the features that performed the best were the text-only and contextual features. In particular, the contextual features were able to achieve error rates close to that of the annotators. They also found that the acoustic features does not significantly improve the classifier's performance.

Neiberg and Gustafson (2010) conducted eight classification tasks using Swedish grunts from the DEAL corpus (Hjalmarsson 2008). The DEAL corpus comprises role-playing human-human face-to-face conversations, where one participant acted as a shopkeeper and the other as a shopper. The Swedish grunts included *fillers* and

Features	Description
text-based	identity of the word; part-of-speech tag of the target word and its preceding and following words; position of the target word in its Inter-Pausal Unit (IPU); position of the target word in its turn; IPU and turn length in words; whether the previous and following turn were uttered by the same speaker
timing	duration of the target word; duration of the IPU and turn in which the target word occurred; duration and type of overlap (if any) of the target word, and its containing IPU and turn with the interlocutor (none, complete, at the beginning, at the end); time elapsed to the target word from the beginning of its IPU and turn, in ms and as a percentage of the duration of its IPU and turn
word acoustic/ prosodic	ratio of voiced and unvoiced frames; minimum, maximum, mean, standard deviation of pitch and intensity; pitch slope, intensity slope, and stylized pitch slope, each calculated over the whole, or over its last 100, 200 and 300 ms; location of syllable boundary, and of the accented syllable for *okay* only
previous-turn acoustic/prosodic	for target words occurring in turn-initial IPUs, the same pitch and intensity features described above but this time calculated over the last IPU of the interlocutor's preceding turn

Table 1.17. – Gravano, Beňuš, Hirschberg, et al. (2007)'s features for classifying affirmative words.

feedback responses such as *mhm*, *eh*, *okej*, *ja*, *nej*, *a*, *nä*, *njo* and *jaä*. These grunts were categorized as either feedback responses or *fillers*. The feedback responses were further classified by annotators into three attitudes: *news receiving*, *dis-preference*, or *general feedback*. The eight classification tasks were as follows:

1. Fillers vs Feedback where all attitudes are merged
2. Feedback attitudes (3 classes)
3. Same as 2 but only for the Other Speaker condition
4. Same as 2 but only for the Same Speaker condition
5. The turn-taking effect of feedback (Other Speaker/Same Speaker/Simultaneous Start)
6. Same as 5, but with Simultaneous Starts removed
7. Same as 5, but Other Speaker and Simultaneous Start are merged
8. Same as 5, but Same Speaker and Simultaneous Start are merged

The term 'Other speaker' refers to the interlocutor who did not produce the grunt, while 'Same Speaker' refers to the interlocutor who did. The classification tasks were performed using linear discriminant analysis (LDA). The input features included timing features, with the grunt represented by suprasegmental prosodic signal representations based on Time Varying Constant-Q Cepstral Coefficients (TVCQCC). The context preceding the grunt was also represented using TVCQCC (left context). Figure 1.3, illustrates three of these timing features (a-c). The fourth timing feature was the duration of the gap or overlap; if the grunt occurred in complete overlap, then the duration of the grunt itself was used as the fourth feature.

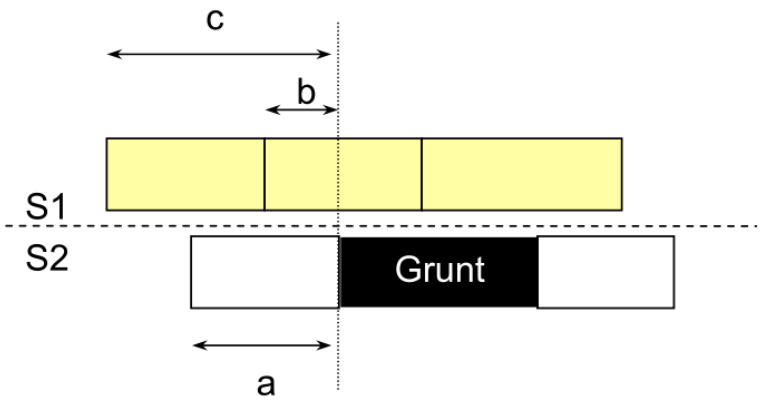


Figure 1.3. – Neiberg and Gustafson (2010)'s illustration of three timing features used in classifying Swedish grunts. The illustration is from Neiberg and Gustafson (2010).

For tasks 1-5, the feature for left context was not used. In all tasks, the classifier performed above random recall rates. The results demonstrate that by representing

the grunt using suprasegmental prosodic signal representations with or without timing features, the classifier distinguished more successfully between *fillers* and feedback responses. The timing and suprasegmental prosodic signal representations of the grunt also effectively differentiated between the three feedback attitudes. Overall, adding the timing features improved the classifier's performance in all tasks.

Neiberg and Gustafson (2010) also visualized the grunts in order to analyze their spectrograms. They found that feedback responses typically have a rising pitch, whereas *fillers* generally have either a flat or falling pitch. When examining the feedback responses for specific attitudes in the 'Other Speaker' condition, they noted that feedback responses expressing *dis-preference* were longer in duration compared to those conveying *news receiving* and *general feedback* attitudes. Additionally, there were notable differences in duration between feedback responses that were turn-initial and those that functioned as *continuers*[2]. Specifically, turn-initial position feedback responses were shorter in duration and had a final energy rise compared to *continuers*. Furthermore, differences in pitch were observed among the various feedback attitudes; *news receiving* feedback responses had the highest mean pitch and a strong pitch rise, whereas *dis-preference* feedback responses had the lowest mean pitch and weaker pitch rise.

In a two-step approach, Prévot, Gorish, et al. (2015) classified the communicative functions of French feedback responses extracted from two sources: the Conversation Interaction Data (CID) (Bertrand, Blache, et al. 2008) corpus and the task-oriented Aix-MapTask Remote (MTR) (Bard et al. 2013) corpus. Feedback responses included terms such as *ouais*, *mh*, *d'accord*, *laughter*, *oui*, *euh*, *uh*, *ok*, *ah*, and *voil*. Annotators classified these responses according to communicative functions listed in Table 1.9. Initially, annotators identified the base function of each feedback response, the options were: *contact*, *acknowledgment*, *evaluation-base*, *answer*, *elicit*, or *other*. If *evaluation-base* was selected, the annotators further classified the feedback response into one of the evaluation functions: *approval*, *expectation*, *amusement*, or *confirmation/doubt*.

Similarly, the automatic classification tasks followed this two-step approach. Although Prévot, Gorish, et al. (2015) experimented with various classifiers, such as Naive Bayes, Decision Trees, SVMs, Ada Boosts, and Random Forests, they report results from the Random Forest classifiers. All classifiers were compared against a baseline that simply selected the majority class. Prévot, Gorish, et al. (ibid.) also experimented with different input features for the classification tasks, which are listed in Table 1.18. Given that the study utilized two corpora, the classification results were analyzed separately for each corpus and for the combined data to determine which features performed better for each corpus.

For the base function classification task, the only classifier that performed worse than the baseline was the one that used POS features for classifying feedback responses from the MTR corpus. All other classifiers outperformed the baseline. For the second task, which involved assigning evaluation functions, the classifiers that performed

2. It is unclear if *continuers* and *general feedback* refer to the same feedback responses, as none of the attitudes were explicitly defined.

worse than the baseline were: (1) a classifier using only POS feaures for the CID corpus, (2) a classifier using only ACO features for the CID corpus, and (3) a classifier using only POS features for the combined data. These results indicate that certain features are more effective for specific corpora than for others.

Features	**Description**
LEX	transcription string, presence vs absence of frequent lexical markers
ACO	pitch (min/max/stdev/height/span/steepness/slope/NaN-ratio), intensity (quartiles Q1, Q2, Q3), average aperiodicity, formants (F1, F2, F3) and duration
POS	speech environment in terms of speech/pause duration before/after the item for both the speaker and the interlocutor; including overlap information
CTX-AUT	first/last tokens and bigrams of previous utterance and interlocutor previous utterance
MAN	function of the interlocutor's previous utterance (assert, question, feedback, try, unintelligible, incomplete)
META	Corpus, Speaker, Session, Role

Table 1.18. – Prévot, Gorish, et al. (2015)'s features for classifying communicative functions of French feedback responses.

In summary, these classification studies have focused either on classifying specific lexical forms into their corresponding communicative functions or on classifying different lexical forms into their corresponding communicative functions. However, the classification categories lack nuance due to the limitations of the annotation schemes, as previously discussed in Section 1.2. In Chapter 2, we present our classification study, where we classify all potential feedback responses (all lexical forms) in the Switchboard corpus into their respective communicative functions based on the functions outlined in our proposed coding scheme (see Chapter 2 for the coding scheme used to annotate the communicative functions of feedback responses).

1.4. Alignment and feedback responses

When feedback responses are described, one characteristic often mentioned is that they do not disrupt the flow of the conversation. It has been suggested that their unobtrusiveness is due to *prosodic alignment* (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Beňuš, et al. 2015b). But what exactly is *alignment*? In this section, we explain what alignment is, how it is measured, and review studies that have investigated prosodic alignment between feedback responses and the preceding turns of the interlocutor.

1.4.1. What is alignment?

The term *alignment* is often referred to by other names such as *accommodation*, *convergence*, *entrainment*, *mimicry* or *synchrony*, depending on the field of study (Wynn et al. 2022). These terms all describe the phenomenon where interlocutors become similar in behavior during conversation. For example, interlocutors may align in various verbal behaviors, such as phonetic (Pardo 2006; Olmstead et al. 2021; Huttner et al. 2023), acoustic-prosodic (R. Levitan and Hirschberg 2011; R. Levitan, Beňuš, et al. 2015a; S. I. Levitan et al. 2018), lexical (Friedberg et al. 2012; Weise et al. 2021; Kruyt, Beňuš, et al. 2022), and syntactic (Branigan et al. 2000; Hartsuiker et al. 2004; Allen et al. 2011). Alignment is also observed in non-verbal behaviors, such as breathing (Offrede et al. 2024), gestures (Holler et al. 2011; Louwerse et al. 2012; Mol et al. 2012) and facial expressions (McIntosh 2006; Y. Huang et al. 2016; Drimalla et al. 2019). There are two prominent accounts explaining why humans align.

The first account, Communication Accommodation Theory (CAT), proposed by Giles, Mulac, et al. (1987), posits that interlocutors, either consciously or unconsciously, adjust their behavior to "create, maintain, or decrease social distance in interaction" (Giles, Ogay, et al. 2007). Interlocutors use two main strategies to change their social distance: they either *converge* (i.e., become more similar) by adapting behaviors of their interlocutor, or they *diverge* (i.e., become less similar) by emphasizing their differences (Giles, Coupland, et al. 1991). According to Giles, Ogay, et al. (2007), *convergence* can help interlocutors gain approval or improve communication effectiveness, whereas *divergence* can highlight differences (e.g., establishing different social identities) or encourage their interlocutor to change their behavior. For example, Akira, who is Japanese, may begin speaking softly and whispering to prompt his American friend Chase, who is speaking loudly, to be less conspicuous on a train in Japan.

The second account, the Interactive-alignment model (IAM), proposed by Pickering et al. (2004), posits that during conversations interlocutors align their linguistic representations through priming. Priming is the process of activating linguistic representations upon the exposure to a stimulus, such as speech. Pickering et al. (ibid.) argues that the process of aligning is automatic and unconscious. Moreover, Pickering et al. (ibid.) states that alignment percolates, meaning alignment at lower levels lead to alignment at higher linguistic levels, the linguistic levels are shown in Figure 1.4. However, this may not be the case, according to Ostrand et al. (2021). This model assumes that the processes of production and comprehension are tightly coupled due to the use of the same representations, in contrast to the traditional view that treats production and comprehension as isolated processes. For example, in the isolated view, a speaker will first encode their message into a linguistic representation and then articulate the sounds. Whereas, the listener decodes the sounds by converting them into linguistic representations. However, according to IAM, if interlocutor A pronounces "pizza" as \piːksə\instead of \piːtsə\, this activates interlocutor B's linguistic representation and primes interlocutor B to then to pronounce "pizza" as \pɪksə\. This activation of representations is priming. Note that with IAM, it does not necessarily

mean that interlocutor B will imitate interlocutor A's pronunciation, although it makes it likely that they will do so.

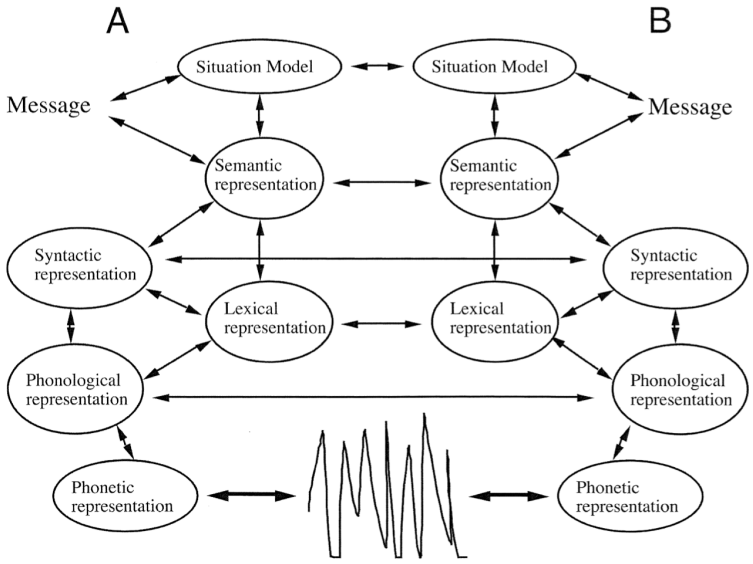


Figure 1.4. – Pickering et al. (2004)'s diagram of the coupled production and comprehension processes of interlocutor A and interlocutor B according to the interactive-alignment model. The illustration is from Pickering et al. (2004).

These two theories, CAT and IAM, present opposing views on whether alignment is intentional or automatic. However, as argued by Coles-Harris (2017), these two theories of alignment may not be opposing but rather complementary. To illustrate this, let's consider two brief scenarios, as shown in Table 1.19 and Table 1.20. Maria is American with Mexican heritage, while Martijn is Dutch. In the first scenario (Table 1.19), from the IAM perspective, Maria might repeat Martijn's pronunciation without conscious thought. From the CAT perspective, she might be being polite, choosing not to pronounce "açai" as she would normally, due to her Mexican heritage, in an effort to be liked. In the second scenario (Table 1.20), from CAT'S perspective, Maria deliberately diverges to emphasize her social identity as a member of the Latin American community, thereby implicitly asserting that her group affiliation gives her greater authority on the pronunciation a Brazilian speaker would use. As noted by Pickering et al. (2006), non-alignment can "occur when interlocutors refer to the same entities but refuse to adopt the same terms"; Maria must make an effort to not align.

It has been suggested that feedback responses are unobtrusive to the flow of conversations due to prosodic alignment (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Beňuš, et al. 2015b). While it would be interesting to investigate whether prosodic alignment between the listener's feedback responses and the interlocutor's preceding turn is an automatic process, we focus on measuring it in order to model it in conversational agents. Following Rasenberg et al. (2020), we use the term "alignment" with an agnostic view to both theories. In Chapter 4 we measure prosodic alignment

Speaker	Dialogue
Martijn	Shall we have \əkaɪ\bowls for breakfast?
Maria	Ooh, yes let's get some \əkaɪ\bowls for breakfast!

Table 1.19. – Example of alignment.

Speaker	Dialogue
Martijn	Shall we have \əkaɪ\bowls for breakfast?
Maria	Ooh, yes let's get some \asai\bowls for breakfast!

Table 1.20. – Example of misalignment.

between feedback responses and the interlocutor's preceding and following utterance across all feedback functions in our annotation scheme.

1.4.2. Measuring alignment in speech

In this section, we will discuss the different types of alignment that are measured in speech. Edlund, Heldner, and Hirschberg (2009) proposed two ways of measuring alignment, which describe the ways in which people can become similar. Alignment can be measured in terms of *convergence*, where parameters (e.g., prosodic features such as pitch) start in different directions and become similar over time. Additionally, alignment can be measured in terms of *synchrony*, which captures similarity in relative terms, specifically how parameters change in relation to each other. Figure 1.5 illustrates convergence and synchrony, with each line representing an interlocutor.

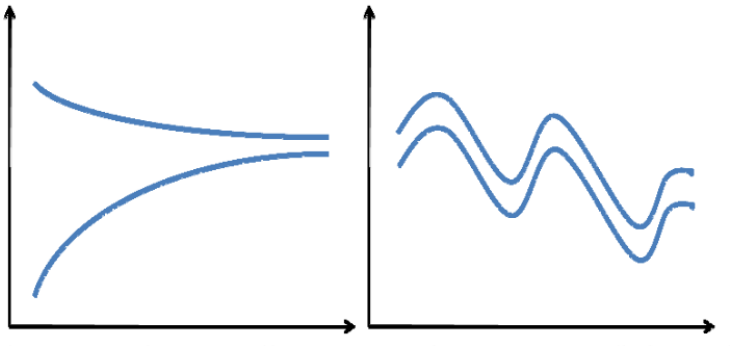


Figure 1.5. – Edlund, Heldner, and Hirschberg (2009)'s illustration of convergence on the left and synchrony on the right. The illustration is from Edlund, Heldner, and Hirschberg (2009).

R. Levitan and Hirschberg (2011) introduced a third way of measuring alignment, *proximity*. Unlike synchrony, proximity measures the absolute similarity between two interlocutors. For example, if interlocutor A raises their pitch to 220 Hz, interlocutor B might adjust their pitch to 190 Hz to match interlocutor A's pitch. If interlocutor A then lowers their pitch to 120 Hz, interlocutor B might alter theirs to 150 Hz to align again. Proximity measures absolute difference (e.g. 220 Hz - 190 Hz), whereas synchrony captures the dynamic changes (e.g. both interlocutors raise or lower their pitch). Figure 1.6 illustrates proximity, convergence, and synchrony, with different symbols representing each interlocutor.

R. Levitan, Beňuš, et al. (2015a) also distinguished between *global* and *local alignment*. Global alignment refers to changes that occur over a longer time period, such as throughout an entire conversation, while local alignment refers to changes that occur at turn exchanges. According to R. Levitan and Hirschberg (2011), convergence and proximity can be measured at both the global and local levels, but synchrony can only be measured at the local level.

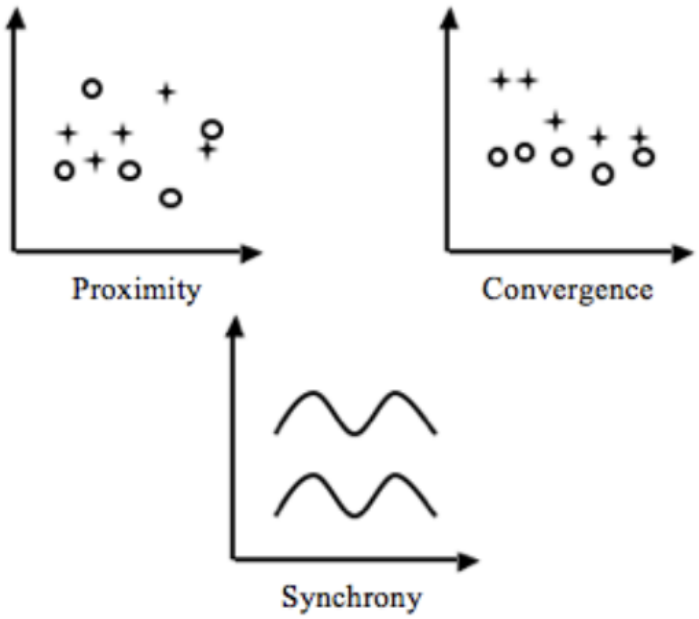


Figure 1.6. – R. Levitan and Hirschberg (2011)'s illustration of proximity, convergence, and synchrony. The illustration is from R. Levitan and Hirschberg (2011).

Building on the work of R. Levitan and Hirschberg (ibid.), Wynn et al. (2022) proposed a framework that expands upon the five types of alignment identified by R. Levitan and Hirschberg (2011) (global/local proximity, global/local convergence, and local synchrony). This framework was developed to encompass additional measures of speech alignment that were not covered in R. Levitan and Hirschberg (ibid.)'s original classification. The framework classifies alignment by three factors: *class*, *level*,

and *dynamicity*; Figure 1.7 summarizes the eight types of alignment based on the combinations of these three factors.

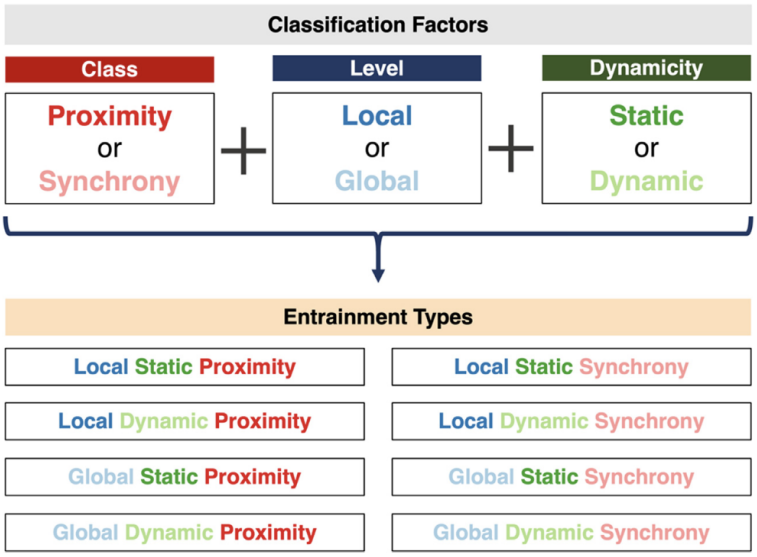


Figure 1.7. – Wynn et al. (2022)'s proposed framework for classifying alignment types. The illustration is from Wynn et al. (2022).

Class refers to either *proximity* or *synchrony*. Wynn et al. (2022) defines *proximity* as the "similarity of speech features between interlocutors". For example, if interlocutor A and interlocutor B both use a similarly fast speech rate, then there would be high proximity. Conversely, if interlocutor A uses a fast speech rate and interlocutor B uses a slower speech rate, this would indicate low proximity. Wynn et al. (ibid.) defines *synchrony* as the "similarity in movement (i.e., direction and magnitude of change) of speech features between interlocutors, regardless of the actual raw feature values of each interlocutor." For example, if interlocutor A increases their speech rate, and interlocutor B does the same—even though interlocutor B's rate remains slower than interlocutor A's—the parallel movement results in high synchrony, despite the low proximity. Figure 1.8 from Wynn et al. (ibid.) illustrates the different combinations of proximity and synchrony, demonstrating the distinctions between the two measurements.

Level refers to either *global* or *local*, indicating the timescale of measuring alignment. Wynn et al. (ibid.) defines *local alignment* as "similarity that occurs between units equal to or smaller than adjacent turns". Conversely, *global alignment* is defined as "similarity that occurs across any time scale greater than adjacent turns." For example, if pitch values are compared across turns that are not immediately adjacent to one another, this would be classified as global alignment, rather than local alignment.

Dynamicity refers to either *static* or *dynamic*, that is accounting for the change or lack of thereof of alignment over time in a conversation. Wynn et al. (ibid.) defines



Figure 1.8. – Wynn et al. (2022)'s illustration of high/low proximity/synchrony combinations. The illustration is from Wynn et al. (2022).

static alignment as "similarity without the statistical consideration of changes across time". For example, the proximity and synchrony of pitch between two interlocutors can be measured without assessing whether these attributes increase or decrease throughout the conversation. Conversely, Wynn et al. (2022) defines *dynamic alignment* as "a change in similarity across time". An example of dynamic proximity is if interlocutor A starts the conversation with a faster speech rate and interlocutor B with a slower one. As the conversation progresses, interlocutor A may slow down while interlocutor B speeds up, leading both to eventually converge at a mid-range speech rate by the end of the conversation. This type of change, known as *convergence* (Edlund, Heldner, and Hirschberg 2009; R. Levitan and Hirschberg 2011), is considered a form of proximity by Wynn et al. (2022), rather than a separate class. Dynamic synchrony, on the other hand, might occur if interlocutor A increases their speech rate while interlocutor B decrease their speech rate at the beginning of the conversation, but over time, they begin to synchronize their rates such that if interlocutor A changes their speech rate, interlocutor B mirrors this adjustment. Figure 1.9 illustrate all eight types of alignment in Wynn et al. (ibid.)'s framework.

A second motivation for Wynn et al. (ibid.) to propose this framework was the varied terminology used across different fields, which impeded the comparison of results related to alignment in speech. For instance, one study might measure global static synchrony while another assesses local static proximity, leading to different results under the umbrella of "prosodic alignment". By adopting the framework proposed by Wynn et al. (ibid.), a consensus on terminology is established, which helps make alignment research in speech more comparable across studies.

Note that in this section, we have not reviewed the methodologies for measuring these types of alignment due to their variety. However, Wynn et al. (ibid.) provides brief

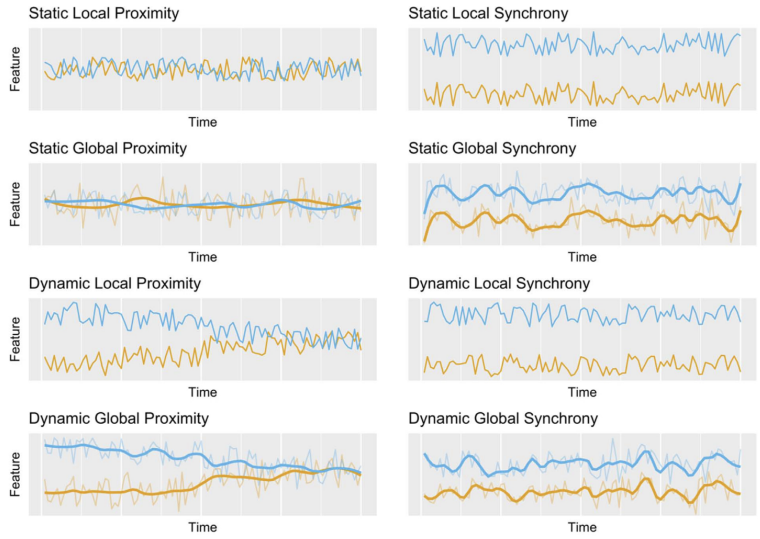


Figure 1.9. – Wynn et al. (2022)'s illustration of all eight types of alignment. The illustration is from Wynn et al. (2022). The bold lines represent similarity aggregated across larger portions of the conversation, the thinner lines represent similarity across adjacent turns.

descriptions and references for some methodologies used in measuring alignment in speech. Similarly, Kruyt, Jong, et al. (2023) reviews and applies 12 methodologies to measure prosodic alignment within a single corpus of conversations. In the next subsection, we will review studies that have investigated prosodic alignment between feedback responses and the interlocutor utterances that precede them, we will describe their methodologies and attempt to apply Wynn et al. (2022)'s classification framework.

1.4.3. Previous studies

Several studies have investigated prosodic alignment between feedback responses and the preceding utterances of interlocutors. Table 1.21 lists these studies and the types of alignment measured, according to Wynn et al. (ibid.)'s classification framework. We distinguish between studies conducted in U.S. English and those conducted in other languages.

1.4.3.1. Studies in U.S. English

Heldner, Edlund, and Hirschberg (2010) investigated alignment in terms of pitch between the preceding utterance of the interlocutor and the feedback response. They used the U.S. English Columbia Games Corpus, which had previously been annotated with different discourse functions of affirmative words (see Table 1.7 for all the labels). The type of feedback responses analyzed by Heldner, Edlund, and Hirschberg (ibid.)

Publication	Alignment type
Heldner, Edlund, and Hirschberg 2010	Static Local Proximity
R. Levitan, Gravano, et al. 2011	Static Global Proximity, Dynamic Global Proximity
R. Levitan, Beňuš, et al. 2015b	Static Local Proximity
Beňuš 2016	Static Local Proximity
Gustafson et al. 2010	Dynamic Local Synchrony
Kawahara, Yamaguchi, Uesato, et al. 2015	Static Local Proximity, Static Local Synchrony
Kawahara, Uesato, et al. 2015	Static Local Proximity, Static Local Synchrony
Ochi et al. 2024	Static Local Synchrony

Table 1.21. – List of previous studies investigating prosodic alignment between feedback responses and the preceding utterance of the interlocutor and corresponding alignment type they measured.

were *backchannels*, defined as responses "to another speaker's utterance that indicated only I'm still here / I hear you and please continue".

Heldner, Edlund, and Hirschberg (2010) compared the difference in pitch (i.e., pitch distance) between backchannels and the preceding utterances of the interlocutor (BC) to other turn transitions: smooth switches (S), pause interruptions (PI), and utterances following backchannels (X2). Smooth switches were defined as turn transitions where interlocutor A finishes speaking without overlapping speech from interlocutor B. Pause interruptions occurred when interlocutor A did not complete their utterance and there was no overlapping speech from interlocutor B. In their analysis, smooth switches (S) and pause interruptions (PI) were combined into one category (S+P1). X2 transitions were defined as speech from interlocutor A following backchannels from interlocutor B. Figure 1.10 visually illustrates these transitions. The mean pitch was calculated from the last 500 ms of the preceding turn and the first 500 ms of the following turn. An ANOVA test was used to compare the transition types, and revealed significant differences in pitch dependent on the type of transition. A Tukey HSD post hoc test revealed that all transition types were significantly different from each other. The results showed that the pitch distance between backchannels and the interlocutor's preceding utterance was smaller compared to other transitions (S+P1, X2), indicating greater pitch similarity between backchannels and the preceding utterances than between other turn transitions.

Heldner, Edlund, and Hirschberg (ibid.) also compared pitch distances between backchannels and the interlocutor's preceding utterance (BC), and the pitch distances between AFFCUE (the other discourse functions of affirmative words listed in Table 1.7) and the interlocutor's preceding utterance. Figure 1.11 illustrates these comparisons. This comparison was conducted because other discourse function of af-

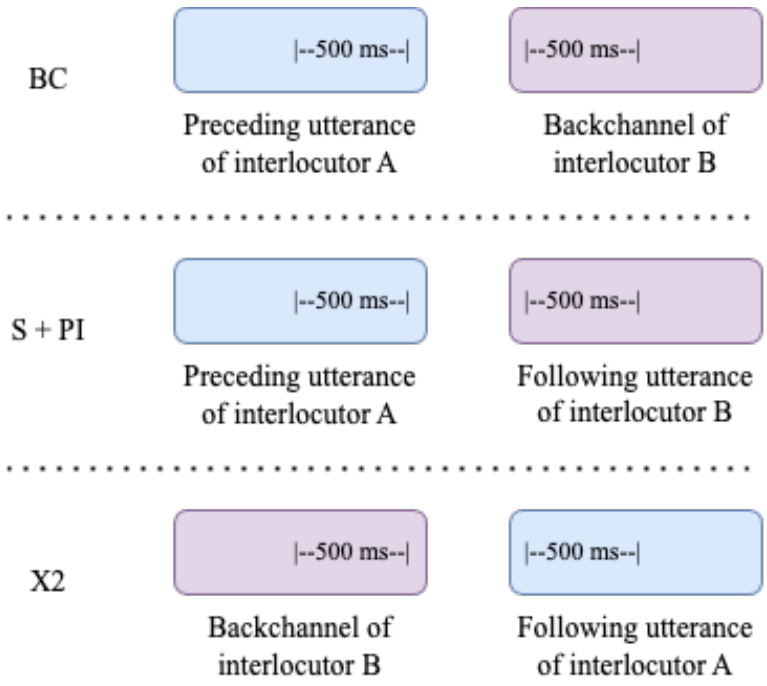


Figure 1.10. – Heldner, Edlund, and Hirschberg (2010)'s visualization of comparison of three transition types BC, S+P1, X2. The boxes represent the turns.

firmative words share some lexical forms with backchannels. An ANOVA test revealed significant differences in pitch between BC and AFFCUE. The results showed that backchannels were significantly more similar in pitch to the interlocutor's preceding utterance than other discourse functions of affirmative words. Heldner, Edlund, and Hirschberg (2010) suggested that the pitch similarity observed between backchannels and the interlocutor's preceding utterance might explain why they are unobtrusive, as the listener producing the backchannels attempts to match the pitch of their interlocutor's preceding utterance. These comparisons are classified as measuring static local proximity according to Wynn et al. (2022)'s classification framework.

Using the same corpus and annotations as in Heldner, Edlund, and Hirschberg (2010), R. Levitan, Gravano, et al. (2011) investigated prosodic alignment between backchannels and the preceding utterances of the interlocutors for the following acoustic-prosodic features: intensity, pitch, pitch slope, duration, and noise/harmonic ratio (NHR). Table 1.22 lists these features along with the lengths of the windows from which they were extracted. The features were extracted from inter-pausal units (IPUs), an IPU is defined as "a pause-free segment of speech from a single speaker, where pause is defined as a silence of 50 ms or more from the same speaker." Note that it is not clear if these features were extracted from the entire backchannel or from the windows listed in Table 1.22. R. Levitan, Gravano, et al. (ibid.) measured two types of alignment according to Wynn et al. (2022)'s classification framework.

The first alignment type measured was static global proximity. R. Levitan, Gravano,

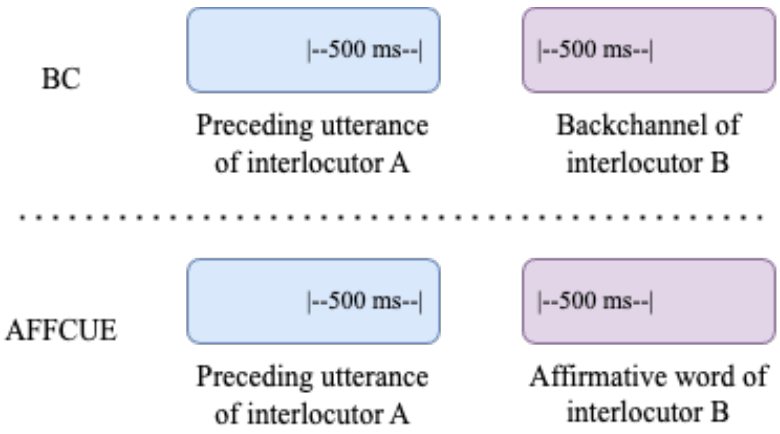


Figure 1.11. – Heldner, Edlund, and Hirschberg (2010)'s visualization of comparison of backchannels and affirmative words.

et al. (2011) compared the differences for each acoustic-prosodic feature between backchannel-preceding IPUs of conversational partners, and the average differences between each interlocutor and the other speakers in the corpus who were not conversational partners. Figure 1.12 illustrates these comparisons between conversational and non-conversational pairs. Results from paired t-tests indicated that the differences in intensity 500/1000 ms, pitch 500/1000 ms, and NHR 1000 ms were smaller among conversational pairs than non-conversational pairs.

Feature	**Description**
Intensity 500	mean intensity over the final 500 ms
Intensity 1000	mean intensity over the final 1000 ms
Pitch 500	mean pitch over the final 500 ms
Pitch 1000	mean pitch over the final 1000 ms
Pitch Slope 200	pitch slope over the IPU final 200 ms
Pitch Slope 300	pitch slope over the IPU final 300 ms
Duration	IPU duration in seconds
NHR 500	Noise/harmonic ratio over the final 500 ms
NHR 1000	Noise/harmonic ratio over the final 1000 ms

Table 1.22. – List of acoustic-prosodic features used in R. Levitan, Gravano, et al. (2011)'s alignment study.

The second alignment type measured was dynamic global proximity. R. Levitan, Gravano, et al. (2011) compared the mean differences for each of the acoustic-prosodic features between conversational partners from the first half of the conversation to the second half of the conversation. Figure 1.13 visualizes this comparison. The t-test

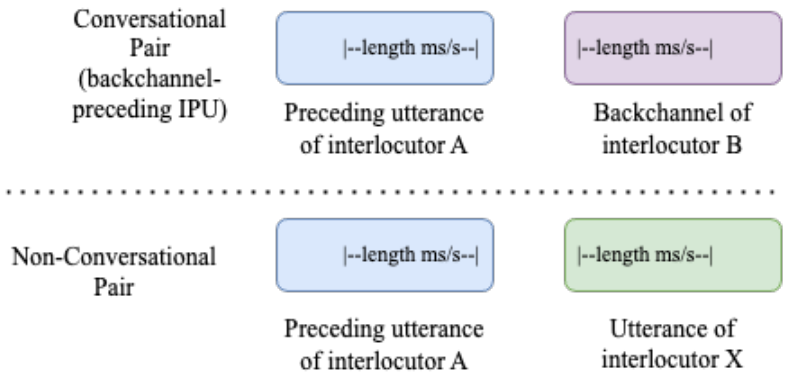


Figure 1.12. – R. Levitan, Gravano, et al. (2011)'s visualization of comparison of conversational pairs and non-conversational pairs. Interlocutor X are the speakers from the corpus that did not actually speak with interlocutor A. The boxes represent IPUs.

results showed that the mean differences in pitch and intensity, extracted from the 1000 ms window of the IPUs, were smaller at the beginning of the conversation than in the latter half. These results suggest that the interlocutor's alignment increased as the conversation progressed.

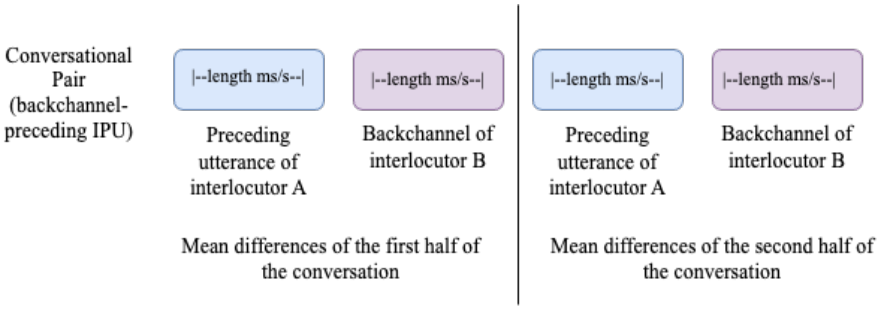


Figure 1.13. – R. Levitan, Gravano, et al. (2011)'s visualization of comparison between mean differences in the first half of the conversation vs. those in the second half of the conversation. The boxes represent IPUs.

In a subsequent study, R. Levitan, Beňuš, et al. (2015b) measured alignment, which would be classified as static local proximity according to Wynn et al. (2022) classification framework. The acoustic-prosodic features investigated included mean and max intensity, mean and max pitch, jitter, shimmer, NHR, and speaking rate. It is not specified whether these features were extracted from a specific time window or from the entire inter-pausal unit(s).

R. Levitan, Beňuš, et al. (2015b) compared the differences between backchannels and the preceding utterances (transition type BC, as shown in Figure 1.10) and the differences in smooth switches (similar to transition type S+P1, as shown in Figure 1.10).

Unlike the alignment measurements of these transitions in Heldner, Edlund, and Hirschberg (2010), R. Levitan, Beňuš, et al. (2015b) compared BC transitions with smooth switches that had the same latencies. R. Levitan, Beňuš, et al. (ibid.) define latency as "the difference between the start time of a turn and the end time of the previous turn". They matched the latencies of BC transitions because backchannels typically have shorter latencies than smooth switches. T-test results revealed significant differences in all acoustic-prosodic features except for speaking rate. The results indicated that the acoustic-prosodic features of BC transitions are more similar than those in smooth switches, suggesting greater alignment in BC transitions.

R. Levitan, Beňuš, et al. (ibid.) also compared the alignment of the acoustic-prosodic features between backchannels and the following utterance of the interlocutor (transition type X2, as shown in Figure 1.10) and smooth switches. They noted that the latency of X2 transitions is shorter than that of smooth switches. When latencies were not matched, t-test results showed that all features, apart from speaking rate, differed significantly, with smaller differences observed in X2 transitions compared to smooth switches. However, when latencies were matched for smooth switches to those in X2, t-test results revealed that only mean intensity, maximum pitch, jitter, shimmer, and NHR were significantly different from each other. For these acoustic-prosodic features, interlocutors aligned more in X2 transitions than in smooth switches. The results suggested greater alignment in X2 transitions than in smooth switches.

1.4.3.2. Studies in Slovak, Swedish, and Japanese

Similarly to the Columbia Games Corpus, Beňuš (2016) used the SK-Games corpus to investigate prosodic alignment between Slovak backchannels and the preceding utterances of the interlocutor. The SK-Games corpus was annotated with affirmative words, each assigned corresponding discourse functions, one of which was backchannels (see Table 1.8 for the full list of discourse functions). Specifically, the prosodic features investigated were pitch and intensity; these features were extracted from the final 500 ms of the interlocutor's preceding turn and over the entire length of the affirmative words. According to Wynn et al. (2022)'s classification framework, Beňuš (2016) measured static local proximity. They compared the similarity of features between backchannels and the preceding utterances of the interlocutor, and the similarity of features between other discourse functions of affirmative words and the preceding utterances of the interlocutor (see Figure 1.11). Their t-test results revealed no significant differences between the comparisons, suggesting that there was no prosodic alignment. However, visual analysis of the pitch contours showed that both backchannels and the interlocutor's preceding utterances typically had rising pitch—a pattern not observed in the other discourse functions of affirmative words that share some lexical forms with backchannels.

Using a corpus of a Swedish radio show, Gustafson et al. (2010) investigated prosodic alignment between feedback responses from the radio host and the preceding utterances of the callers. Specifically, they examined prosodic alignment in terms of intensity and pitch slope, which were extracted from the last syllable of the feedback

response and from the last syllable of the caller's preceding IPU. Feedback responses were annotated as either passive, acknowledging that the radio host was listening, or active, where the host expressed interest and encouraged the caller to continue speaking. However, these two types of feedback responses were not differentiated in the prosodic alignment analysis. Using a linear regression model, they predicted the intensity or pitch slope of the last syllable in the feedback response given the intensity or pitch slope of the last syllable in the preceding caller's IPU, with R^2 used to measure alignment. The feedback responses were categorized based on their occurrence at the start, middle, or end of the conversation. This type of method would be classified as dynamic local synchrony according to Wynn et al. (2022)'s framework. Gustafson et al. (2010) found evidence of alignment in terms of intensity, with alignment increasing as the conversation progressed. For pitch slope, they noted low R^2 values at the beginning, an increase in the middle, and a decrease towards the end of the conversation. They conclude that while intensity alignment might occur due to habit, pitch slope alignment is more challenging to detect and may be related to engagement, suggesting that as engagement wanes, so do the R^2 values.

Three studies have investigated prosodic alignment between feedback responses and the preceding utterances of interlocutors in Japanese. The first study, conducted by Kawahara, Yamaguchi, Uesato, et al. (2015), analyzed a corpus of counseling dialogue where the feedback responses were produced by two counselors and the preceding utterances by college students. The feedback responses were categorized into two types: acknowledging tokens (*u:n*, *un*, *un un*, *un un un*, *un un un un*) and reactive tokens (*a:* and *ha:*). The prosodic features investigated included max pitch, mean pitch, pitch range, max power, mean power, and power range. The values for pitch and power were extracted from the final 500 ms of the turn. However, the paper did not specify whether these prosodic features were extracted from the 500 ms of the feedback response or from the entire feedback response. Kawahara, Yamaguchi, Uesato, et al. (ibid.) measured two types of alignment: static local proximity and static local synchrony.

Static local proximity is measured by comparing the differences (i.e., distance) in features between the feedback responses of the counselor and the preceding utterance of the college student, and the difference between the preceding utterance of the college student and the following utterance of the counselor (which Kawahara, Yamaguchi, Uesato, et al. (ibid.) refer to as a normal turn). Kawahara, Yamaguchi, Uesato, et al. (ibid.) found no significant difference in terms of pitch between the two types of turns; however, a significant difference was observed in terms of power. Their results indicate that feedback responses are more similar to the preceding utterance than to normal turns in terms of power. Note that these findings differ from those in studies conducted in U.S. English, where evidence of alignment in terms of pitch was found (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Beňuš, et al. 2015b).

Static local synchrony is measured by performing correlations between the features of the feedback responses of the counselor and the preceding utterances of the college student. In this analysis, feedback responses are analyzed separately by lexical token and type (acknowledgement token, reactive token). The correlation results revealed

that there was more significant correlation in terms of power than pitch. Specifically, for reactive tokens, significant correlations were found in terms of both pitch and power. Pitch and power were significantly correlated for short acknowledging tokens: *u:n, un, un un*. However, only significant correlation was found in terms of power for the longer acknowledgement tokens: *un un un* and *un un un un*.

Using the exact same corpus and measuring the same alignment types as their previous work, Kawahara, Uesato, et al. (2015) extended their investigation of prosodic alignment. Their results for measuring static local proximity remained consistent with those reported in Kawahara, Yamaguchi, Uesato, et al. (2015). However, for static local synchrony, they expanded their analysis to include correlations between different prosodic features compared to the first study. Specifically, Kawahara, Uesato, et al. (2015) computed the correlation between the pitch of the college students and the power of the counselors, and vice versa, the power of the college students and the pitch of the counselors. Note the results from previous correlations (power vs. power and pitch vs. pitch) remained unchanged as they utilized the same dataset. However, with these new correlations between mixed prosodic features, significant correlations were found for *u:n, un*, and *un un un un*. For the reactive tokens, significant correlations were observed where the college students' pitch affected the power of the counselors' feedback responses. Note that in both studies, Kawahara, Yamaguchi, Uesato, et al. (2015) and Kawahara, Uesato, et al. (2015), the specific statistical tests were not reported.

The third study investigating prosodic alignment between Japanese feedback responses and the interlocutor's preceding utterance was conducted by Ochi et al. (2024). This study used a conversational corpus from (Inoue et al. 2020), in which participants interacted with the android ERICA in a Wizard of Oz setup. Human operators were instructed to produce feedback responses while the participants conversed with ERICA. Table 1.23 lists the prosodic features investigated, which were extracted from time windows of 1000, 2000, 4000, and 8000 ms of the participant's preceding utterance; however, Spearman's rank sum correlation results were reported only for the 8000 ms window. The correlations were categorized by three lexical forms of feedback responses: un, un un, and hai. Similar to Kawahara, Uesato, et al. (2015), Ochi et al. (2024) computed correlations between mixed prosodic features. Their results revealed significant correlations for the power features of feedback responses and the preceding utterances of the participants, with stronger correlations than those with pitch features. Unlike the U.S. English studies that showed evidence of prosodic alignment in terms of pitch (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Gravano, et al. 2011; R. Levitan, Beňuš, et al. 2015b), such similarity was not observed in Japanese. However, it is important to note that the U.S. English studies measured different types of proximity, whereas Ochi et al. (2024) measured static local synchrony as per Wynn et al. (2022)'s classification framework. Furthermore, Ochi et al. (2024) found that pitch features of the feedback responses were highly correlated with the power features of preceding utterance, contrasting the findings of Kawahara, Uesato, et al. (2015). Ochi et al. (2024) noted that, unlike in English, where pitch and power are correlated, in Japanese—being a pitch-accent language where pitch carries linguistic

information—there may not be such a correlation.

Statistic	F_0	power
Mean	$mean_f$	$mean_p$
Median	med_f	med_p
Standard deviation (SD)	SD_f	SD_p
Interquartile range (IQR)	IQR_f	IQR_p
Maximum value	max_f	max_p
Minimum value	min_f	min_p
Slope of the first-order regression line	$slope_f$	$slope_p$

Table 1.23. – List of prosodic features used in Ochi et al. (2024)'s alignment study.

To summarize, studies investigating prosodic alignment between U.S. English feedback responses and the preceding utterances of the interlocutor have been limited to a single task-oriented corpus, the Columbia Games Corpus. Visualizations of the pitch contours of Slovak feedback responses and their preceding utterances both show a rising pitch. This raises the question whether these task-oriented studies are measuring prosodic alignment or merely capturing prosodic patterns typical of question-answer pairs. Notably, none of the U.S. English studies have investigated prosodic alignment in spontaneous speech. Furthermore, these studies have not measured synchrony, unlike their Japanese counterparts. Moreover, many of these studies have focused on a single type of feedback responses—those that signal the interlocutor to continue speaking. Feedback responses that convey other communicative functions such as surprise, sympathy, or disagreement have not been addressed. Therefore, in chapter 4, we present our study investigating prosodic alignment (static local synchrony) between feedback responses and both the preceding and following utterances of the interlocutor. Our study explores prosodic alignment for all communicative functions found in the Switchboard corpus, a spontaneous speech corpus.

1.5. Modeling feedback responses in conversational agents

Corpus studies on human-human conversations have been conducted to understand and analyze the cues speakers produce that elicit feedback responses from listeners (Ward 1996; Koiso et al. 1998; Bertrand, Ferré, et al. 2007; Gravano and Hirschberg 2009; Truong, R. W. Poppe, IA de Kok, et al. 2011). These studies have provided valuable insights for researchers aiming to implement feedback responses in conversational agents. In fact, modeling feedback responses may be a growing area of research, especially with the increasing integration of conversational agents in our homes and daily lives.

In this section, we focus specifically on works that model vocal feedback responses. For studies that focus solely on modeling visual feedback responses, we refer the reader to the survey papers of Axelsson et al. (2022) and Iwan de Kok et al. (2012a). A quick search on Google Scholar with the keywords "modeling backchannels/feedback responses" or "predicting backchannels/feedback responses" yields various references. However, it raises questions: What aspects of feedback responses are modeled? What exactly are these prediction models predicting? We will review many of these works in this section, although our coverage may not be exhaustive.

1.5.1. Modeling vocal and visual feedback responses

Many studies have implemented both vocal and visual feedback responses in conversational agents. For example, using a rule-based approach based on the interlocutor's pause duration, Thórisson (1996) incorporated vocal and visual feedback responses in a humanoid named Gandalf, and Cassell et al. (1999) incorporated nods and the feedback response "mhm" in the virtual agent named REA. Kopp, Stocksmeier, et al. (2007) and Kopp, Allwood, et al. (2008) used a combination of rules and a Bayesian Network based on the part-of-speech information from the user's input text to incorporate various vocal and visual feedback responses in a virtual human named Max. Unlike these studies, R. Poppe, Truong, Reidsma, et al. (2010) applied a rule-based approach, using the speech and gaze of the user to generate "uh-huh" feedback responses, blinks, and nods in a virtual human. Additionally, Iwan de Kok et al. (2012b) used a Support Vector Machine to predict when a virtual agent should produce nods and vocal feedback responses, leveraging both auditory and visual information from the user. Further, Inden et al. (2013) used a probability distribution based on the user's vowel onset and rhythmic prominence to implement head nods and "um" feedback responses in a virtual human.

As observed, many of these studies incorporate only one form of vocal and visual feedback response. Some, such as Iwan de Kok et al. (2012b), do not specify which vocal feedback responses are included. Regarding vocal feedback responses, several studies (Cassell et al. 1999; R. Poppe, Truong, Reidsma, et al. 2010; Iwan de Kok et al. 2012b; Inden et al. 2013) also fail to mention whether they use pre-recorded or synthesized feedback responses. Furthermore, many of these studies do not report results from subjective evaluations (Cassell et al. 1999; Kopp, Stocksmeier, et al. 2007; Kopp, Allwood, et al. 2008; Iwan de Kok et al. 2012b), thus offering no insights into how human-like the agents' listening behaviors are perceived to be.

Perception studies have analyzed the combination of various visual and vocal feedback responses in virtual humans (Bevacqua et al. 2010; Oertel, Lopes, et al. 2016), while other research has focused on predicting the timing of these responses (Penteado et al. 2019; Blache et al. 2020) for both modalities. Notably, neither Penteado et al. (2019) nor Blache et al. (2020) performed subjective evaluations. Penteado et al. (2019) used a Random Forest model that used temporal features to determine whether there should be a feedback response, whereas Blache et al. (2020) used a rule-based approach based on prosody, discourse, syntax, and semantic features. However, in a

perception study where modalities of feedback responses were switched, R. Poppe, Truong, and Heylen (2011) found that changing vocal feedback responses and bimodal feedback responses to nods slightly increased the perceived inappropriateness. However, they reported a higher perceived inappropriateness when nods were switched to vocal feedback responses. Consequently, they suggested that future work should focus on predicting feedback response opportunities (i.e., timing) for specific types of feedback responses (vocal or visual).

Recently, Jain et al. (2021) used a neural network model to predict whether a visual feedback response, a vocal feedback response, or both should be used. They evaluated their visual feedback responses such as nods, head-shakes, smiles, and Hindi vocal feedback responses such as "hmm", "haan" ("okay" in English), and combination of the previously mentioned feedback responses using Apple's Memoji feature.

Often, modeling vocal feedback responses is equated with merely predicting the timing of when to insert the feedback responses. However, other important aspects include choosing the appropriate lexical form, determining the communicative function the conversational agent should convey which also requires generating the appropriate prosody. In the subsequent sections, we will review studies that present models predicting these various aspects of vocal feedback responses.

1.5.2. Predicting the timing of feedback responses

As mentioned, most research on modeling feedback responses has focused on predicting the timing. In other words, the primary question these models address is, "Should the conversational agent generate a feedback response at this point in the conversation?" Table 1.24 lists the publications, along with the type of models and features used to train these models. For the model types, we only make high-level distinctions, as many of the neural network-based (NN) models have multiple components. Similarly, for the input features of the interlocutor, we provide high level distinctions. For instance, if the studies used mean, minimum, and/or maximum of F_0 values, we simply list F_0. This approach is intended to provide a high-level overview of features commonly shared across different models.

In this section, we do not compare the performance of these models. Instead, we highlight the types of evaluations used and their implications. As Iwan de Kok et al. (2012a) note, comparing feedback response models is challenging due to different corpora, languages, and metrics used for evaluation. Information about the languages and corpus types used to train the models listed in Table 1.24 can be found in Table B.1 in Appendix B. Details about the metrics used for evaluating these models are provided in Table C.1 in Appendix C. Of all the models listed in Table 1.24, only three have been subjectively evaluated: (Mueller et al. 2015; Lala, Milhorat, et al. 2017; Ruede et al. 2017). The rest have been assessed using only objective measures such as precision, recall and F-scores, although not all models have been evaluated with all of these metrics.

One reason these models may not have been subjectively evaluated is the high cost in terms of both time and money involved in assessing various combinations of input

Publication	**Model type**	**Input features**
Ward and Tsukahara 2000	RB	F_0
Cathcart et al. 2003	RB	POS, pause duration
Solorio et al. 2006	LWLR	filter responses of F_0 and energy
Truong, R. W. Poppe, and Heylen 2010	RB	pause duration, pitch slope,
Sano et al. 2011	NN	F_0, power, gaze
Mueller et al. 2015	NN	F_0, power
Lala, Milhorat, et al. 2017	LR	F_0, intensity, pitch slope, silence durations, voice activity, overlap of speaker and listener
Ruede et al. 2017	NN	F_0, power, FFV, word2vec
Hara et al. 2018	NN	F_0, intensity, voice activity, spectral stability
Ruede et al. 2019	NN	F_0, power, FFV, MFCC, word2vec
Adiba, Homma, Bertero, et al. 2021	NN	F_0, power, MFCC, word embedding
Ishii et al. 2021	NN	VGGish embedding, BERT embedding, ResNET embedding from speaker and listener
Jinhan Wang et al. 2024	NN	HuBERT, GPT2 embedding, RedPajama embedding

Table 1.24. – List of publications predicting the timing of vocal feedback responses. Abbreviations: Rule-based (RB), Neural Network based (NN), Locally Weighted Linear Regression classifier (LWLR), Logistic Regression (LR), Part-of-speech (POS), Fundamental frequency variation (FFV), Mel-cepstral coefficients (MFCC).

features. Consequently, it is often quicker to evaluate multiple iterations of a model using objective measures. Another potential reason is that there are more opportunities for feedback responses than there are in the original conversations (Heldner, Hjalmarsson, et al. 2013). Feedback responses are optional, and different individuals might choose different placements for them, a variability often not accounted for by these models. Rather, the models are considered to predict correctly if they predicted the timing of the feedback response near to where they occurred in the conversation.

For the subjective evaluation, Mueller et al. (2015) had participants listen to segments of conversations from the Switchboard corpus. They inserted "mhm" or "uh-huh" feedback responses based on the model's predictions. However, it is unclear whether these responses were originally from the Switchboard corpus, synthesized, or if the same recording was repeatedly inserted. Participants rated whether the number of feedback responses was too few or too many and assessed the placement of the responses, including whether there were other possible placements that the model did not predict. They also evaluated the overall audio for naturalness versus artificiality. Similarly, Ruede et al. (2017) asked participants to listen to portions of Switchboard conversations and randomly selected feedback responses from the Switchboard corpus which were then placed in the conversations based on the timings of the model. Participants were asked to rate the overall naturalness of the recording and the appropriateness of the feedback response timing.

Unlike Mueller et al. (2015) and Ruede et al. (2017), Lala, Milhorat, et al. (2017) focused on having participants rate the appropriateness of the timing for each individual feedback response, rather than the entire conversation. Participants evaluated pre-recorded feedback responses in terms of coherence and timing in the conversations. As highlighted by Iwan de Kok et al. (2012a), definitions of backchannels (i.e., feedback responses) often vary. In a related context, Ortega, Meyer, et al. (2023), who proposed a model for predicting different types (i.e., the communicative functions) of feedback responses, noted:

> *On the one hand, the lumping approach focuses on BC placement, i.e. where within the speaker's speech the BCs occur. On the other hand, the splitting approach aims at explaining BC functionalities and categorizing them according to their type, function and discrete specific responses.*

However, this raises a question: have previous studies modeling feedback responses considered those conveying attitudinal information, such as "wow", "oh", and "aww", or have they simply only focused on feedback responses such as "mhm" and "uh-huh"? The review in the previous section on modeling vocal and feedback responses suggests that most have focused on the latter. In the next section, we will explore work being done on predicting the lexical form of feedback responses.

1.5.3. Predicting the lexical form of feedback responses

While predicting the timing of feedback responses has received considerable attention, less focus has been placed on predicting the lexical forms. Enabling con-

versational agents to generate varied lexical forms of feedback responses can make interactions less monotonous. Moreover, determining the lexical form is crucial because the communicative function of feedback responses is partly conveyed by the lexical form and partly by the prosody. For example, the lexical form "wow" can be interpreted as expressing surprise.

To the best of our knowledge, only three studies exist that address this issue. Table 1.25 lists these studies, detailing the type of model used, the input features, and the types of responses generated. All three studies were conducted in Japanese and included subjective evaluations. The objective metrics used by Kawahara, Yamaguchi, Inoue, et al. (2016) and Yamaguchi et al. (2016) are detailed in Table C.2 in Appendix C. The type of corpora used to train the models are described in Table B.2 in Appendix B.

Fujie et al. (2004) and Fujie et al. (2005) used a Gaussian distribution based on the user's F_0 and log power to determine when to generate a feedback response. They evaluated their models on the humanoid robot ROBISUKE. However, they did not provide information about the specific lexical forms predicted, nor did they describe the synthesis method used to generate the synthetic feedback responses.

Both Kawahara, Yamaguchi, Inoue, et al. (2016) and Yamaguchi et al. (2016) predict Japanese continuers *un*, *un un* (unx2), *un un un* (unx3). Additionally, Kawahara, Yamaguchi, Inoue, et al. (2016) includes an *assessment* category, defined as "backchannel types commonly used to express the listener's reaction, or assessment such as surprise, interest and empathy". Yamaguchi et al. (2016) includes a *reactive* category, defined as "backchannels, which react strongly to the new information, such as *wow* in English and *hee* in Japanese". It remains unclear whether, in their subjective evaluations, these categories represent a single lexical form or multiple forms from each category. This again highlights a recurring theme: all feedback responses conveying attitudinal information are grouped into a single category.

Unlike Fujie et al. (2004) and Fujie et al. (2005), Kawahara, Yamaguchi, Inoue, et al. (2016) and Yamaguchi et al. (2016) used pre-recorded feedback responses for each lexical form, resulting in a lack of prosodic variability. However, variability in both lexical form and prosody is important to prevent conversational agents from sounding monotonous. As noted earlier, to accurately convey the communicative function of feedback responses appropriate to the context, the prosody must also be considered. In the following section, we will review studies focused on predicting the communicative function of feedback responses.

1.5.4. Predicting the communicative function of feedback responses

Recently, there has been a growing body of work focused on predicting the type of feedback response (i.e., the communicative function). Essentially, these studies address the question: *"What communicative function should the conversational agent convey that is appropriate at this moment in the conversation?"*. The publications presenting models that predict the communicative function of feedback responses are listed in Table 1.26,1.27, and 1.28. These tables include the type of model, input

Publication	**Model type**	**Input features**	**Predicted categories**
Fujie et al. 2004, Fujie et al. 2005	FST	MFCC	Unknown
Kawahara, Yamaguchi, Inoue, et al. 2016	LR	type of boundary clause, number of buntsetsu phrase units, depth/width of parsed tree, ending word, POS ending word, previous clause boundary type, previous backchannel form, duration F_0, power creakiness	un, unx2, unx3, assessments not-to-generate
Yamaguchi et al. 2016	SVM	boundary strength, last word, POS last word, boundary syntactic-type label, number/depth/ width of phrase, boundary strength type before previous backchannel, morphological pattern of previous backchannel	un, unx2, unx3, reactive

Table 1.25. – List of publications predicting the lexical form of feedback responses. Abbreviations: Finite State Transducer(FST), Logistic Regression (LR), Support Vector Machine (SVM), Mel-cepstral coefficients (MFCC), Part-of-speech (POS).

features, and the categories (i.e., communicative function) predicted by each model. Note that the model types are categorized broadly into neural networks or traditional machine learning methods; for detailed descriptions of the architectures, see the respective references. The input features are described at a high level to highlight the commonalities across the models; for more specific details, refer to the corresponding references. For information on the language and corpus types used to train the models, see Table B.3 in Appendix B.

Publication	**Model type**	**Input features**	**Predicted categories**
Ortega, C.-Y. Li, et al. 2020	NN	word2vec, F_0, MFCC, loudness, voicing probability, listener embedding	Continuer, Assessment, No-BC
Adiba, Homma, and Miyoshi 2021	NN	word2vec, MFCC, PLP	Responsive interjections, Expressive interjections, non-BC
Boudin et al. 2021	LR	pause, tones POS, discourse markers, positive/negative valence, concreteness, laughs, smiles, nods	Generic, Specific
Jang et al. 2021	NN	KoBERT embedding, MFCC	Continuer, Understanding, Empathic Response, No BC
Lala, Inoue, et al. 2022	LR	F_0, power, % voiced window, % silences	Responsives, Expressives, Shared laughs

Table 1.26. – List of publications from 2020-2022 predicting the communicative functions of feedback responses. Abbreviations: Neural network based models (NN), Logistic Regression (LR), Mel-cepstral coefficients (MFCC), Perceptual Linear Prediction (PLP), Part-of-speech (POS).

Many models predict various categories or types of feedback responses (i.e., their communicative functions), and include an option for not generating a feedback re-

Publication	**Model type**	**Input features**	**Predicted categories**
Onishi et al. 2023	NN	head motion, gaze behavior, action units, intensity, MFCC, LPC, probability of voicing, F_0, POS	Neutral word, Positive word, Non-positive word, Emotional word, Anticipation, Confirmation, Repetition, Summary, Other
Ortega, Meyer, et al. 2023	NN	speaker embedding, listener embedding, speaker-listener encoding, MFCC	Yeah, Uh-huh, No-BC

Table 1.27. – List of publications from 2023 predicting the communicative functions of feedback responses. Abbreviations: Neural network based models (NN), Mel-cepstral coefficients (MFCC), Part-of-speech (POS), Linear predictive coding (LPC).

Publication	Model type	Input features	Predicted categories
Boudin et al. 2024	LR	pause, tone, speech rate, overlap, POS positive/negative valence, concreteness, nodding, laughter, smiles, duration since listener's previous feedback response, type of listener's response	Model 1: Specific, Generic; Model 2: Positive, Negative; Model 3: Given, New
Choi et al. 2024	NN	speech embedding	Continuer, Assessment, Non-BC
Park, Liermann, Choi, and Lee 2024	NN	wav2vec, BERT KorBERT, previous utterance embedding, speaker embedding	Korean Model: Continuer, Understanding, Empathic U.S. English Model: Continuer, Assessment NoBC
Park, Liermann, Choi, Kim, et al. 2024	NN	Same features as above, person/topic/ progress embedding	Continuer, Understanding, Empathetic, NoBC

Table 1.28. – List of publications from 2024 predicting the communicative functions of feedback responses. Abbreviations: Neural network based models (NN), Logistic Regression (LR), Part-of-speech (POS).

sponse, often labeled as No-BC, non-BC, or NoBC. However, studies such as those by Boudin et al. (2021), Boudin et al. (2024), and Lala, Inoue, et al. (2022) use external models to make this binary decision about whether or not to generate a feedback response. Unlike all the models listed in Table 1.28, Boudin et al. (2024) proposed four distinct models: one for the binary decision on whether a feedback response should be generated, one to determine if the response is *specific* or *generic*, another to classify if it is a response to a *positive* or *negative* context, and a fourth to decide if it is a response to a *given* or *new* information.

All of these models share the same overarching goal, but the categories they predict vary, as evidenced by the definitions provided in Appendix D. The variation in categories stems from the lack of consensus on what the communicative functions of feedback responses are (see Section 1.2) and the lack of consensus in definitions of the terminologies used for feedback responses (see Section 1.1). Many of these publications use the term *backchannel type*, although many would argue that responses such as "wow" are not to be considered backchannels.

All of these models have a common limitation: they typically group feedback responses containing attitudinal information into a single category. An exception is the model proposed by Ortega, Meyer, et al. (2023), which does not consider attitudinal responses at all. For example, feedback responses conveying disagreement, sympathy, or surprise would all fall under one category. This grouping could pose issues if these models were implemented in conversational agents, as they would struggle to generate contextually appropriate feedback responses. In fact, among all the proposed models, only Lala, Inoue, et al. (2022) has conducted a subjective evaluation of their model. The others have assessed their models solely by objective metrics such as Accuracy, F-scores, precision, and recall. A summary of the objective measures used for evaluating each model is available in Table C.3 in Appendix C.

In their subjective evaluation of appropriateness, Lala, Inoue, et al. (ibid.) found that their model, which generated *expressive* feedback responses, performed no better than a baseline model that predicted feedback responses randomly. This lack of improvement could be attributed to the feedback's communicative function being inappropriate for the context or possibly due to the prosody of feedback responses being inappropriate. Although Lala, Inoue, et al. (ibid.) noted that the feedback responses were generated using a text-to-speech (TTS) system with variations, the selections were made randomly. As the prosody of feedback responses is an important factor in the perceived appropriateness of dialogue speech, it is important to consider their prosody when conducting subjective evaluations. In the next section, we review works that predict the prosody of feedback responses.

1.5.5. Predicting the prosody of feedback responses

While there has been some work on predicting the prosodic features of feedback responses, such models are relatively rare. However, these models are crucial for accurately modeling prosodic alignment, as discussed in Section 1.4. Predicting the prosodic features of feedback responses is crucial for conveying the appropriate

prosody for the context, since the communicative function of feedback responses is partly conveyed by the prosody and partly by the lexical form. These models address the important question: *"How should the feedback responses of the conversational agent sound?"*.

Table 1.29 presents these publications along with the types of models used, the input features, the predicted features, and the objective measures used for evaluating the models. Details about the languages and corpus types used in these studies can be found in Table B.4 in Appendix B. Although Nath et al. (2022), predicted the prosody of discourse markers and not feedback responses, we include this study because the lexical forms overlap with those of feedback responses. Nath et al. (ibid.) extracted discourse markers based on the lexical form but did not distinguish between actual discourse markers and other forms (e.g., feedback responses).

All of these models have been evaluated only by objective measures and predict prosodic features at the "word-level". However, Nath et al. (ibid.), suggests that future studies should aim for more detailed predictions, such as at the "frame-level". Currently, the prosodic features of feedback responses are represented as discrete features, but it is feasible to represent this information using embeddings, as proposed by Qian et al. (2024). For subjective evaluations of models that predict the prosodic features of feedback responses, the feedback responses would need to be synthesized with the predicted prosodic features. Another alternative approach involves maintaining a database of feedback responses from which the response closest to the predicted or required features is selected (Campbell 2007; Oertel, Jonell, et al. 2017). However, this method is only feasible if there are multiple recordings of the same lexical forms, each with prosodic variations provided by the voice talent (the conversational agent's voice). But what work has there been on synthesizing feedback responses? We will explore this topic in the following section.

1.5.6. Synthesis of feedback responses

As mentioned previously, one approach to generating feedback responses with prosodic variability in conversational agents involves creating a database filled with numerous samples of feedback responses. From this database, the response that best matches the target prosodic realization for the given context is selected (Campbell 2007; Oertel, Jonell, et al. 2017). However, this approach is only feasible when a comprehensive database is available. Moreover, it may not be entirely effective if the goal is to have conversational agents adapt their prosodic features to those of the user—a phenomenon known as prosodic alignment. When only a few feedback response samples are available, synthesizing feedback responses becomes a more practical method for creating prosodic variability.

Synthetic feedback responses have been used in a few studies on modeling feedback responses (Fujie et al. 2004; Fujie et al. 2005; Kopp, Stocksmeier, et al. 2007; Kopp, Allwood, et al. 2008; Lala, Inoue, et al. 2022). However, these studies often do not specify the synthesis methods used to generate these feedback responses. More recently, feedback responses have been incorporated in Amazon's Alexa through the

Publication	**Model type**	**Input features**	**Predicted features**	**Evaluation metric**
Kawahara, Uesato, et al. (2015)	MF	F_0, power	mean/max log F_0/power	MSE
Nath et al. (2022)	MLR	log energy, lengthening, peak disalignment, creakiness, pitch lowness/highness, narrow/wide pitch, speaking rate	mean/max log energy, cepstral flux, pitch, harmonic ratio	RMSE
Ochi et al. (2024)	SVR	mean, median, SD, IQR, max, min of F_0/power, slope of first-order regression line	median, SD of power/F_0	MAE

Table 1.29. – List of publications predicting the prosody of feedback responses. Abbreviations: mathematical formula (MF), (LR), Multivariate linear regression (MLR), Support Vector Regression (SVR), standard deviation (SD), Interquartile Range (IQR), Mean Square Error(MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE).

development of an Alexa skill (Motalebi et al. 2019; Cho et al. 2022). While it is probable that these responses were synthesized using a text-to-speech (TTS) system, the use of pre-recorded, or "canned" responses remains a possibility.

One popular method for synthesizing feedback responses with variable prosodic realizations is through signal processing. In their study, Åsa Wallers et al. (2006) used the Multi-Band Resynthesis OverLap Add (MBROLA) (Dutoit et al. 1996), a diphone-based speech synthesizer, to explore the perception of Swedish monosyllabic feedback responses "m" and "a". They manipulated the pitch peak position, pitch peak height, and duration of each token. Figure 1.14 displays visualizations of these stimuli. In an initial study, the stimuli were inserted into the sentence "On my left I have a X house...", where X was replaced with one of the colors red, yellow, or blue. Participants were asked to interpret the meaning of the Swedish feedback responses, the interpretations are summarized in Table 1.30. In a subsequent study, new participants were presented with the same stimuli but were asked to choose among the potential meanings from the first study.

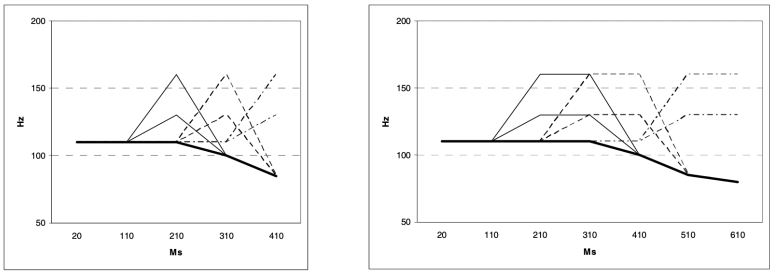


Figure 1.14. – Åsa Wallers et al. (2006)'s stimuli for Swedish feedback responses *m* and *a*. The left pane is for the shorter duration feedback responses and the right pane is for the longer duration feedback responses.

Interpretation
Good, you are in the right place.
Oh, NOW I understand where you are.
Really? That was unexpected.
Oh, you are in the wrong place.
Okay, but I need more information.

Table 1.30. – Åsa Wallers et al. (2006)'s perceived meanings of Swedish feedback responses *m* and *a*.

Also using MBROLA, Stocksmeier et al. (2007) synthesized variations of the German feedback response "ja". They manipulated the duration of the entire feedback response, the ratio of the phones \j\and \a:\, and modified the pitch contour into

different patterns, which are displayed in Figure 1.15. Participants listened to these feedback responses in isolation (i.e., without context) and rated them on a five-point scale across pairs of opposite traits: happy vs. sad, brave vs. anxious, certain vs. hesitant, approving vs. rejecting, pushing vs. not pushing, surprised vs. bored, angry vs. balanced. Additionally, participants were asked to write down their perceived meanings, revealing varied interpretations that highlighted the importance of context in understanding these responses.

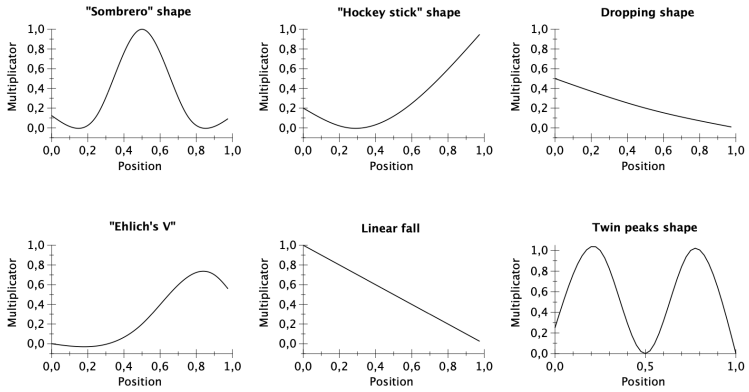


Figure 1.15. – Stocksmeier et al. (2007)'s pitch curves for the German feedback response *ja*.

In a perception study of U.S. English feedback responses, "uh-huh" and "yeah", Chandler (2023) used Amazon's Polly to synthesize these utterances. Using Praat (Boersma et al. 2024), they then manipulated the pitch and energy curves, as well as the duration. For the feedback response "uh-huh", the durations were changed per syllable, whereas for "yeah" the duration was changed over the entire utterance. Participants listened to the re-synthesized feedback responses in isolation and were asked to rate them on a five-point scale for the following communicative functions: agreement, concern, interest, and understanding.

Pammi, Schröder, Charfuelan, et al. (2010) and Pammi and Schröder (2011) extended the capabilities of MaryTTS, a unit-selection TTS system, to synthesize feedback responses with variable prosody for a female British vocal talent. This vocal talent recorded numerous feedback responses rich in attitudinal information, which were annotated with details on their lexical forms, meanings, intonation, and voice quality (see Table 1.31). Pammi, Schröder, Charfuelan, et al. (2010) and Pammi and Schröder (2011) modified the MaryTTS framework to accept MaryXML requests specifying target features for feedback responses, as listed in Table 1.31. The system then selected a suitable candidate unit (i.e., a feedback response recording) with approximate suitable features, and an intonation contour. Signal processing was applied to impose the target contour onto the candidate unit using three methods: Mel-Generalized Log Spectral Approximation (MLSA), Frequency domain pitch synchronous overlap-add (FD-PSOLA), and Harmonics Plus Noise Model (HNM). Using this approach, three

lexical forms (*mhm*, *really*, and *oh*) were imposed with three varying intonation contours. Participants rated the naturalness of each response on a 5-point Likert scale and assessed the feedback responses on several meanings: anger, sadness, amusement, happiness, contempt, solidarity, antagonism. They also evaluated the responses on dimensions with opposite meanings: certain vs. uncertain, agreeing vs. disagreeing, interested vs, uninterested, and anticipated vs. unexpected. Note that unit-selection refers to concatenating units of speech together, however, Pammi, Schröder, Charfuelan, et al. (2010) and Pammi and Schröder (2011) treated the feedback response as an entire unit, rather than synthesizing them from phones or diphones.

Attribute	**Possible values**
meaning	anger, sadness, amusement, happiness, contempt, certain, uncertain, agreeing, disagreeing, interested, uninterested, low-anticipation, high-anticipation, low-solidarity, high-solidarity, low-antagonism, high-antagonism
intonation	rising, falling, high, low
voice quality	modal, creaky, whispery, breathy, tense, lax
name	yeah, yes, mhmh, mhm, right, tsright, tsyeah, aha, (snort), (sigh), definitely, really, gosh, ah I see, oh god (gasp), yeah absolutely

Table 1.31. – Pammi, Schröder, Charfuelan, et al. (2010)'s MaryXML values for feedback responses.

As noted by Pammi, Schröder, Charfuelan, et al. (2010), signal processing degrades the audio quality. In contrast, Oertel, Gustafson, et al. (2016) used a statistical parametric speech synthesis system to synthesize the English feedback response "mhm" using the Clustergen method in Festival (Taylor et al. 1998). This TTS system took as input the text (i.e., the lexical form of the feedback response), a target duration, and target power for each syllable of the feedback response. It then synthesized feedback responses conditioned on these targeted duration and power features. Participants listened to these feedback responses within the context of conversations from the KTH-IDIAP corpus (Oertel, Funes Mora, et al. 2014), placed where original feedback responses occurred. They were asked to rate the level of attentiveness expressed by the feedback response "mhm".

Fernandez et al. (2022) synthesized interjections, which they defined as "non-lexical items that are used as exclamations, fillers, and backchannels, and which are prevalent, and indeed necessary, in conversational settings". Even though Fernandez et al. (ibid.) did not explicitly focus on synthesizing feedback responses, we include their work in this section since their definition encompasses backchannels. Fernandez et al. (ibid.) collected a database from a voice talent who recorded speech containing interjections. This data was used to train a non-attentive Tacotron 2 TTS system proposed by J. Shen et al. (2018). In order to synthesize interjections for other speakers, Fernandez et al. (2022) used voice conversion to generate augmented data, which then was used to

train the TTS system for the other speakers. However, from their audio samples, it seems that interjections were only synthesized within full utterances, such as "*aha*, I think I figured it out! I think you're trying to use the wrong email address to log in." It is unclear if their TTS system could synthesize isolated feedback responses.

Most recently, Mitsui et al. (2023) [3] proposed a CHatty Agents Text-to-Speech (CHATS) system, comprised of three modules: a speech-to-unit module, a unit language model, and a unit-to-speech module. This TTS system can synthesize speech for both the speaker and the listener, automatically generating feedback responses by the listener without needing to specify it in the input text. This allows the TTS system to autonomously decide when to generate, which feedback responses (i.e., lexical forms) to generate, and how they should sound. However, a significant downside of this approach is the lack of control over the timing, lexical form, and prosody of the feedback responses, potentially leading to responses that are inappropriate for the context.

Modeling all aspects of feedback responses is crucial for conversational agents to generate human-like listening behavior. However, most existing research has focused primarily on predicting the timing of these responses. Currently there are many limitations in modeling feedback responses. For example, models that predict timing, lexical form, communicative functions, and prosody of feedback responses are rarely evaluated subjectively, which may explain the limited work on synthesizing feedback responses using newer TTS systems. Additionally, feedback responses are often restricted to *continuers* such as "mhm" and "uh-huh", while those conveying richer attitudinal information are either ignored or lumped into a single category. The exception being perception studies investigating how prosody affects the communicative functions of feedback responses. This lack of diversity and fine-grained communicative functions in feedback response modeling might also account for the scarcity of systems capable of synthesizing responses that accurately reflect varied communicative functions.

In Chapter 5, we transplant the prosody of feedback responses from listeners in the Switchboard corpus onto a voice talent's voice, using a signal processing approach and a sequence-to-sequence TTS system. We evaluate the naturalness, prosodic appropriateness of the feedback responses, as well as to what extent these synthesis methods are able to preserve the intended communicative functions of feedback responses. However, before synthesizing feedback responses that convey various communicative functions, it is essential to first investigate which functions occur in human-human conversations. In Chapter 2, we present our annotations of the communicative functions found in the Switchboard corpus.

3. This model was trained on Japanese data.

2. Annotation scheme for communicative functions of feedback

Summary

2.1. Introduction and Motivation . . . 100
2.2. The Switchboard corpus . . . 102
2.3. Previous annotations of backchannels in the Switchboard corpus . . . 102
2.4. Finding feedback responses and clip creation . . . 103
2.5. Proposed annotation scheme . . . 104
2.6. Results . . . 109
2.6.1. Inter-annotator agreement . . . 109
2.6.2. Lexical distribution . . . 111
2.6.3. Prosodic Analysis . . . 113
2.6.3.1. Duration . . . 113
2.6.3.2. Mean Pitch . . . 114
2.6.3.3. Pitch Slope . . . 115
2.6.3.4. Pitch Range . . . 115
2.7. Discussion . . . 116
2.8. Summary . . . 117

2.1. Introduction and Motivation

This chapter is based on our publication *"Annotation of Communicative Functions of Short Feedback Tokens in Switchboard"* (Figueroa, Adigwe, et al. 2022), which was presented at the 13th Conference on Language Resources and Evaluation (LREC) 2022, in Marseille, France.

The main goal of this work was to address two of our research questions:

R1 *What are the various communicative functions of feedback responses in human-human conversations that a conversational agent may be required to convey?*

R2 *Per each communicative function found in human-human conversations what are the different distributions in terms of lexical form and prosody?*

These two research questions are motivated by the vocal feedback model proposed in Section 0.5. The communicative functions, lexical forms, and prosodic features of the feedback responses are the three components from the proposed vocal feedback model addressed by R1 and R2; these components are circled in Figure 2.1.

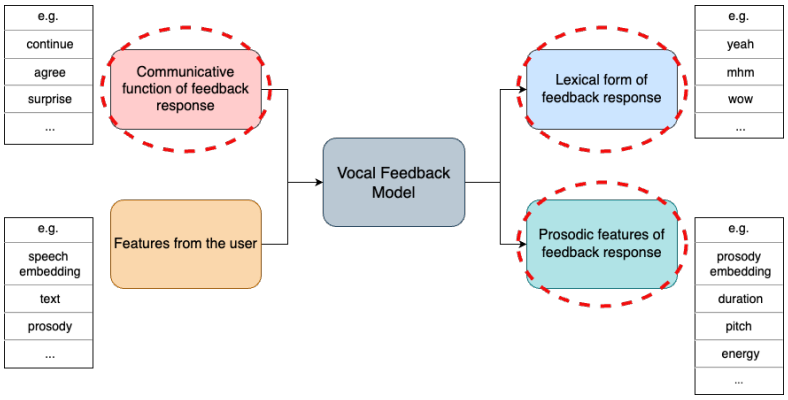


Figure 2.1. – Components from the vocal feedback model that are addressed by R1 and R2.

Communicative functions of feedback responses: Our proposed vocal feedback model takes as input the communicative function that the conversational agent needs to convey. This communicative function may be specified by the interaction designer or predicted by an external model. To understand the potential communicative functions required, we annotate various feedback responses with their corresponding communicative function in the Switchboard corpus. While numerous coding schemes exist for annotating communicative functions of feedback responses (see Chapter 1.2), there is no standardized annotation scheme. Existing coding schemes use identical labels with different descriptions or different labels for the same description, and they typically lack nuance, often categorizing all attitudinal functions under a single label. Consequently, we developed our own coding scheme, combining function labels from existing schemes with new labels based on our analysis of feedback responses in the Switchboard corpus.

Lexical forms of feedback responses: Our proposed vocal feedback model outputs a predicted lexical form based on the feedback function it needs to convey and the user's features. To prevent the conversational agent from sounding monotonous, there should be variability in the predicted lexical forms. For example, if the agent needs to express surprise, possible lexical forms might include "oh", "wow", or "really". To achieve this variability, we investigate the distributions of lexical forms for each communicative functions.

Prosodic features of feedback responses: Our proposed vocal feedback model also outputs predicted prosodic features. These prosodic representations can be either discrete (e.g., pitch, energy, duration) or distributed, such as prosody embeddings. While lexical variability helps prevent a conversational agent from sounding

monotonous, prosodic variability is equally important. For instance, if someone continuously responds with *yeah* in a flat tone, it might seem like they are bored. However, varying the prosody (e.g., “yeah,", “yeah!”, “yeah?”) conveys different attitudes and emotions, indicating they are engaged. To determine if each communicative function has distinguishable prosodic realizations, we performed an analysis of the different communicative functions.

To summarize, this chapter presents our coding scheme for annotating the communicative functions of feedback responses, the distribution of lexical forms of feedback responses found in the Switchboard corpus, and their prosodic analysis.

2.2. The Switchboard corpus

As mentioned in the introduction, we chose the Switchboard (Godfrey et al. 1992) corpus for our study of vocal feedback responses in human-human conversations. The Switchboard corpus comprises approximately 2,500 telephone conversations between dyadic pairs, with each conversation lasting between 3 to 10 minutes. All of the participants were native U.S. English speakers, totaling 500 individuals. Although the participants were given topic prompts, they were free to deviate, resulting in spontaneous conversations. The audios were recorded at a sampling rate of 8000 Hz, with each participant recorded on a separate channel. The recordings were transcribed and time-aligned at the word level for each speaker.

2.3. Previous annotations of backchannels in the Switchboard corpus

In previous work, Jurafsky et al. (1998) annotated *backchannels*, or lexical tokens related to backchannels, with dialog tags in the Switchboard corpus. They annotated four subcategories of backchannels with dialog act tags: *continuers* (indicating the other interlocutor should continue speaking), *assessments* (responses to something that the interlocutor said), *incipient speakership* (the listener attempting to take the floor), and *agreements* (agreements with opinions or acceptance of proposals). Additionally, they noted a fifth category, *yes-answers* (responses to yes/no questions), which shared the same lexical forms but were not considered backchannels. All of these annotations were based on text transcriptions, without the annotators listening to the speech.

As we know, prosody plays a crucial role in conveying the communicative functions of feedback responses (Beňuš et al. 2007; Neiberg, Salvi, et al. 2013; Prévot, Gorish, et al. 2015). To understand the role of prosody in assigning dialog acts, one of the authors from Jurafsky et al. (1998), who had originally participated in the annotations, relabeled 44 randomly selected conversations by listening to the audio files. During the relabeling process, the original labels were available for reference. A clear pattern emerged: prosody affects the disambiguation between continuers and agreements.

Some backchannels initially labeled as continuers were reclassified as agreements after listening to the audio.

Similar to other coding schemes for annotating the communicative functions of feedback responses reviewed in Chapter 1.2, there is only one category (assessments) that subsumes all attitudinal feedback responses. Jurafsky et al. (1998) lists the assessments adjectives as being: *great*, *good*, *nice*, *wonderful*, *cool*, *fun*, *terrible*, *exciting*, *interesting*, *wild*, *scary*, *hilarious*, *neat*, *funny*, *amazing*, *tough*, *incredible*, *awful*. However, we find this annotation scheme still lacks nuance, both in functions but also in lexical forms. For example, in their annotations they do not distinguish "mhm" from "uh-huh". Given the drawbacks of Jurafsky et al. (ibid.)'s annotations of backchannels in the Switchboard corpus, we chose to identify feedback responses differently.

2.4. Finding feedback responses and clip creation

Our first step in annotating feedback responses in the Switchboard corpus was to identify these responses within the conversations. As mentioned in Chapter 1.1, feedback responses can vary in length, ranging from short to long utterances depending on the terminology and definition used. In this thesis, however, we focused on synthesizing short feedback responses, specifically single tokens. The most common lexical forms for vocal feedback responses in English that are cited in the literature are *mhm*, *uh-huh*, *yeah*, *right*, and *okay*, see Tables A.1, A.2, and A.3 in Appendix A. However, we aimed to capture a broader range of lexical forms beyond those typically categorized as backchannels or continuers.

To automatically identify potential feedback responses in the conversations, we used the word level timestamps in the Switchboard corpus. Each word and silence in the conversations was annotated with start and end times. Using these timestamps, we searched for lexical tokens surrounded by silences of at least 5 seconds. Initially, we tried silences of duration of at least 1 second, but this included too many non-feedback responses. We also tried silences of at least 10 seconds, but this excluded too many potential feedback responses. Ultimately, silences of at least 5 seconds proved to be a good compromise. We selected lexical tokens that appeared 10 times or more in the corpus. Additionally, we searched for stretches of conversation lasting 20 seconds or more in which at least two lexical tokens occurred. These stretches of conversation became our clips for annotation.

One downside of this method was that the feedback responses that tend to occur consecutively are continuers. Therefore, we conducted a second search for potential feedback responses such as *gosh*, *golly*, and *pardon*, which tend to occur by themselves. In total, we had 7,545 clips, in which we found 26,981 potential short feedback responses. In this publication (Figueroa, Adigwe, et al. 2022), 1,627 short feedback responses were annotated.

From the beginning, we intended to have clips where one interlocutor would be the "speaker" and the other the "listener" for future experiments (see Chapter 5). For example, we had clips where Interlocutor A was the "speaker" and Interlocutor B

the "listener", producing feedback responses. Conversely, we had clips where Interlocutor B was the "speaker" and Interlocutor A the "listener", producing feedback responses. For our future listening tests, we planned to replace the Switchboard feedback responses with synthesized feedback responses, as detailed in Chapter 5.

2.5. Proposed annotation scheme

Our proposed annotation scheme for annotating the communicative functions of feedback responses aims to consolidate the concepts from existing schemes reviewed in Chapter 1.2 into a single scheme. As we were listening to the feedback responses in Switchboard, we kept in mind the communicative functions from Table 1.2 and labels used in previous annotation schemes Chapter 1.2. However, our annotations focused primarily on attitudinal information, that is, we annotated the feedback responses with the communicative functions from a perspective of generation. For example, when the feedback response "ugh" conveyed disgust or disapproval based on the prosody and context, we labeled it as disapproval rather than the four communicative functions (*Contact*, *Perception*, *Understanding*, *Attitudinal reactions*) proposed by Allwood, Nivre, et al. (1992).

Label	**Function Definition**
(C) Continue	Continue speaking. I hear you and I'm listening but not necessarily agreeing/disagreeing.
(U) Non-Understanding	I'm uncertain I understood/heard what you said.
(A) Agree	I agree with what you said.
(D) Disagree	I doubt what you said is true. I disagree with what you said.
(Y) Yes	I am giving a positive response/answer to your yes/no question.
(N) No	I am giving a negative response/answer to your yes/no question.
(S) Sympathy	I'm expressing sympathy/pity/sorrow/concern/ compassion to a negative statement.
(Ds) Disapproval	I am showing disapproval/disgust.
(MS) Mild Surprise	I am showing mild surprise, showing slight interest.
(SS) Strong Surprise	I am showing strong surprise; I am impressed.
(O) Other	Not a feedback. Filler, listener trying to take turn, or not enough context to determine function.
(?)	A feedback that doesn't fit into the current labels.

Table 2.1. – Labels and function descriptions in our proposed annotation scheme.

Table 2.1 provides descriptions of the communicative functions identified in the

1,627 clips that we[1] annotated. We identified ten communicative functions: *(C) Continue*, *(U) Non-understanding*, *(A) Agree*, *(D) Disagree*, *(Y) Yes*, *(N) No*, *(S) Sympathy*, *(Ds) Disapproval*, *(MS) Mild Surprise*, and *(SS) Strong Surprise*. It is important to note that this list is not exhaustive; it only accounts for the functions found in the clips we annotated, and other functions may exist depending on the corpus. For examples of feedback responses and their context for each function, refer to https://carolfigphd.github.io/FeedbackAnnotationScheme/.

The distinction between short feedback responses and other short vocalizations is not always clear. For example, when a speaker poses a rhetorical question and the listener response, should these responses be considered a turn/not a turn or a feedback response? Following the definition of feedback responses by Allwood, Nivre, et al. (1992), we include brief yes/no responses in our annotation scheme. Although yes/no responses are often not regarded as feedback responses, we believe that they should be classified appropriately from the perspective of a conversational system due their similar lexical forms. In our scheme, we do not classify incipient speakership lexical tokens as short feedback responses; instead we categorize them under the *(O) Other* category. Table 2.2, provides an example of an incipient speakership lexical token in context, as detailed by Jurafsky et al. (1998).

Spkr	**Dialog Act**	**Utterance**
B	Wh-Question	Now, how long does it take for your contribution to vest?
A	Statement	God, I don't know /
A	Statement	<laughter> It's probably a long time <laughter>.
A	Statement	I'm sure it's not till
A	Statement	like twenty-five years, thirty years.
B	Incipient	Yeah,/
B	Statement	the place I work at's, health insurance is kind of expensive./

Table 2.2. – Example of incipient speakership lexical token by Jurafsky et al. (1998).

H. H. Clark (1996) introduced the concept of two tracks of actions that people engage in during a conversation. In the first track, people talk about official business, while in the second track, they talk about their communicative acts. In the second track, "contributors ask for confirmation or invite completions, and respondents provide acknowledgment and other evidence of understanding" (ibid., p. 252). Table 2.3 illustrates an example of these two tracks as described by H. H. Clark (ibid., p. 243). In his explanation of this example, H. H. Clark (ibid.) explains that the response *yes* does not mean that "yes, it was a lovely day" but rather it means "yes, I understand what you mean by that". However, we can think of this *yes* as both an agreement–"yes, I agree it is a lovely day"–and an acknowledgment–"yes I understand what you said".

1. This is a singular we, à la Ayn Rand's Anthem style.

Utterance	Track 1	Track 2
A: it was uh it was a lovely day B: yes	1. [I assert] it was a lovely day 2. [I ratify your assertion]	1. [Do you understand this?] 2. yes [I understand that.]

Table 2.3. – Example of the two tracks of actions by H. H. Clark (1996).

As mentioned in Chapter 1.2, many existing schemes for annotating the communicative functions of feedback responses are based on the four basic communicative functions–*Contact*, *Perception*, *Understanding*, and *Attitudinal reactions*–proposed by Allwood, Nivre, et al. (1992). These four feedback functions are respectively related to the four levels of joint actions of an addressee proposed by H. H. Clark (1994), see Table 2.4. In his description of grounding, H. H. Clark (1996) introduces two concepts: *upward completion*, which he defines as "in a ladder of actions, it is only possible to complete actions from the bottom level up through any level in the ladder" and *downward evidence*, which he defines as "in a ladder of actions, evidence that one level is complete is also evidence that all levels below it are complete" (ibid., pp. 147–148). In our annotation scheme, we will use the concept of *downward evidence*.

Level	H. H. Clark 1994	Allwood, Nivre, et al. 1992
4	Uptake	Attitudinal reactions
3	Understanding	Understanding
2	Identification	Perception
1	Attention	Contact

Table 2.4. – Joint action of the addressee at each grounding level by H. H. Clark 1996 and corresponding feedback communicative function by Allwood, Nivre, et al. 1992.

In our annotation scheme, we refer to the levels of joint actions of the addressee proposed by H. H. Clark (1994) and the four feedback functions proposed by Allwood, Nivre, et al. (1992) as grounding levels: *Contact*, *Perception*, *Understanding*, and *Acceptance/Attitude*. Utilizing the concept of *downward completion*, each level below is implied by the level above. Since we regard these as grounding levels rather than functions, we do not use *Contact*, *Perception*, *Understanding*, and *Attitudinal reactions* to label feedback responses in the Switchboard corpus. In other words, we think of our ten function labels as being in these grounding levels. Table 2.5 shows our ten functions labels in their corresponding grounding levels.

In their annotations of vocal feedback responses, Malisz, Włodarczak, Buschmeier, Skubisz, et al. (2016) (see Table 1.6) observed a lot of disagreement in their inter-annotator agreements. From their confusion matrix of inter-annotator agreement,

Grounding Level	Positive	Negative
Acceptance/Attitude	A, Y, S, SS, MS	D, N, Ds
Understanding	C	U
Perception		
Contact		

Table 2.5. – Labels in Grounding Framework.

they reported that annotators tended to favor the label representing the middle ground level, P2 (perception and understanding). They suggested that this tendency arises because "feedback functions form a hierarchy of grounding strength" (Malisz, Włodarczak, Buschmeier, Skubisz, et al. 2016), indicating that difficulties in discerning grounding strengths could lead to high levels of disagreement.

In our annotation scheme, we avoid making a fine-grained distinction between *Contact*, *Perception*, and *Understanding*. We believe these distinctions are challenging to discern solely based on the lexical form of the short vocalizations and it would largely depend on contextual information. As Schegloff (1982) argued, tokens such as *uh-huh*, *mhm*, and *yeah* only claim attention and/or understanding. For instance, if a listener responds with convincing *mhms* or *uh-huhs* while distracted, the interlocutor or a third party observer might infer that they have established contact, perception, and understanding, when in reality, there might only be contact, or just contact and perception.

Some schemes either eliminate the *Contact* function or merge it with other grounding levels. For instance, Neiberg, Salvi, et al. (2013) represented contact and perception with a neutral function. In the scheme by Allwood, Cerrato, et al. (2007), the *Basic* function had two features CP (contact and perception) and CPU (contact, perception, and understanding). In our annotation scheme *(C) Continue* can be interpreted as encompassing contact, perception, or contact, perception and understanding, depending on the context of the feedback response.

In their annotation scheme Allwood, Cerrato, et al. (ibid.) differentiate *Acceptance* and *Attitude* into distinct categories. Conversely, in the coding schemes of Buschmeier et al. (2011) and Malisz, Włodarczak, Buschmeier, Skubisz, et al. (2016), the function *Attitude* can either be a modifier to a category or be a category by itself. For example, a feedback response might be labeled as A to express an attitude, or as P3A, where A modifies P3 (perception, understanding, and acceptance).

In contrast, in our scheme, the highest grounding level is *Acceptance/Attitude*, where we have merged *Acceptance* and *Attitude* into a single grounding level. Similarly, Prévot, Gorish, et al. (2015) and Prévot, Gorisch, et al. (2016) include *Approval/non-approval*, which can be thought of as *Acceptance*, within their *Evaluation/Attitude* level. We chose not to separate *Acceptance* and *Attitude* due to the concept of *downward completion*, where each lower level is implied by the level above. For instance, if a feedback response is annotated with the function *(MS) Mild Surprise*, it does not nec-

essarily mean that the message was accepted (agreement) or rejected (disagreement), but we can infer that the levels of understanding, perception and contact have been established.

Our scheme is also flexible, allowing us to categorize our communicative functions into the two categories proposed by Bavelas et al. (2000): *generic* listener responses and *specific* listener responses. The function *(C) Continue* can to be thought of as corresponding to the *generic* listener responses category. While, functions such as *(U) Non-understanding* and those within the *Acceptance/Attitude* level correspond to *specific* listener responses.

Additionally, our function labels can be visualized in a valence-arousal space, akin to a circumplex model of affect, as displayed in Figure 2.2. In this space, feedback functions containing attitudinal information are positioned within one of the four quadrants. Those lacking attitudinal or emotional information are placed in the neutral section. It is important to note that our labels in the *Acceptance/Attitude* grounding level are not exhaustive; they represent only the labels identified within the 1,627 short feedback responses annotated in the Switchboard corpus.

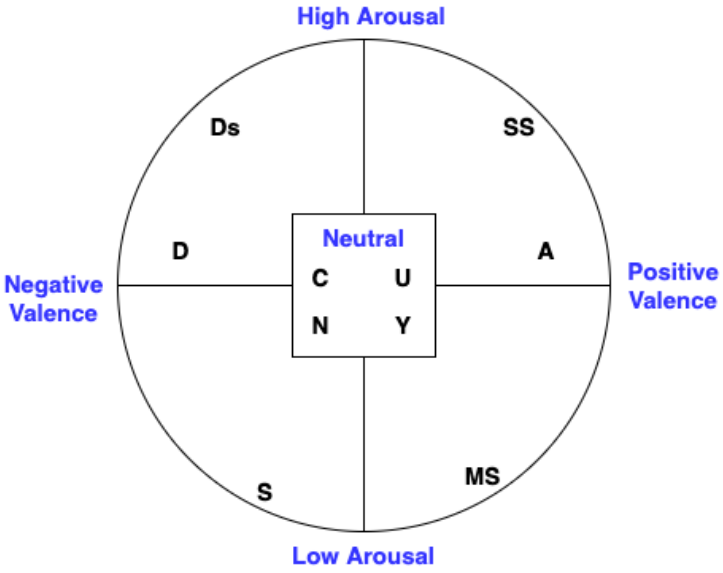


Figure 2.2. – Labels in Valence-Arousal Space.

2.6. Results

2.6.1. Inter-annotator agreement

From the 1,627 short feedback responses initially annotated by the first annotator[2], a subset of 1,124 responses was subsequently annotated by a second annotator[3] to calculate inter-annotator agreement. Prior to beginning the annotations, the second annotator met with the first to gain a thorough understanding of the labeling scheme. During this meeting, they discussed examples for each function. After this training session, the second annotator independently completed the annotations. Both annotators assessed the feedback responses in the Switchboard corpus by listening to the entire clip, replaying it multiple times if necessary, and assigning labels based on the context and prosody of the lexical token. In cases where it was challenging to distinguish between two potential labels, such as deciding whether a response functioned as *(A) Agree* or *(C) Continue*, the annotators assigned a combined label, *'A/C'*. The distribution of the labels by each annotator is illustrated in Figure 2.3.

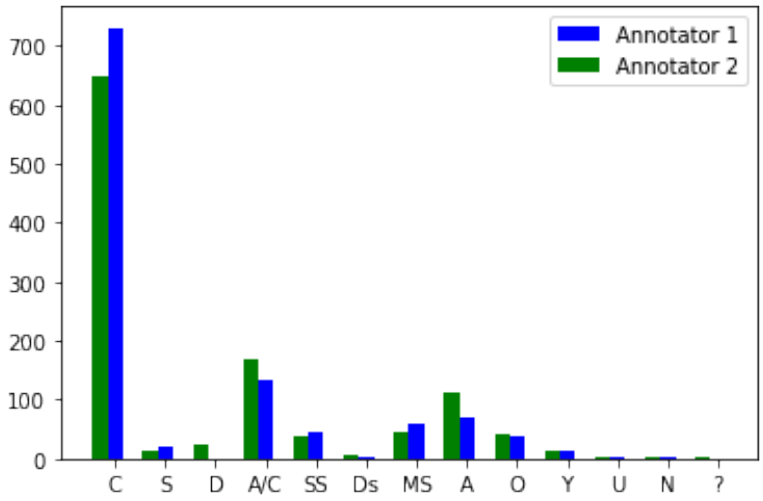


Figure 2.3. – Feedback token labels distribution across annotators.

The annotations were skewed due to the unbalanced distribution of function categories, with some functions having fewer instances than others. For example, in our current annotations, there are relatively few instances of *(D) Disagree*, *(Y) Yes*, *(N) No*, and *(S) Sympathy* compared to *(C) Continue* and *(A) Agree*. Only one annotator used the *"?"* label, but we included this label in our annotation scheme to allow for the identification of new feedback functions in future research. Although the distributions of all function labels are not proportional and this may affect our inter-annotator agreement calculations, we still presented these calculation.

2. Carol Figueroa
3. Adaeze Adigwe

Similar to Kousidis et al. (2013), our main objective is to validate our proposed feedback annotation scheme. Kousidis et al. (ibid.) noted that the Cohen Kappa is a stricter reliability measurement. Following their approach, we validate our scheme using *percentage agreement* scores. However, to provide a more comprehensive evaluation, we also employ the *Cohen Kappa Coefficient*. This dual approach ensures a robust assessment of our annotation reliability.

The first approach calculates the percentage agreement over the samples labeled by both annotators. Within this approach:

— *Absolute*. Only exact matches between the annotators are counted, leading to an agreement of 69.3%. Note that this method is very strict. For example, if one annotator assigns *'A'* and the other annotator assigns *'A/C'*, this approach does not consider the partial agreement for *'A'*.

— *Weighted*. To better account for partial agreement between annotators, we also use a method that assigns a weight of 0.5 to instances of partial agreement, compared to 1.0 for absolute agreement. This weighted percentage approach results in a 78.1% agreement.

— *Liberal*. Lastly, a more liberal approach considers any overlap between the categories of both annotators as a match. For instance, *'A/C'* and *'A'* would be considered a match. This method results in a 86.8% agreement.

We report a summary of the three different types of percentage-agreement metrics in Table 2.6.

Approach	**Agreement Percentage**
Absolute	69.3%
Weighted	78.1%
Liberal	86.8%

Table 2.6. – Inter-Annotator Percentage Agreement Score.

For our second approach, we calculate the widely used Cohen Kappa Coefficient, which measures the inter-rater reliability. To accommodate the fact that annotators were allowed to assign up to two labels to a feedback response, we modified the Cohen Kappa Coefficient calculation slightly. In cases of partial agreement, cells in the confusion matrix are assigned a value of 0.5. For example, if for a given feedback response, *annotator 1* assigns the label *'A'* and *annotator 2* assigns the label *'A/C'*, we populate the cell $M_{A,A} = 0.5$ for partial agreement and $M_{A,C} = 0.5$ for partial disagreement. These values are reported in a confusion matrix illustrated in Figure 2.4. The modified calculation method results in a Cohen Kappa coefficient score of **0.51**, indicating moderate agreement between the annotators. This score is slightly higher than the score of 0.484 obtained with the unmodified Cohen Kappa Coefficient score.

The Confusion matrix (Figure 2.4) highlights areas of agreement and disagreement between the function categories of feedback responses. We observe relatively strong agreement for most function categories between the two annotators. The notable

exception is the function category *(D) Disagree*, where disagreement is observed. This could possibly be explained by the low number of instances of this category in the corpus, as seen in Figure 2.3.

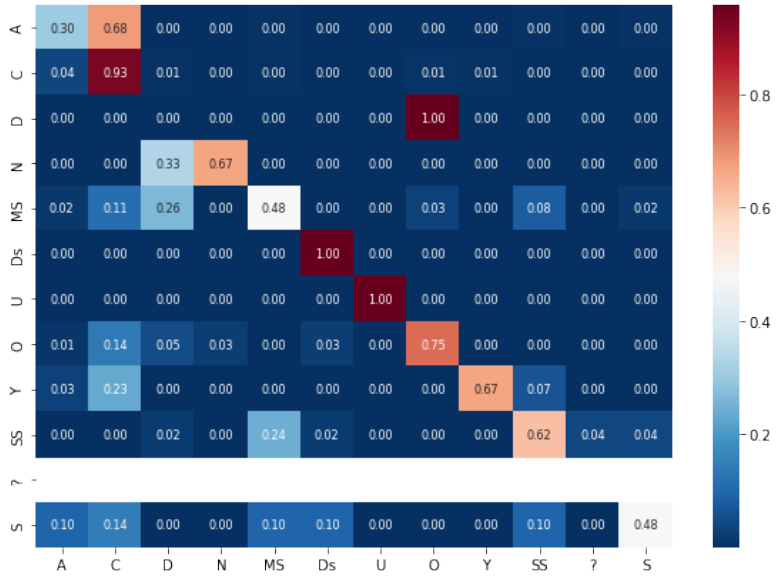


Figure 2.4. – Normalized Confusion matrix between the annotators.

2.6.2. Lexical distribution

As mention in Chapter 1.1, researchers studying vocal feedback responses often do not provide a comprehensive list of the lexical forms used in their studies. Their forms are typically referred to simply as feedback responses, such as *mhm*[4], *yeah*, or *uh-huh* and the like. Since we want conversational systems to generate variability in lexical forms, we have documented all the lexical forms identified in our annotations. In total, we found 59 unique lexical forms that were categorized as short feedback responses. Figure 2.5 illustrates the distribution of the lexical forms that appear at least eight times. Table 2.7 provides a detailed breakdown of the counts for all 59 unique lexical forms and their distribution across each communicative function.

As highlighted by Figure 2.3, function *(C) Continue* is the most frequent feedback function. This aligns with the findings of Jurafsky et al. (1998), who identified continuers as the most common type of backchannel in their annotation of the Switchboard corpus. The function *(C) Continue* also exhibits the most variation in lexical forms; *mhm* has the highest count, followed by *yeah*, *uh-huh*, *right*, *sure*, *mmm*, *hm*, *yes*, etc. Similarly, Beňuš et al. (2007) found that *mhm* was the most common backchannel in their analysis of the Columbia Games Corpus. In their study, the backchannels (their

4. There is no consensus whether vocalizations such as *mhm* and *uh-huh* are lexical or non-lexical.

Function	**Lexical forms and Count**
(C) Continue	mhm (321), yeah (188), uh-huh (100), right (60), sure (13), mmm (12), hm (10), yes (10), okay (9), huh-huh (4), huh (2), ah-hah (1), exactly (1), huh-hm (1), m-kay (1), mhm-hm (1), oh (1), uh-hah (1), uh-hm (1), well (1), yah (1), yuh (1)
A/C	yeah (65), mhm (62), right (45), uh-huh (4), yes (3), sure (1), yah (1), yup (1)
(U) Non-understanding	pardon (21), what (14), huh (5), sorry (1)
(A) Agree	yeah (61), no (59), right (27), absolutely (16), yes (16), exactly (8), mhm (7), indeed (1), nah (1), nuh (1), okay (1), true (1), yea-m (1), yep (1)
(D) Disagree	hm (19), no (12), huh (4), really (3), mmm (2), nah (1), yeah (1)
(Y) Yes	yeah (12), yes (4), mhm (2), absolutely (1), uh-huh (1)
(N) No	no (66), m-m (1), nuh (1), nuh-uh (1)
(S) Sympathy	oh (5), mmm (4), hm (2), uh-oh (2), aww (1), jeez (1), oof (1), uh (1), yeah (1)
(Ds) Disapproval	ooh (13), ugh (8), gosh (7), jeez (5), goodness (4), golly (2), wow (2), gee (1), jee (1), mmm (1), no (1), oh (1), oof (1), uh (1), uh-oh (1), what (1), wooh (1), yuck (1)
(MS) Mild Surprise	hm (34), huh (26), mmm (4), ah (3), goodness (3), oh (3), really (3), uh (3), interesting (2), uh-huh (2), eh (1), huh-uh (1), mhm (1), uh-ho (1)
(SS) Strong Surprise	wow (117), ooh (20), oh (12), huh (5), gosh (4), really (3), hm (2), yeah (2), ah (1), aha (1), goodness (1), jeez (1), mmm (1), uh (1), uh-oh (1), wah (1), what (1), yes (1)

Table 2.7. – Distribution of lexical forms and corresponding count per communicative function.

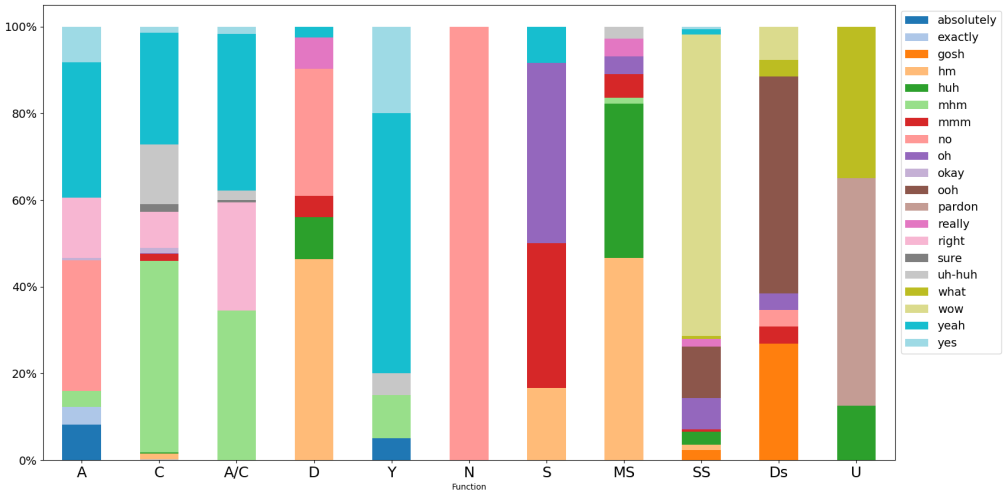


Figure 2.5. – Distribution of lexical forms per communicative function.

continuers) included *mhm*, *uh-huh*, *okay*, and *yeah*, which coincide with our lexical forms for *(C) Continue*.

Echoing the findings of Jurafsky et al. (1998), *yeah* is the most ambiguous lexical form in our annotations. It appears in several communicative function categories, including *(C) Continue*, *A/C*, *(A) Agree*, *(D) Disagree*, *(Y) Yes*, *(S) Sympathy*, and *(SS) Strong Surprise*. Other lexical forms, such as *hm*, also overlap across different functions.

2.6.3. Prosodic Analysis

We also explored the differences in prosody across communicative functions. Specifically, we investigated differences in duration, mean pitch, pitch slope, and pitch range across the functions using one-way ANOVA tests and Tukey Kramer post-hoc tests. This analysis aimed to determine whether communicative functions have distinguishable prosodic features. The prosodic features are extracted at the "word-level". The number of annotations per function, along with their duration, mean pitch, pitch slope, and pitch range, are summarized in Table 2.8.

2.6.3.1. Duration

Due to errors in the start and end times of the word-level alignments, we first revised these timestamps to accurately calculate the duration of each feedback response. According to Table 2.8, feedback responses conveying the function *(S) Sympathy* have the longest average duration (0.580 seconds), while those conveying *(U) Non-understanding* are the shortest, averaging 0.288 seconds.

A one-way ANOVA test revealed a statistically significant difference in mean duration: $F(10,1616)=44.73$, $p<0.05$. Table E.1 in Appendix E presents the results of the Tukey test for duration, showing that functions *(S) Sympathy*, *(SS) Strong Surprise*, and

Function	Duration (s)	Mean Pitch (SD)	Pitch Slope (SD)	Pitch Range (SD)	Count
(C) Continue	0.329	-0.130	-0.092	1.674	740
(U) Non-Understanding	0.288	0.865	0.919	2.460	41
(A) Agree	0.389	-0.181	-0.307	1.795	201
A/C	0.320	-0.192	-0.172	1.595	182
(D) Disagree	0.360	0.240	-0.582	1.989	42
(Y) Yes	0.313	-0.110	-0.211	1.768	20
(N) No	0.357	0.043	-0.314	1.594	69
(S) Sympathy	0.580	-0.226	-0.572	2.061	18
(Ds) Disapproval	0.511	-0.160	-0.303	2.283	52
(MS) Mild Surprise	0.341	0.514	-0.817	2.675	87
(SS) Strong Surprise	0.511	0.065	-0.483	2.403	175

Table 2.8. – Summary of duration, mean pitch, pitch slope, pitch range, and the count of annotations per communicative function.

(Ds) Disapproval have significantly longer duration compared to the other functions. In her study of *right* and *really* in the Switchboard corpus, Lai (2009) found that these lexical forms which express surprise had long duration. In our results, *(SS) Strong Surprise* are the second function category with the longest duration. When comparing the durations of *agreements* and *backchannels* (i.e., continuers), Beňuš et al. (2007) in the Columbia Games corpus, they found them to be similar. However, Jurafsky et al. (1998) observed that *continuers* had shorter duration than *agreements* in the Switchboard corpus. In our findings, *(A) Agree* also has a significantly longer duration than *(C) Continue*.

2.6.3.2. Mean Pitch

Pitch values were extracted using the Python implementation (Dinh et al. 2019) of the WORLD vocoder (Morise et al. 2016), transformed to logarithmic scale, and z-score normalized per speaker, based on their entire speech. Due to technical problems during pitch extraction, we eliminated 33 instances of *(C) Continue*, 10 of *(A) Agree*, 10 of *A/C*, 1 of *(Y) Yes*, 1 of *(N) No*, 1 of *(S) Sympathy*, 2 of *(Ds) Disapproval*, and 2 of *(SS) Strong Surprise*.

A one-way ANOVA test revealed a statistically significant difference in mean pitch: $F(10,1556)=19.13$, $p<0.05$. Table E.2 in Appendix E presents the results of the Tukey test for mean pitch. The results indicate that functions *(U) Non-understanding* and *(MS) Mild Surprise* have significantly higher mean pitch than *(SS) Strong Surprise*, *(N) No*, *(Y) Yes*, *(Ds) Disapproval*, *(A) Agree*, and *(S) Sympathy*. Additionally, *(D) Disagree* and *(SS) Strong Surprise* are found to have significantly higher mean pitch than *(C)*

Continue and *(A) Agree*.

In our analysis, we find that *(MS) Mild Surprise* and *(SS) Strong Surprise* have high mean pitch, similarly, Lai (2009) found that *right* and *really* which convey surprise have higher mean pitch. In their analysis, Beňuš et al. (2007) found that *backchannels* (i.e., continuers) have a higher pitch than *agreements*. This finding contrasts with Jurafsky et al. (1998), who noted that *continuers* have a lower pitch compared to *agreements*. A potential reason for the differing observations between Beňuš et al. (2007) and Jurafsky et al. (1998) could be that Beňuš et al. (2007) focused solely on affirmative words, excluding uses of *no*, which can be used to agree with negative statements and was not annotated in the Columbia Games corpus. In our analysis, we observed no significant difference in mean pitch between *(C) Continue* and *(A) Agree*.

2.6.3.3. Pitch Slope

To calculate pitch slope, we subtracted 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 response. A one-way ANOVA revealed a statistically significant difference in pitch slope, with $F(10, 1556)=23.63$, $p<0.05$. Table E.3 in Appendix E presents the results of the Tukey test for pitch slope; the function *(U) Non-Understanding* is the only one with a rising pitch slope and it is significantly different from all the other functions. Additionally, functions *(U) Non-understanding*, *(C) Continue*, *(Y) Yes*, *(Ds) Disapproval*, *(A) Agree*, *(N) No*, and *(SS) Strong Surprise* have significantly higher pitch slopes than *(MS) Mild Surprise*.

Jurafsky et al. (1998) observed that *continuers* have a flatter pitch than agreements, a finding that contrasts with Beňuš et al. (2007). In their comparison of *agreements* and *backchannels* (i.e., continuers), Beňuš et al. (ibid.) reported that backchannels exhibit a higher pitch slope than agreements. The discrepancy between these observations may be attributed to differences in lexical forms annotated and in the corpora used. From our analysis, we note that the function *(C) Continue* is significantly different from *(A) Agree*, with *(C) Continue* having a flat pitch slope.

2.6.3.4. Pitch Range

From our results, we observe that the function *(MS) Mild Surprise* has the most variability in pitch, while *(N) No* shows the least. A one-way ANOVA test revealed that a statistically significant difference in pitch range: $F(10,1556)=13.23$, $p<0.05$. Table E.4 in Appendix E presents the results of Tukey test for pitch range. The results show that functions *(MS) Mild Surprise*, *(U) Non-understanding*, and *(SS) Strong Surprise* have a significantly bigger pitch range than *(A) Agree*, *(C) Continue*, and *(N) No*. Additionally, we observe that function *(Ds) Disapproval* has a significantly higher pitch range than *(C) Continue*. Lai (2009) found that pitch range, specifically a bigger pitch range is the best prosodic cue for conveying surprise; we observe that *(MS) Mild surprise* and *(SS) Strong Surprise* have the most variability in pitch, as well as *(U) Non-understanding*.

2.7. Discussion

From our analysis, it is clear that both the lexical form and the prosodic features of feedback responses are crucial in determining their communicative function. Lai (2009) characterized *right* and *really* conveying surprise as having high mean pitch, high mean intensity, bigger pitch range, and longer duration than *right* and *really* not conveying surprise. Jurafsky et al. (1998) describe *continuers* as having shorter durations, flatter pitch contours, lower fundamental frequency F_0, and less energy compared to *agreements*. In contrast, Beňuš et al. (2007) characterize *backchannels* (i.e., continuers) as having higher pitch, pitch slope, and intensity, and similar in durations to *agreements*. The discrepancies between these findings could stem from the differences in the corpora used: Switchboard is a spontaneous conversational corpus, whereas the Columbia Games Corpus is task-oriented, which potentially causes differences in the types of feedback responses. This is supported by the recent findings from Spaniol et al. (2024), which compared German backchannels from task-oriented speech and spontaneous speech. They found that backchannels in the task-oriented speech had predominantly rising intonation, whereas the backchannels in the spontaneous speech had falling intonation. Additionally, Beňuš et al. (2007) annotated only affirmative words in the Columbia Games corpus, which could contribute to the observed differences.

In our results, we note both similarities and differences in comparison to the findings of Jurafsky et al. (1998), despite analyzing the same corpus. These differences might be attributed to the fact that the majority of their annotations were made without listening to the audio, which could lead to classification errors between *continuers* and *agreements*. In this paper, we did not analyze differences in energy or intensity, but intensity will be addressed in the next chapter (see Chapter 3).

Our proposed coding scheme includes functions with attitudes, which have not been annotated in such fine-grained categories or analyzed previously. Our annotations have an unbalanced distribution of communicative functions. For example, we have 740 instances of *(C) Continue* but 18 instances of *(S) Sympathy*. This imbalance may be due to our method of searching for feedback responses. If we had annotated by listening to entire conversations rather than searching for clips of consecutive feedback responses, we might have identified more non-continuer feedback responses, as these tend to occur in isolation. Additionally, our automatic search method might have missed some lexical forms of feedback responses. However, by focusing on clips, we were able to annotate a greater number of conversations. Listening to full conversations would have been more time-consuming, given that there are 2,438 telephone conversations in the Switchboard corpus, and would have resulted in fewer conversations being annotated.

As more feedback responses are annotated using our coding scheme, we hope to identify clearer prosodic characteristics and more distinct distributions of the lexical forms per function category. In this study, we analyzed the prosodic features within each communicative function, which include a variety of lexical forms. However, as we annotate more feedback responses, it will become possible to conduct comparisons

of the same lexical form across different communicative functions. For example, we could investigate whether a *yeah* that conveys *(C) Continue*, *(A) Agree*, *(Y) Yes*, and *(SS) Strong Surprise* differ in prosodic features. This deeper analysis could reveal nuanced variations in how a single lexical form is prosodically realized based on their corresponding communicative function.

2.8. Summary

In this chapter, we have presented the main contribution of our work (Figueroa, Adigwe, et al. 2022): our coding scheme for annotating the communicative functions of feedback responses. Given the absence of a standard coding scheme, we propose our own, which aims to integrate the ideas from existing schemes into a single coding scheme. Our proposed coding scheme encompasses functions with attitudinal information, thereby providing a nuanced range of feedback functions.

From our analysis of the distribution of lexical forms, we have observed that they can be ambiguous. Thus, it is crucial to consider both the lexical form and the prosodic features of feedback responses to accurately determine their communicative function. We also discovered significant differences in duration, mean pitch, pitch slope, and pitch range among the communicative functions, suggesting that these prosodic features can potentially be used to distinguish them.

Manually annotating the 1,627 feedback responses in the clips was a time-consuming process, yet there remain 26,981 potential feedback responses that have not been annotated. In the following chapter, we will use the annotations from this work to train classifiers to annotate the remaining potential feedback responses in the Switchboard corpus.

3. Automatic classification of feedback functions

Summary

3.1. Introduction and Motivation . . . 118
3.2. Related Work . . . 120
3.3. Manual annotations of feedback responses . . . 122
3.4. Input features . . . 122
3.4.1. Feedback features . . . 123
3.4.2. Context features . . . 124
3.5. GPT-3 . . . 124
3.6. Classifiers . . . 128
3.7. Results . . . 128
3.7.1. Performance of classifiers . . . 129
3.7.2. Labeling remaining potential feedback responses . . . 131
3.7.3. Investigating sex differences . . . 133
3.7.4. Analysis of Overlap . . . 134
3.8. Discussion . . . 135
3.9. Summary . . . 136

3.1. Introduction and Motivation

This chapter is based on our publication *"Classification of Feedback Functions in Spoken Dialog Using Large Language Models and Prosodic Features"* (Figueroa, Ochs, et al. 2023), which was presented at the 27th Workshop on the Semantics and Pragmatics of Dialogue (SemDial) 2023, in Maribor, Slovenia.

The main goal of this work was to address the following research question:

R3 *Is it possible to automatically classify the communicative functions of feedback responses based on their lexical form, prosodic features, and preceding context?*

As mentioned in the previous chapter, manually annotating feedback responses in the Switchboard corpus was a time-consuming process. Therefore, in order to annotate the remaining potential feedback responses, we proposed automating this task, which was the primary motivation for research question R3. To develop an automated

method for classifying the communicative functions of feedback responses, we reflected on our manual annotation process to identify how it could be replicated. While listening to the conversations in the Switchboard corpus and manually annotating the feedback responses, we considered three main factors: the lexical form of the feedback response, its prosody, and its preceding context.

These three factors are components we introduced in our proposed vocal feedback model. Figure 3.1 highlights the components addressed in R3: the communicative functions, the lexical, forms, and the prosodic features of feedback responses, and the preceding context. Instead of predicting the lexical form and prosodic features of the feedback responses based on the communicative function and preceding context, we rearranged these components to create a classification task. For this task, the model takes as input the lexical form and prosodic features of the feedback response, along with the preceding context, and predicts the communicative function of the feedback response. Figure 3.2 illustrates this classification model.

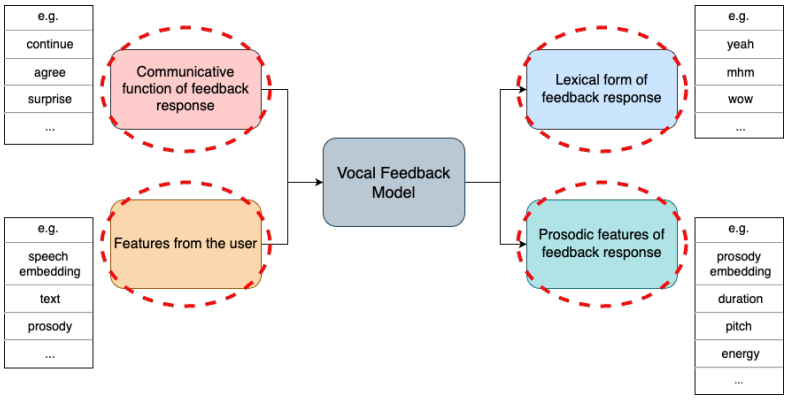


Figure 3.1. – Components from the vocal feedback model addressed by R3.

Although the primary motivation for automatically classifying the communicative functions of feedback responses was to expedite the annotation process, such a model has several other potential applications. For instance, when deployed online within a spoken dialogue system, it can classify user-generated feedback responses. By doing so, the conversational agent can determine whether the user intends to interrupt and take control of the conversation or if it should continue speaking. When utilized offline, the annotated feedback responses from such a model could provide insights into human conversational behavior. For example, analyzing these responses could reveal whether there are prosodic features specific to certain communicative functions, which could then inform the synthesis of feedback responses in text-to-speech systems.

To summarize, this chapter presents the results of our experiments in which we explored the performance of various models and combinations of features for classifying the communicative functions of feedback responses in the Switchboard corpus. For the classification task, we used lexical and prosodic features from the feedback responses, as well as contextual features from interlocutor's preceding utterance. Given

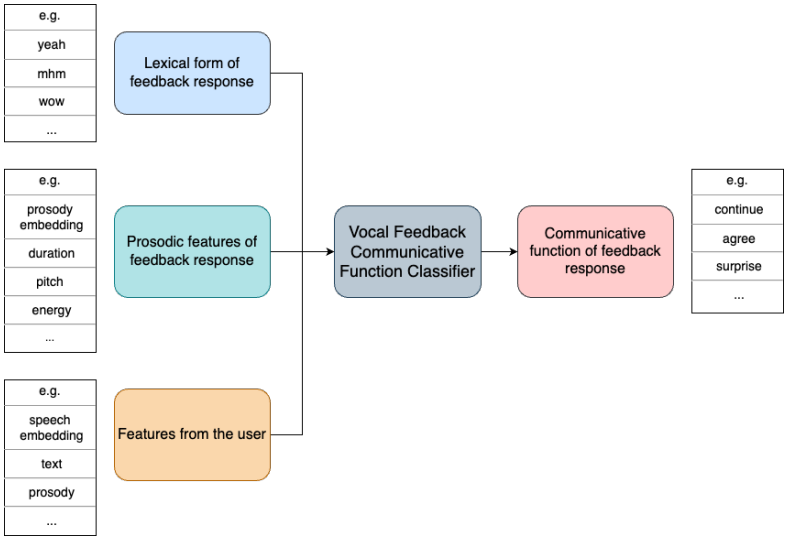


Figure 3.2. – Visualization of the classifier that classifies the communicative function of feedback responses.

the complexity of representing dialogue context, especially with the limited amount of annotated data available, we also explored the use of probability distributions from a pre-trained large language model (GPT-3) as input to a Support Vector Machine (SVM) classifier, in conjunction with the previously mentioned features. To the best of our knowledge, using probability distributions from GPT-3 as input to another machine learning algorithm for this task has not been done before.

3.2. Related Work

The classification of feedback functions is related to the broader problem of Dialog Act classification, where the goal is to identify the communicative function of an utterance in dialogue. However, most Dialog Act classification schemes do not include specific labels for the various types of feedback responses. Typically, there is only one single dialog act label, *backchannels*, which are generally treated as a single category without any fine-grained distinctions (Stolcke et al. 2000; Dielmann et al. 2008; Liu 2006). As reviewed in Subsection 1.3.3, some studies have classified the communicative functions of feedback responses, thereby creating more fine-grained distinctions.

When discussing the classification of the communicative functions of feedback responses, other related work on *identifying* or *detecting* feedback responses or *predicting* their communicative functions is often cited. To clarify any misconceptions, *identifying/detecting* feedback responses, *predicting* the communicative function of a feedback response, and *classifying* the communicative function of feedback responses are related but distinct tasks.

Identifying or *detecting* feedback responses can essentially be reduced to a yes/no question: "Is this a feedback response?" In human-machine interactions, this involves determining whether a user has produced a feedback response, such as *uh-huh*, *okay*, or *wow* to decide whether the system should stop talking. To the best of our knowledge, no study has been conducted in which a robot or virtual agent actively detects user vocal feedback responses while it is speaking to make decisions about whether to pause or continue speaking, as well as classifies the communicative function of the user's feedback response. In a corpus study, this task often involves determining whether a lexical form or utterance qualifies as a feedback response or another type of short utterance (Sajjanhar et al. 2006; Heldner, Edlund, Al Moubayed, et al. 2010; Kholiavin et al. 2020; Müller et al. 2022). For instance, Edlund, Heldner, and Pelcé (2009) introduced an auxiliary unit *very short utterances* (VSUs) [1], which are automatically extracted from corpora using duration and voicing ratio thresholds. These VSUs "are intended to capture a large proportion of interactional dialogue phenomena such as those commonly referred to as backchannels, feedback, continuers, et cetera, at zero manual effort" (ibid.). In a subsequent study, Heldner, Edlund, Al Moubayed, et al. (2010) trained a classifier on duration and inter-speaker relative loudness to explore how effectively these VSUs and Non-VSUs captured *backchannels* and *non-backchannels*. They employed a K-Nearest Neighbors classifier to perform a binary classification determining whether a VSU or NonVSU was a *backchannel* or *non-backchannel*.

Predicting the communicative function of a feedback response can be framed as the question: "What communicative function should the conversational agent convey that is appropriate at this moment in the conversation?" For example, Ortega, C.-Y. Li, et al. (2020) trained convolutional neural networks (CNNs) on lexical and acoustic features to predict whether the system should generate a *backchannel-continuer*, or generate a *backchannel-assessment*, or should not generate a *backchannel*. As discussed in Section 1.5, where we reviewed models predicting the timing and communicative functions of feedback responses, many of these prediction models focus on predicting either *continuers* or *assessments*, but these categories are not sufficiently fine-grained. Future work should aim to predict communicative functions of feedback responses that allow conversational agents to express nuanced attitudinal information, such as *surprise*, *agreement*, *sympathy*. For instance, if a user says something sad, the feedback response generated by the conversational agent should convey *sympathy*.

Classifying the communicative function of the feedback response can described as asking, "This may [2] be a feedback response; what is its meaning, what is it conveying?" In Subsection 1.3.3, we reviewed studies that either classified specific lexical forms into their corresponding communicative functions or focused on classifying different lexical forms into their respective communicative function. For example, Prévot, Gorish, et al. (2015) used two Random Forest classifiers to classify French feedback responses. The first classifier determined the *base functions* of the feedback responses, and if the

1. In the publication, no examples are provided of what VSUs or Non-VSUs are.
2. This is a 'maybe' because some classification studies include an 'other' category capture utterances that share the lexical form of feedback responses but are not feedback responses.

function *evaluation-base* was selected, a second classifier categorized the response into one of the functions in the *evaluation* set, which included *approval*, *expectation*, *amusement*, and *confirmation/doubt*. Instead of this two-step approach, we use a single classifier to automatically annotate the remaining potential feedback responses in the Switchboard corpus by classifying them into one of the ten communicative functions identified during our manual annotations: *Continue*, *Non-understanding*, *Agree*, *Disagree*, *Yes*, *No*, *Sympathy*, *Disapproval*, *Mild Surprise*, *Strong Surprise*, and an additional category for non-feedback *Other*.

3.3. Manual annotations of feedback responses

As mentioned in the previous chapter, we manually annotated 1,627 feedback responses in the Switchboard corpus. However, in total, we identified 26,981 potential feedback responses from the clips. When we searched through the transcriptions of the conversations for the lexical forms listed in Table 2.7, we found 85,956 potential feedback responses. After our paper was published (Figueroa, Adigwe, et al. 2022), we continued with manual annotation, which is why the counts in Table 3.1 do not match those in Table 2.8. In total, we manually annotated an additional 552 feedback responses, bringing up our total count to 2,179 feedback responses, or 2,256 if we include those categorized as *(O) Other*.

During our initial annotation, we sometimes assigned multiple function labels when we could not decide on a single communicative function, such as 'A/C'. For this classification task, however, we decided to allow only one function label. To achieve this, we listened to the feedback responses in context and assigned the function that matched our first reaction. These 2,256 manual annotations, listed in Table 3.1, were used for training and test sets and for evaluating the performance of the classifiers. More details about this process will be provided below. Our manual annotations of the communicative functions of feedback responses in the Switchboard corpus can be found at https://github.com/carolfigPhD/FeedbackAnnotationScheme/blob/main/manual_function_annotations_swb_feedback_.csv.

3.4. Input features

Unlike most perception studies where participants assign communicative functions by listening to feedback responses in isolation (see Subsection 1.3.2), we annotated the feedback responses by listening to them in context. Specifically, the factors we considered when assigning the communicative functions included the lexical form of the feedback response, its prosody, and both preceding and following context. By context, we mean not only what the 'speaker' said but also how it was said (i.e., the prosody). We aimed for the classifier's input features to reflect the factors we used in our manual annotation, hence we represented features from both the feedback response and its context. However, because we want this classifier to be applicable in both offline and online settings, we only used features from the preceding context of

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	56	Yes-answer
(N) No	114	No-answer
(S) Sympathy	82	Sympathy
(Ds) Disapproval	65	Reproach
(MS) Mild Surprise	103	Interest
(SS) Strong Surprise	191	Surprise
(O) Other	77	Other

Table 3.1. – Feedback communicative functions, count of manually annotated data, and corresponding labels in GPT-3 prompt.

the feedback response. For instance, when a user is interacting with a conversational agent, the agent cannot know for certain what the user will say next, although a language model may make predictions. In the following sections, we describe the features used in our classification task.

3.4.1. Feedback features

For the feedback responses, we use both the lexical form and its prosodic realization as features.

The **lexical** forms of the feedback responses are encoded using one-hot encodings via the scikit-learn Python library (Pedregosa et al. 2011). It's important to note that while feedback responses such as *mhm* and *hm* are considered non-lexical, others like *yeah* and *right* are deemed lexical. However, we use the term *lexical form* to encompass both types. This is because feedback responses such as *mhm* and *hm* do not necessarily represent phonetic transcriptions (e.g., IPA transcriptions), and referring to them as *written form* may not be accurate due to the variability in how they are transcribed (e.g., asking individuals to transcribe what they hear).

Five **prosodic** features were extracted from each feedback response: duration, mean pitch, pitch slope, pitch range, and mean intensity. These features were extracted at the "word level." We used Parselmouth (Jadoul et al. 2018) to extract pitch (F_0 Hz) and intensity (dB) values. The pitch values were first transformed to logarithmic scale and then z-score normalized; intensity values were also z-score normalized. This normalization was performed per speaker, using the mean and standard deviation of each speaker computed from their entire conversation. Pitch slope was calculated by subtracting the mean of the z-score normalized pitch values in the second half of the feedback response from the mean of the z-score normalized pitch values in the first

half of the feedback response.

3.4.2. Context features

Context features were extracted from 4000 ms of the interlocutor's utterance preceding the feedback response. Previous work in feedback modeling has extracted features from context by setting either an arbitrary window length or a specific number of words. We experimented with window lengths ranging from 1500 to 4000 ms and found that 4000 ms often captured complete sentences. As previously mentioned, we chose to use only features from the preceding utterance of the interlocutor (and not any future context) to ensure the model's applicability for online classification.

Part-of-speech (POS) tags of the preceding utterance were extracted using the spaCy Python library. The spaCy library offers both fine-grained POS tags, using the *tags* option, and coarse-grained POS tags, using the *pos* option. We opted for the *tags* option. From these tags, POS bigrams were created and then ranked according to their term frequency-inverse document frequency (TF-IDF), treating the ten feedback functions as documents and the POS bigrams as terms. From this sorted list, the top 30 bigrams were selected and encoded as one-hot features. These are listed in Table F.1 in Appendix F.

Dialog Acts were automatically assigned to the interlocutor's utterance using DialogTag (Malik 2020), a Python library that categorizes utterances into 39 dialog acts. Table G.1 in Appendix G provides an example utterance for each dialog act. However, we consolidated the following dialog tags into a single 'Question' tag: 'Yes-No-Question', 'Declarative Yes-No-Question', 'Rhetorical-Question', 'Wh-Question', and 'Tag-Question', since we did not find it necessary to distinguish among these types. The dialog tags were then one-hot encoded.

Sentence Embedding of the preceding utterance was obtained using SimCSE (Gao et al. 2021), an auto-encoding embedding technique based on contrastive learning. During training, SimCSE utilizes BERT encodings of the input and fine-tunes parameters using a contrastive learning objective. This objective encourages the model to draw semantically similar pairs closer together while pushing apart dissimilar pairs. We employed the 'sup-simcse-bert-base-uncased' pre-trained model from SimCSE, which is readily available on Github [3].

3.5. GPT-3

In this work we also experimented with using GPT-3, a large language model developed by OpenAI (Brown et al. 2020), designed to generate human-like text based on input from a human (a prompt). GPT-3, a Transformer model, was trained on a vast array of diverse texts, including content from Wikipedia, the internet, books, and other sources. Due to its extensive training data, the main highlight of GPT-3 is its ability to perform various natural language processing tasks.

3. https://github.com/princeton-nlp/SimCSE

Therefore, we used GPT-3 as a classifier to annotate the remaining potential feedback responses in the Switchboard corpus according to their communicative functions. For classification, we explored three different approaches: *zero-shot*, *few-shot*, and *fine-tuning*. Additionally, as an alternative to the previously mentioned context features, we tested the use of the probability distributions from GPT-3 as inputs for a main classifier. In other words, we use GPT-3 to encode both the interlocutor's preceding utterance and the lexical form of the feedback response.

For *zero-shot* classification, we provided GPT-3 with a prompt similar to the one shown in Table 3.2. In the *zero-shot* approach GPT-3 receives only the instructions, the description of the dialog acts (our communicative functions), the interlocutor's preceding utterance (A's utterance), and the lexical form of the feedback response (B's utterance). The prompt ends with an opening bracket, and GPT-3 is tasked with predicting the next token (marked in bold). This is done using the davinci-003 model via OpenAI's API.

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 3.2. – Zero-shot approach: Prompt given to GPT-3. GPT-3 completes the bold text.

For *few-shot* classification, similarly to the *zero-shot* approach, we provide GPT-3 with the task instructions and descriptions of the dialog acts. However, for the *few-shot* approach, we additionally provide examples for each communicative function (example dialogues). Table 3.3 shows an example of the prompt for this *few-shot* approach. Note that for both the *zero-shot* and *few-shot* approaches, we do not use our manual annotations to train GPT-3 since we employ GPT-3 directly as a classifier.

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] A: i didn't see them B: pardon [misunderstand] A: i don't think they will win B: no [agree] A: running is good for you B: yeah [agree] A: i think deep down they are kind B: hm [disagree] A: have you seen star wars? B: yeah [yes-answer] A: can you swim B: no [no-response] A: i lost my keys and wallet B: oh [sympathy] A: he ran over five people B: ooh [reproach] A: adding wine changes the flavor B: hm [interest] A: she won give medals B: wow [surprise] A: i know but they B: well [other] A: what do you think they don't vote B: um [**other]**

Table 3.3. – Few shot approach: Prompt given to GPT-3. GPT-3 completes the bold text.

prompt	completion
speaker A: that uh or a client that lives on uh Lake Tawakoni speaker B: mhm function:	Continuer###
speaker A: i think it's a little bit inefficient but but it's better than nothing speaker B: yeah function:	Agree###
speaker A: okay and you know and it was just the poor cat was just a mess speaker B: mmm function:	Sympathy###

Table 3.4. – Example of the columns of a csv for fine-tuning GPT-3.

The third approach involves *fine-tuning* GPT-3, wherein we use our manual annotations to fine-tune the model. We opted for the davinci base model, as davinci-003 was not available for fine-tuning at the time of our initial publication. Unlike the other approaches, fine-tuning does not involve instructions or examples within the prompt; instead, the model is trained directly on example pairs, the input text (the interlocutor's preceding utterance and the lexical form of the feedback response) with the corresponding output (the function label). Table 3.4 illustrates the format of the csv [4] used for fine-tuning GPT-3. The csv format includes two columns: the first contains the prompt, and the second specifies the completion that GPT-3 should generate, previously requiring a termination symbol–hence our use of three hashtags at the end of each completion.

Note that in the prompts for *zero-shot* (Table 3.2) and *few-shot* (Table 3.3) classification, we modified the feedback function labels from those listed in the first column of Table 3.1. Since GPT-3 generates word pieces, we changed the feedback function labels in the prompts to avoid starting with the same first letters [5]. This change allowed us to easily identify the function based on the first generated word piece from GPT-3.

As previously mentioned, for the *zero-shot* and *few-shot* approaches, we also explored using the prediction as an input feature for a main feedback function classifier instead of using it directly. For this, we use the probability distribution that GPT-3 outputs for potential function labels (or rather their prefix). From GPT-3, we obtain the top five predicted labels (word pieces) and their corresponding probabilities. For

4. OpenAI's API previously had a pre-processing command for converting a csv file to the required json file format for fine-tuning. This is no longer supported, and the format fine-tuning newer models has changed.

5. Although *Sympathy* and *Surprise* begin with the same letter, they do not have the same first letters.

example, given the feedback response *yeah*, GPT-3 might 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 the feedback function *(A) Agree* is assigned 77%, *(C) Continue* 1.7%, *(Y) Yes-answer* 21.3%, and all other functions, including the *Other* category are assigned 0%. These probability distributions are then used as input features for 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.

3.6. Classifiers

The task of the main classifier is to classify the communicative function of the feedback response using the features listed above. As explained earlier, 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. Given the limited size of our dataset, we opted for three machine learning models that can handle small data sets (Forman et al. 2004): Support Vector Machine (SVM), Logistic Regression, and Random Forest, using the classifiers implemented in scikit-learn a Python library (Pedregosa et al. 2011). As mentioned in Subsection 1.3.3, similar classifiers, such as SVM (Lai 2008), Decision tree (Heylen et al. 2007; Truong and Heylen 2010), and Random Forest (Prévot, Gorish, et al. 2015), have been used in previous classification studies.

For the three classifiers, we set the class_weight parameter to 'balanced' to adjust for our imbalanced annotations. We used a linear kernel for the SVM classifier and, although we experimented with a radial basis function kernel, the linear kernel yielded the best results. For the Logistic Regression classifier, we set max_iter to 200. Apart from these specified adjustments, all other parameters for each classifier were kept to their default settings. Overall, the SVM classifier with the linear kernel outperformed the others, and therefore, we only report results from this classifier, alongside the results of using GPT-3 as a main classifier.

We experiment with various combinations of input features and evaluate our SVM classifiers using 10-fold cross validation. To assess model performance, we used the F1 score. In cases where we used GPT-3 directly as a classifier without fine-tuning, we did not use cross-validation; instead, we applied the model to our entire annotated dataset.

3.7. Results

In this section we present the results of our experiments with various combinations of input features. We also use one of our best-performing classifier models to classify the communicative functions of the remaining potential feedback responses. By using both manual and automatic annotations, we explore human conversational behavior by investigating sex differences and conducting an overlap analysis of feedback

responses in the Switchboard corpus. These examples demonstrate the utility of our annotations.

3.7.1. Performance of classifiers

Table 3.5 summarizes the F1 scores of the various models using different combinations of input features. For comparison, we report results from a majority-class baseline, which selects the most frequent class label. Additionally, we also report our inter-annotator agreement annotations from 1,124 annotations from our first publication (Figueroa, Adigwe, et al. 2022) for comparison. In cases where the annotators could not decide on a single function (e.g. 'A/C'), we randomly selected one of the functions when calculating the F1 score. This procedure was averaged over 10 iterations.

When only lexical features are used (Model 1), we achieve a relatively high F1 score (0.63), which surpasses the baseline. However, the prosodic features (Models 2-6) are not very informative, as they perform worse than the baseline. Although combining all prosodic features (Model 15) does result in a performance better than the baseline, adding these prosodic features to the lexical features (Model 16) does not improve the performance further, as the F1 score remains the same (0.63).

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 particularly useful (Models 10,11,13,14). This is notable since GPT-3 also encodes lexical information about the feedback response, yet its performance is on par with Model 1, which only uses lexical features. Moreover, there is no significant difference between the zero-shot or few-shot features (Model 13,14) ($t(18) = 1.585$; $p = 0.13$). In contrast, combining lexical features with SimCSE results in better performance: Model 19 performs as well as the inter-annotator agreement score. Figure 3.3 shows the confusion matrix of Model 19, which was trained on 1804 examples and evaluated on a test set of 452. Specifically, Model 19 performs poorly in classifying functions *(D) Disagree*, *(Ds) Disapproval*, and *(Y) Yes*, which could be due to the limited number of training examples.

To further improve this score, we included the prosodic features, the dialog act, and the GPT-3 distributions to the lexical and SimCSE features (Model 20). Although Model 20 shows a statistically significant improvement over Model 19 ($t(18) = 2.509$; $p = 0.02$), the margin of improvement is modest given the substantially larger feature set. Figure 3.4 shows the confusion matrix for Model 20, trained on the same 1804 examples and evaluated on the same 452 test set. The addition of the prosodic features, the dialog act, and the GPT-3 distribution features has improved the classification for *(C) Continue*, *(SS) Strong Surprise*, *(MS) Mild Surprise*,*(D) Disagree*, *(U) Non-Understanding*, and *(Y) Yes*. However, it also reduces the classification performance for *(A) Agree*, *(O) Other*, and *(S) Sympathy*.

The best performing model is the GPT-3 fine-tuned classifier (Model 12), which performs significantly better than Model 20 ($t(18) = -2.803$; $p = 0.01$). This model achieves a F1 score higher than the inter-annotator agreement. Figure 3.5 shows the

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

confusion matrix for Model 12, which was trained on the 1804 examples and evaluated on the 452 test set. Compared to Model 20, Model 12 improves the classification for *(C) Continue*, *(SS) Strong Surprise*, *(S) Sympathy*, *(O) Other*, *(U) Non-understanding*, *(DS) Disapproval*, and *(Y) Yes*. However, it also reduces the classification performance for *(MS) Mild Surprise*, *(D) Disagree*, and *(N) No*.

In Appendix H, we display all the confusion matrices for the remaining models listed in Table 3.5. All of these models are trained on the same 1804 examples and are evaluated on the same 452 test set. However, since Models 10 and 11 use GPT-3 directly as a classifier, we do not train these models, but we do report their confusion matrices for the same 452 test set. To summarize, Models 12, 19, and 20 are all viable classifiers for feedback functions, each perform on par with the inter-annotator agreement F1 score. The choice of classifier depends on specific requirements, such as whether it will be used in an offline or online setting and whether access to GPT-3 is available.

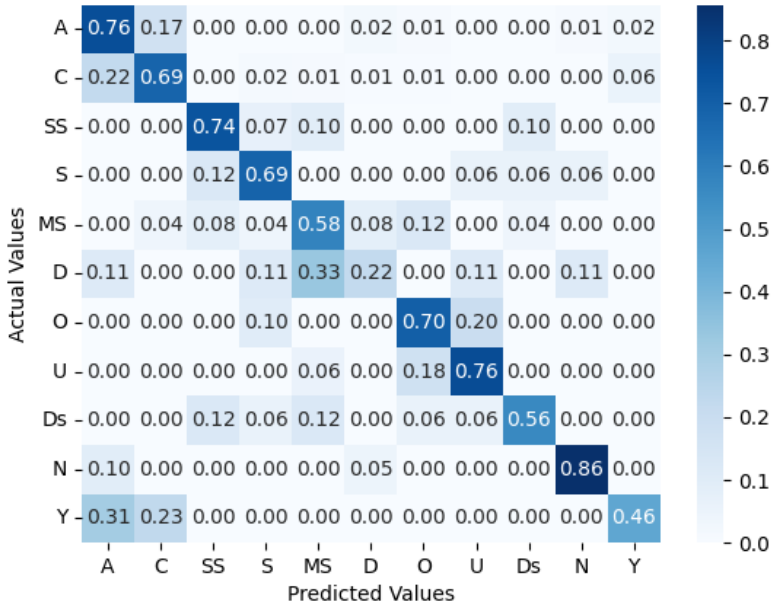


Figure 3.3. – Confusion matrix for Model 19.

3.7.2. Labeling remaining potential feedback responses

To gain insights into the overall distribution of communicative functions of feedback responses in the Switchboard corpus, we applied one of our best performing classifiers to the remaining 83,700 potential feedback responses. Among the three viable classifiers (Models 12, 19, 20), we chose Model 19 because it offers performance on par with human inter-annotator agreements and more importantly it is cost-effective. The distributions are shown in Table 3.6, which includes the distributions of the 2,256 manually annotated feedback responses and the 83,700 automatically annotated feedback responses. In total, there were 74,106 instances of actual feedback responses (not

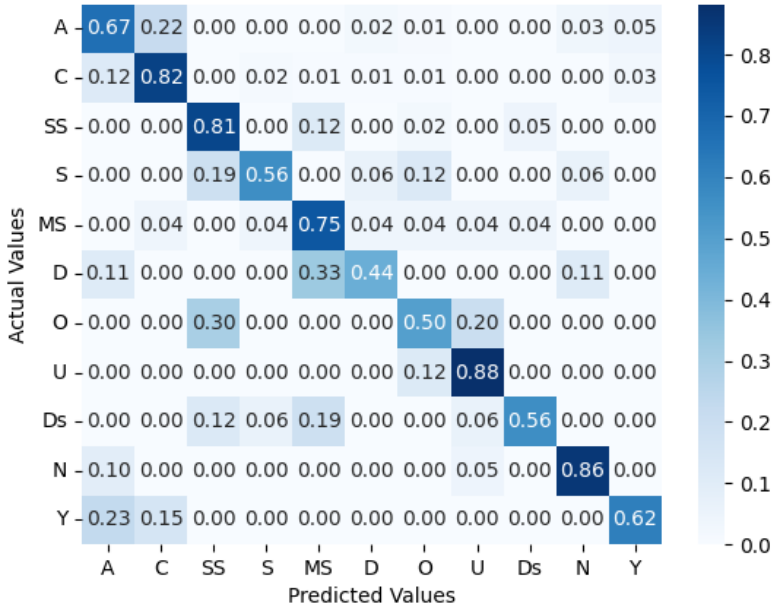


Figure 3.4. – Confusion matrix for Model 20.

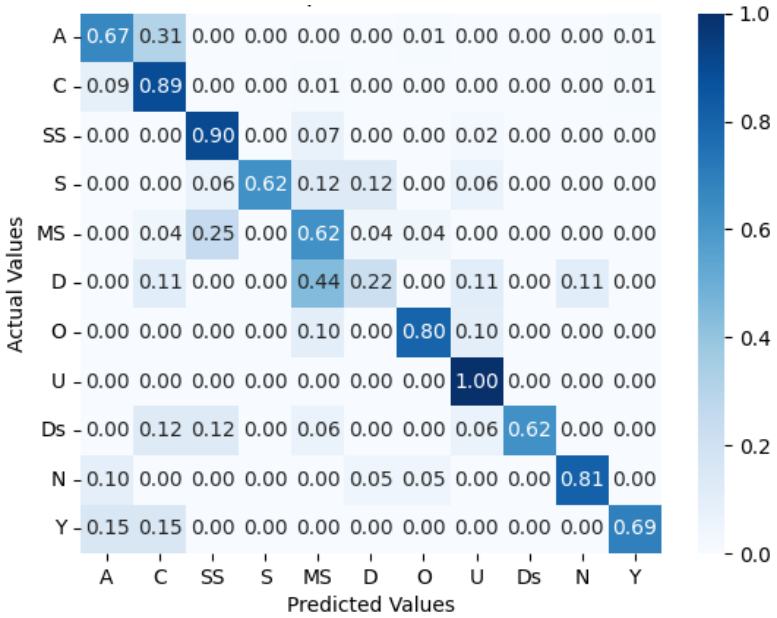


Figure 3.5. – Confusion matrix for Model 12.

including *(O) Other*), according to the classifier. As shown, *(C) Continue* and *(A) Agree* are the most frequent feedback functions. Our classification results from Model 19 on the communicative functions of feedback responses in the Switchboard corpus can be found at https://github.com/carolfigPhD/FeedbackAnnotationScheme/blob/main/automatic_function_annotations_swb_feedback_model_19.csv.

Function	Total	Tot. %	F %	M %	FF %	FM %	MM %	MF %	Ov %
(C)	39499	51.8	**54.2**	**48.8**	54.7	53.6	48.4	49.3	45.3
(U)	342	0.45	**0.34**	**0.58**	0.31	0.39	0.56	0.60	26.9
(A)	22809	29.9	**26.3**	**34.3**	26.3	26.3	**35.3**	**33.2**	45.3
(D)	986	1.29	**1.12**	**1.51**	**0.96**	**1.32**	1.56	1.44	36.0
(Y)	4101	5.38	**4.93**	**5.93**	5.05	4.77	5.90	5.97	45.1
(N)	787	1.03	1.02	1.05	0.96	1.08	**0.90**	**1.24**	34.6
(S)	1775	2.33	**3.19**	**1.25**	3.29	3.06	**1.01**	**1.55**	46.7
(Ds)	638	0.84	0.81	0.86	0.78	0.85	0.80	0.94	39.8
(MS)	2325	3.05	3.03	3.07	2.82	**3.31**	3.01	3.13	37.7
(SS)	3023	3.96	**5.03**	**2.64**	4.81	5.31	2.62	2.66	41.8
(O)	9671								

Table 3.6. – Distribution of manually annotated and automatically annotated feedback responses 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 responses, etc. Bold numbers denote significant deviations from the expected distribution ($\alpha = 0.05$). Ov=Overlap.

3.7.3. Investigating sex differences

To illustrate how the annotations obtained from our classification task can be used to further analyze human conversational behavior, we broke down these numbers based on the sex of the listener (i.e., the interlocutor producing the feedback responses), as provided in the Switchboard corpus. This is shown in Table 3.6. Note that the metadata of the Switchboard corpus provides only two options for sex: female and male. A chi-square test revealed that sex influences the communicative function of feedback responses ($\chi^2(9) = 1165.71, p < .001$). Analysis of the standardized residuals ($\alpha = 0.05$) revealed significant differences in most feedback functions, as detailed in Table 3.6. Notably, the use of *(S) Sympathy*, and *(SS) Strong Surprise* occurs much more frequently among females than males. To further investigate whether these effects are also influenced by the sex of the interlocutor receiving the feedback responses, we analyzed the data split by the sex of both interlocutors, as shown in Table 3.6. Chi-square tests confirmed such effects, both when feedback responses were 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, *(S) Sympathy* is used less frequently compared to compared male-female conversations.

Additionally, our analysis shows that in general, females produce 2.73 feedback responses per minute, whereas males produce 2.23 feedback responses per minute. Conflicting results have been found in other studies. For example, Ueno (2004) found that Japanese women provide more feedback responses than Japanese men, and Kjellmer (2009) observed similar results for British women. However, Stubbe (1998) found that Maori and Pakeha women produced fewer feedback responses than their male counterparts. It is important to note that our findings are specific to the Switchboard corpus and may not be generalizable to other corpora. Moreover, we refrain from making blanket statements based on these results to avoid overgeneralizing about sex differences.

3.7.4. Analysis of Overlap

Another example in which our automatic classification of the communicative functions of feedback responses can be used to gain insights into human behavior is to analyze whether certain feedback functions tend to overlap more with the speech of the interlocutor. To determine whether a feedback response was overlapping, we checked the start time of the feedback response against the speech timeline of the interlocutor. If the feedback start timestamp fell during or at the start of the interlocutor's speech, we classified the feedback response as overlapping. Conversely, if the feedback's start time occurred during a pause or laughter from the interlocutor, we classified it as not overlapping. The percentage of overlap for each feedback communicative function is presented in Table 3.6.

Using this method, we find that *(U) Non-understanding*, *(D) Disagree*, *(N) No*, *(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. Similarly, listeners might also delay producing feedback functions with negative connotations such as *(D) Disagree*, *(N) No*, and *(Ds) Disapproval*, to decide whether to take the turn or to further respond to what the interlocutor has said.

We had expected feedback functions such as *(C) Continue*, *(A) Agree*, *(S) Sympathy*, *(MS) Mild Surprise*, and *(SS) Strong Surprise* to overlap with the speech of the interlocutor. However, we found that for *(C) Continue*, *(A) Agree*, *(S) Sympathy*, and *(SS) Strong Surprise*, these feedback functions almost equally overlap and not overlap. Further analysis should examine whether the silences during the interlocutor's speech are short breaths or longer pauses. It would be interesting to conduct an analysis similar to that of C. Goodwin (1986), they compared *assessments* and *continuers*. C. Goodwin (ibid.) found that although both *assessments* and *continuers* occur 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. Such an analysis, considering the conversational units of the interlocutor, may provide

deeper insights into where exactly these feedback functions occur within the interlocutor's turn. It could also reveal whether feedback functions occurring during the interlocutor's silences happen between or within turn-constructional units.

3.8. Discussion

Although fine-tuning GPT-3 (Model 12) performs the best, it may not be suitable for online settings or annotating large corpora. Firstly, not everyone has access to GPT-3 since it is a paid service and relies on OpenAI's API, which can experience downtime and incurs significant costs. Secondly, in terms of reproducibility, GPT-3 is like a black-box; when fine-tuning, we do not know exactly what is fine-tuned. In contrast, Model 19 (SimCSE + Lexical), which classifies feedback functions on par with human annotators, could be a viable alternative for online settings or for annotating large corpora. Furthermore, SimCSE is an open source model and is available at no cost.

However, if costs are not a restriction, further experimentation with GPT-3 could be done. From our results, we observed that without fine-tuning, the probability distributions from zero-shot and few-shot classification approaches did not perform better than the lexical one-hot encodings. Experimentation with different prompts might improve the performance of GPT-3 features. Due to our limited annotations, we could not train GPT-3 on a separate training set or evaluate the probability distributions from the fine-tuned model because of the 10-fold cross-validation constraints; however, this remains a possibility for future work. The probability distribution features from the fine-tuned model could then be integrated with other features. Another area for exploration could involve incorporating prosodic information into the prompts by discretizing them. For example, pitch slope could be discretized by describing it as flat, rising, or falling, thereby enriching the prompts with descriptive prosodic information.

Another feature that did not perform well in the classification task was the dialog acts (Model 8). This poor performance may be attributed to the preceding context being misclassified with the incorrect dialog act. From the confusion matrix in Figure H.8 in Appendix H, it is evident that for almost all the communicative functions of feedback responses, the dialog act feature erroneously classified the response as *Strong Surprise*. Consequently, these errors were propagated through the model. A similar issue of error propagation could have occurred with the probability distributions by GPT-3 (Models 13, 14). For future work, further analysis should be done on how these errors affect the model.

During the manual annotations, the prosody of the feedback response was an important factor in assigning their communicative function. However, from our results, we observed that the selected prosodic features performed poorly (see Figure H.14 in Appendix H), and ultimately did not contribute significantly to the best-performing models. Surprisingly, the best model (Model 12), using a fine-tuned GPT-3, did not use any prosodic features. This finding is intriguing since prosody should aid in disambiguating feedback responses that are not easily classified based solely on textual information, such as *no* used as negative agreement. Alternatively, it could be that

the preceding context provides sufficient information, possibly helping disambiguate a question (preceding a *(N) No*) from a statement (preceding a *(Ds) Disapproval)*. Therefore, our results raise a new question: Were our prosodic features redundant given the context, or were they inadequate representations of prosody?

During Jurafsky et al. (1998)'s initial annotations of *backchannels* in the Switchboard corpus, they did not listen to the audio; annotations were made solely from reading the transcripts of the conversation. However, during subsequent partial re-annotations, they found that *backchannels* initially annotated as *continuers* were reclassified as *assessments*. Similarly, Shriberg et al. (1998) found prosody to be helpful in classifying different dialog acts. In contrast, Gravano, Beňuš, Hirschberg, et al. (2007) who classified affirmative words, found that the most important features were textual and context features; acoustic/prosodic features did not improve the performance of their classifier. Likewise, in Beňuš (2016)'s annotation of discourse functions for affirmative words in Slovak, the annotations were performed using only the transcripts, and audio was consulted only when the text was ambiguous. Thus, it appears that there are feedback responses where the communicative function can be determined without prosody.

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 significantly worse than the distributed representations (SimCSE). Therefore, it would be interesting to plot the prosodic representations of feedback responses from both discrete and distributed representations (e.g., prosody embeddings) in order to visualize potential clusters. Note that the results from our Tukey tests, in which we found significant differences in prosodic features between communicative functions (Tables E.1, E.2, E.3, and E.4), do not always reflect the results in our classification of the 402 feedback responses in the test set (Figures H.2, H.3, H.4, H.5).

During our annotations, one factor we considered but did not include in our input features, was the prosody of the context. Future work should represent prosodic features using distributed representations such as wav2vec (Schneider et al. 2019) or HuBERT (Hsu et al. 2021) embeddings. It would be interesting to investigate whether the prosody of the context would be redundant compared to the SimCSE embeddings or if they would improve the classification performance.

A primary limitation of this work is that our classification models have only been trained and evaluated using the Switchboard corpus. Future work could extend this investigation to include the performance of our best models on other corpora, such as face-to-face corpus like the Candor corpus.

3.9. Summary

In this chapter, we presented the results from our experiments, where we explored the performance of various models and combination of features for classifying the communicative functions of feedback responses in the Switchboard corpus. Specifi-

cally, the feedback responses were classified into ten communicative feedback functions: *continue*, *non-understanding*, *agree*, *disagree*, *yes/no response*, *mild/strong surprise*, *sympathy*, and *disapproval*. We experimented with different combinations of lexical and prosodic features from the feedback responses, as well as context features from the preceding utterance of the interlocutor, using these as input for an 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 these approaches with a fine-tuned GPT-3 classifier.

Our experiments demonstrated that using just lexical features and SimCSE embeddings yields a fairly good performance, on par with inter-annotator agreement. While using GPT-3 in a zero-shot or few-shot manner does not contribute significantly, a fine-tuned GPT-3 model outperforms all other models, even without the use of prosodic features. We conclude that the prosodic features may be redundant given the SimCSE context embeddings, or the prosodic representations used here may be inadequate representations, whereas distributed representations may be better. However, our automatic and manual annotations are available and could be used to train a prediction model to make fine-grained function predictions for a conversational system.

4. Prosodic alignment in different feedback functions

Summary

4.1. Introduction and Motivation . . . 138
4.2. Annotations used for measuring alignment . . . 140
4.3. Extraction of prosodic features . . . 141
4.4. Measuring local alignment . . . 142
4.5. Results . . . 143
4.5.1. Mean Pitch . . . 144
4.5.1.1. Feedback response vs. preceding utterance . . . 145
4.5.1.2. Feedback response vs. following utterance . . . 145
4.5.1.3. Preceding utterance vs. following utterance . . . 145
4.5.2. Mean Intensity . . . 146
4.5.2.1. Feedback response vs. preceding utterance . . . 146
4.5.2.2. Feedback response vs. following utterance . . . 147
4.5.2.3. Preceding utterance vs. following utterance . . . 147
4.5.3. Pitch Slope . . . 148
4.6. Discussion . . . 148
4.7. Summary . . . 149

4.1. Introduction and Motivation

This chapter is based on our publication *"Prosodic alignment in different conversational feedback functions"* (Figueroa, Beňuš, et al. 2023), which was presented at the 20th International Congress of Phonetic Sciences (ICPhS) 2023, in Prague, Czech Republic.

The main goal of this work was to address the following research question:

R4 *Does prosodic alignment occur for different communicative functions of feedback responses?*

This research question uses components from the proposed vocal feedback model: the communicative functions of feedback responses, the prosodic features of feedback responses, and the features from the user. These components are circled in the proposed vocal feedback model in Figure 4.1.

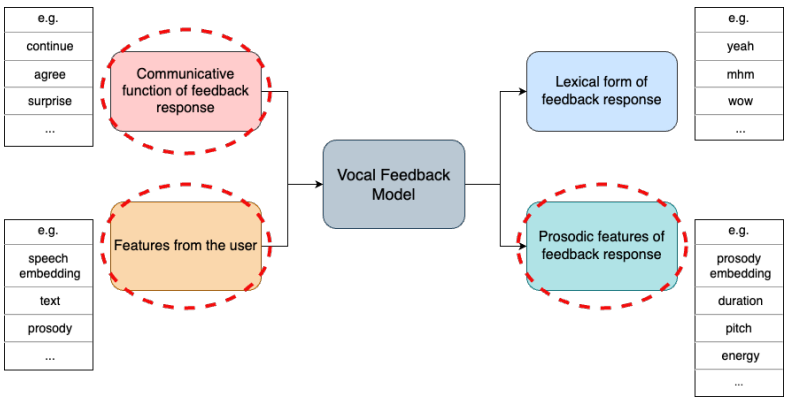


Figure 4.1. – Components from the vocal feedback model addressed by R4.

When feedback responses are described, a key characteristic often highlighted is their unobtrusiveness to the flow of the conversation. Feedback responses do not interrupt the speaker; rather, the listener produces them without disrupting the ongoing conversation. Heldner, Edlund, and Hirschberg (2010) and R. Levitan, Beňuš, et al. (2015b) have suggested that this unobtrusiveness is due to *prosodic alignment*, meaning the listener produces feedback responses with similar prosody to the prosody of the interlocutor's preceding utterance. However, previous studies investigating prosodic alignment between feedback responses and the interlocutor's preceding utterance (see Subsection 1.4.3) have focused solely on feedback responses that signal the listener is attentive and that the speaker should continue speaking (i.e., the *continue* function in our annotation scheme of communicative functions of feedback responses). These studies have excluded other communicative functions, such as sympathy, surprise, and disapproval. Therefore, if conversational agents are to produce feedback responses that adapt their prosody to that of their interlocutors as in human-human conversations, we must first determine if this prosodic alignment occurs for other communicative functions of feedback responses.

There are two other limitations in previous studies investigating prosodic alignment between feedback responses and the interlocutor's preceding utterance, as mentioned in Subsection 1.4.3. The first limitation is that these studies in U.S. English have used the same task-oriented corpus, the Columbia Games Corpus. The findings from task-oriented conversations may not reflect those of spontaneous conversational speech. As Beňuš (2016) pointed out, task-oriented speech might exhibit prosodic patterns similar to those found in question-answer pairs or request-answer pairs.

The second limitation of previous studies in U.S. English is that they have only measured one type of alignment, *proximity*, which measures the absolute difference between prosodic features. For an overview of different types of alignment, see Section 1.4.2. However, studies in other languages have measured *synchrony* (Gustafson et al. 2010; Kawahara, Yamaguchi, Uesato, et al. 2015; Kawahara, Uesato, et al. 2015; Ochi et al. 2024) in spontaneous conversational corpora. Unlike proximity, synchrony measures the similarity in movement. For example, if a female user interacts with a

conversational agent with a male voice, their pitch differences may be substantial (i.e., low proximity) due to the typically higher pitch of female voices compared to male voices. However, if the conversational agent raises its pitch when the user raises hers, and lowers it when she lowers hers, their pitch would have similarity in movement (i.e., high synchrony). For a visualization of combinations of high/low proximity and high/low synchrony, see Figure 1.8 in Subsection 1.4.2.

The validity of conclusions from U.S. English studies on prosodic alignment between feedback responses and preceding utterances is uncertain, given the type of corpus and type of alignment measured. Additionally, Heldner, Edlund, and Hirschberg (2010) found the pitch of *backchannels* to be higher in general, and previous research has shown that a rising pitch can serve as a backchannel-inviting cue (Gravano and Hirschberg 2009), these observations might explain the findings. Consequently, it remains unclear whether interlocutors truly align, in the sense of the listener adjusting the pitch level of the backchannel to match the pitch level of the interlocutor's preceding utterance.

Alignment in human-human conversations has been shown to have a positive effect; for example, when interlocutors align, their interactions are more successful and feel more natural (Nenkova et al. 2008). Beňuš (2014) advocates for implementing speech alignment in interactions with social robots to enhance their naturalness and effectiveness. There has also been interest in incorporating speech alignment behavior in spoken dialogue systems (R. Levitan, Beňuš, Štefan, et al. 2016; Hoegen et al. 2019; Yamazaki et al. 2021). In our study, if prosodic alignment is observed for certain feedback functions in human-human conversations, this alignment could be implemented in spoken dialogue systems or conversational agents, potentially making their feedback responses more human-like. The results of our analyses could inform researchers who are interested in developing adaptive feedback responses in spoken dialogue systems or conversational agents.

Therefore, in this chapter, we present our results from our study investigating prosodic alignment between feedback responses and both the preceding and following utterances of the interlocutor. Furthermore, we explore whether this alignment occurs across the ten communicative functions found in the Switchboard corpus from our previous work (see Chapter 2): *(C) Continue*, *(U) Non-Understanding*, *(A) Agree*, *(D) Disagree*, *(Y) Yes*, *(N) No*, *(S) Sympathy*, *(Ds) Disapproval*, *(MS) Mild Surprise*, and *(SS) Strong Surprise*. Specifically, we measure prosodic alignment for the following prosodic features: mean pitch, mean intensity, and pitch slope.

4.2. Annotations used for measuring alignment

For this study, we used our manual annotations of feedback responses in the Switchboard corpus, as listed in Table 4.1. In total, we had 2,118 manually annotated feedback responses. Note that this number is greater than our initial annotations listed in Table 2.8 because manual annotations continued after the publication of our first paper (Figueroa, Adigwe, et al. 2022). However, the total count is slightly less than those

reported in the previous chapter (see Table 3.1). The slightly lower count results from the fact that this alignment study was conducted simultaneously with the ongoing classification study but was completed before the classification study [1].

Function	Count
(C) Continue	1004
(U) Non-Understanding	61
(A) Agree	413
(D) Disagree	43
(Y) Yes	57
(N) No	107
(S) Sympathy	83
(Ds) Disapproval	65
(MS) Mild Surprise	99
(SS) Strong Surprise	186

Table 4.1. – Feedback communicative functions and count of manually annotated data used in alignment study.

4.3. Extraction of prosodic features

As previously mentioned, we measure prosodic alignment using the following features: mean pitch, mean intensity, and pitch slope. These features are extracted from the entire feedback response, as well as from the last 500 ms of the interlocutor's preceding utterance and the first 500 ms of the interlocutor's following utterance. The 500 ms window length was chosen because it has been commonly used in previous studies (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Gravano, et al. 2011; Kawahara, Yamaguchi, Uesato, et al. 2015; Kawahara, Uesato, et al. 2015; Beňuš 2016). By considering the interlocutor's following utterance, we gain insights into whether the interlocutor aligns to the 'listener'. This approach allows us to explore the directionality of alignment.

To extract pitch (F_0 Hz) and intensity (dB) from the speech signal every 10 ms, we used Parselmouth (Jadoul et al. 2018). F_0 values were first transformed to a log scale and then z-score normalized. Intensity values were similarly z-score normalized. We normalized these features by speaker, using each speaker's mean and standard deviation computed from their entire conversation. Pitch slope was calculated as described previously in Section 2.6.3.3.

1. The alignment study was conducted during my secondment at the Slovak Academy of Sciences in collaboration with Štefan Beňus. Consequently, the classification study was paused during my secondment to submit this alignment study for the ICPhS submission deadline.

To obtain features for the preceding 500 ms of the interlocutor, we first extracted features from a 2000 ms window before the start of the feedback response. From this list of features, we searched for the last voiced frame, then from the last voiced frame we took the preceding 49 frames. Concretely, this means that the preceding 500 ms of the interlocutor consists of the last voiced frame and the preceding 49 frames, 50 frames in total. Similarly, for the following 500 ms of the interlocutor, we first extracted features from a window of 2000 ms after the end of the feedback response. From this list of features we searched for the first voiced frame, then from the first voiced frame we took the following 49 frames, also totaling in 50 frames.

4.4. Measuring local alignment

As mentioned in Subsection 1.4.3, previous studies in U.S. English that investigated prosodic alignment between feedback responses and the interlocutor's preceding utterance have only measured *static local proximity*, *static global proximity*, or *dynamic global proximity* (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Gravano, et al. 2011; R. Levitan, Beňuš, et al. 2015b). These types of alignments are defined within Wynn et al. (2022)'s classification framework, which is reviewed in Subsection 1.4.2.

In this study, the type of alignment we measure is *static local synchrony* between the preceding and following utterance of the interlocutor and the feedback response of the listener. The alignment is considered *static* because it does not take into account how the alignment changes across the entire conversation. It is *local* because we measure alignment between adjacent turns [2]. We measure synchrony, which refers to the similarity in movement. Explanations of these classification factors in Wynn et al. (ibid.)'s framework can be found in Subsection 1.4.2.

R. Levitan, Beňuš, et al. (2015a) describe local alignment as "a continuous reaction to one's interlocutor and updating of one's own output in response to what has just been heard". This is what we aim to model. For instance, if the user has the floor and raises their pitch and it is an appropriate moment for the conversational agent to give a feedback response, the conversational agent should also raise its pitch. Conversely, if the user lowers their pitch, the conversational agent should also lower its pitch when giving feedback responses. In this study, we are interested in synchrony, rather than measuring the absolute difference (distance) in prosodic features, which is referred to as proximity. In their investigation of local alignment in acoustic-prosodic features across various languages, including Chinese, English, Slovak and Spanish, R. Levitan, Beňuš, et al. (ibid.) found that synchrony was a more useful measurement as it characterized "the way in which human interlocutors adjust to each other at the local level" than proximity. In some languages, there was no evidence of proximity, but evidence for synchrony was found.

We measure static local synchrony using Pearson's correlation coefficient of prosodic features between:

2. We acknowledge that there is no consensus on whether feedback responses constitute a separate turn.

— the preceding utterance of the interlocutor and the feedback response
— the following utterance of the interlocutor and the feedback response
— the preceding utterance of the interlocutor and the following utterance of the interlocutor

Figure 4.2 visually illustrates the correlations; note the figure does not depict overlapping speech, although there were instances where overlapping speech could have occurred. We chose Pearson's correlation to measure synchrony because correlations have been used in other studies (Kawahara, Yamaguchi, Uesato, et al. 2015; Kawahara, Uesato, et al. 2015; Ochi et al. 2024) and it captures the following patterns:

— Positive correlation: prosodic features move up or down together, positive alignment
— Negative correlation: prosodic features move in opposite direction, negative alignment (i.e., misalignment or divergence)

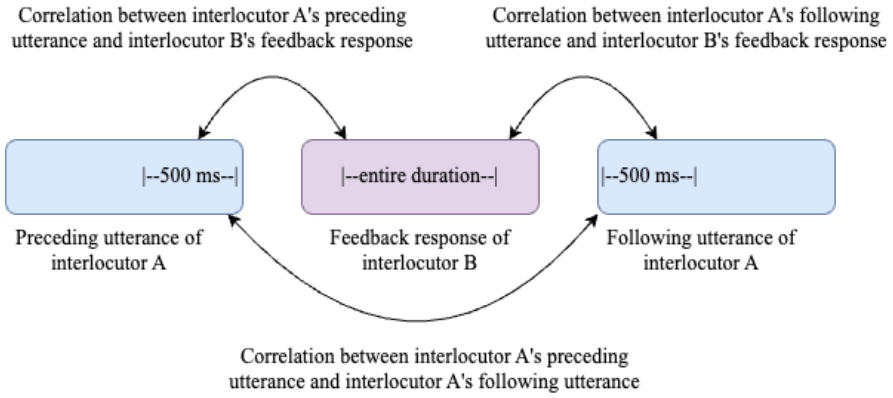


Figure 4.2. – The three correlations computed. Interlocutor A is the "speaker" and Interlocutor B is the "listener". The boxes represent speech.

We will interpret the correlations in the following ways:

— Correlation between the listener's feedback response and the preceding utterance of the interlocutor indicates that the listener aligns to the interlocutor.
— Correlation between the listener's feedback response and the following utterance of the interlocutor indicates that the interlocutor aligns to the listener's feedback.
— Correlation between the preceding and the following utterance of the interlocutor does not indicate any alignment between the interlocutors, but the feedback response can be seen as less obtrusive.

4.5. Results

In this section, we report the Pearson correlation coefficients for the prosodic features between different feedback functions and the 500 ms preceding and following utterance of the interlocutor. Given that we make 30 comparisons for each prosodic

feature (10 functions * 3 correlations), we apply a Bonferroni correction to mitigate the risk of Type I errors. With the Bonferroni correction, correlations with $p < .0016$ ($\alpha = .05$) are considered significant. However, since the Bonferroni correction is a very conservative correction, we also report correlation coefficients which could be considered marginally significant ($p < .05$ and $p < .01$), and which could warrant further investigation in future studies. We denote the strengths of the correlation coefficients as: strong correlation ($r > .5$), moderate correlation ($.3 < r < .5$), and weak correlation ($0 < r < .3$).

We also conducted the analyses excluding feedback responses that contained overlapping speech. However, this exclusion did not alter the observed patterns in the results. Therefore, we report the results using the original 2,118 annotations of feedback responses.

4.5.1. Mean Pitch

Figure 4.3 displays the significant results of Pearson's correlations for mean pitch between the feedback functions and the preceding and following 500 ms utterances of the interlocutor, as well as between the interlocutor's preceding and following utterances. The results for all Pearson's correlations that were not significant, along with the significant results in Figure 4.3 for mean pitch can be found in Section I.1 in Appendix I.

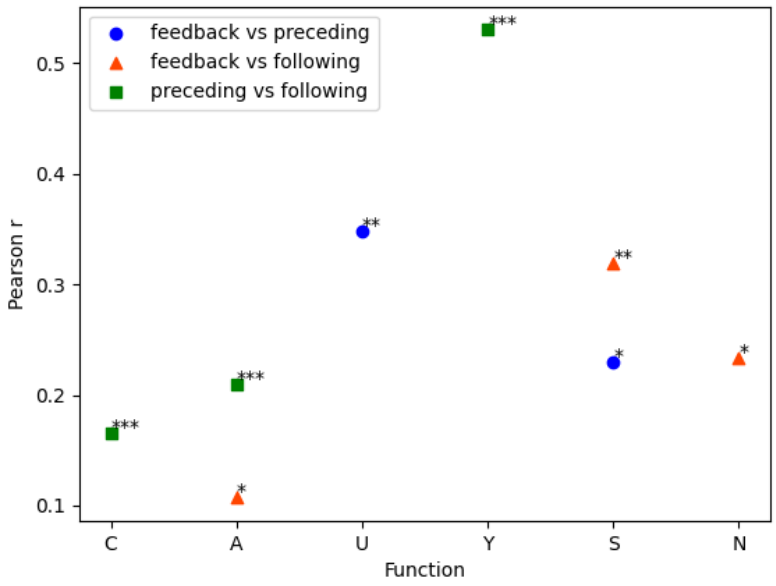


Figure 4.3. – Pearson's correlation coefficients between the mean pitch of feedback functions and preceding/following 500 ms of the interlocutor. *$p < .05$, **$p < .01$, ***$p < .001$

4.5.1.1. Feedback response vs. preceding utterance

Using Bonferroni correction, we did not find any significant results indicating that the listener aligns their mean pitch of their feedback response to the mean pitch of the preceding utterance of the interlocutor. The feedback function *(C) Continue* in our annotation scheme (Figueroa, Adigwe, et al. 2022) serves the same communicative function as *backchannels*. Heldner, Edlund, and Hirschberg (2010) reported that backchannels have more similar pitch to the preceding utterance of the interlocutor than other turn-shifts. Similarly, R. Levitan, Beňuš, et al. 2015b reported that backchannels have more similar pitch to the preceding utterance of the interlocutor than smooth switches, suggesting an indication of prosodic alignment.

These observations contrast with our results regarding feedback functions *(C) Continue*. The discrepancy in findings may be attributed to differences in the corpora used, as both Heldner, Edlund, and Hirschberg (2010) and R. Levitan, Beňuš, et al. (2015b) used the task-oriented Columbia Games Corpus. As noted by Beňuš (2016), task-oriented speech might exhibit prosodic patterns similar to those found in question-answer pairs or request-answer pairs, and these prosodic patterns may not be similar in spontaneous conversational speech. Additionally, both studies using the Columbia Games corpus, Gravano and Hirschberg (2009) found that the utterances of the interlocutor preceding backchannels had a rising pitch and Heldner, Edlund, and Hirschberg (2010) found that in general backchannels had a rising pitch. The difference in results could also stem from types of alignment measured. We measured synchrony, whereas Heldner, Edlund, and Hirschberg (2010) and R. Levitan, Beňuš, et al. (2015b) measured proximity.

4.5.1.2. Feedback response vs. following utterance

We find only marginally significant results for the feedback functions *(A) Agree*, *(S) Sympathy*, and *(N) No*. Both *(A) Agree* and *(N) No* have positive weak correlations, with ($r = 0.107$, $p < .05$) and ($r = 0.234$, $p < .05$) respectively. Meanwhile, *(S) Sympathy* is moderately correlated ($r = 0.319$, $p < .01$). These results suggest that the interlocutor may align to the listener in terms of pitch for these specific feedback functions.

When comparing the similarity of mean pitch between backchannels and the following utterance of the interlocutor to that between smooth switches, R. Levitan, Beňuš, et al. (2015b) found that the interlocutor's utterances following backchannels were more similar in terms of mean pitch compared to smooth switches. However, this significance disappeared when the latency between these transitions were matched. For the feedback function *(C) Continue* (i.e., backchannels, we find no significant evidence of the interlocutor aligning to the listener.

4.5.1.3. Preceding utterance vs. following utterance

We find significant evidence suggesting that the mean pitch of functions *(C) Continue*, *(A) Agree*, and *(Y) Yes* can be seen as less obtrusive to the conversation. There is a weak positive correlation between the mean pitch of the preceding and following

utterances for the feedback function *(C) Continue* ($r = 0.165$, $p < .001$). Similarly, for the feedback function *(A) Agree*, there is a significant positive correlation between the preceding and following utterance ($r = 0.209$, $p < .001$). For the feedback function *(Y) Yes*, there is a strong positive correlation ($r = 0.53$, $p < .001$) between the preceding and following utterance. In the case of *(Y) Yes*, after the listener gives a positive response, the interlocutor may continue with the same pitch in a follow-up question, or continue speaking if the listener responded to a rhetorical question. For feedback function *(C) Continue*, although no evidence of alignment in terms of mean pitch was found, we do find evidence suggesting that it is unobtrusive to the conversation.

4.5.2. Mean Intensity

Figure 4.4 displays the significant results of Pearson's correlation for mean intensity between the feedback functions and the preceding and following 500 ms utterance of the interlocutor, as well as between the interlocutor's preceding and following utterance. The results for all Pearson's correlations that were not significant, along with the significant results in Figure 4.4 for mean intensity can be found in Section I.2 in Appendix I.

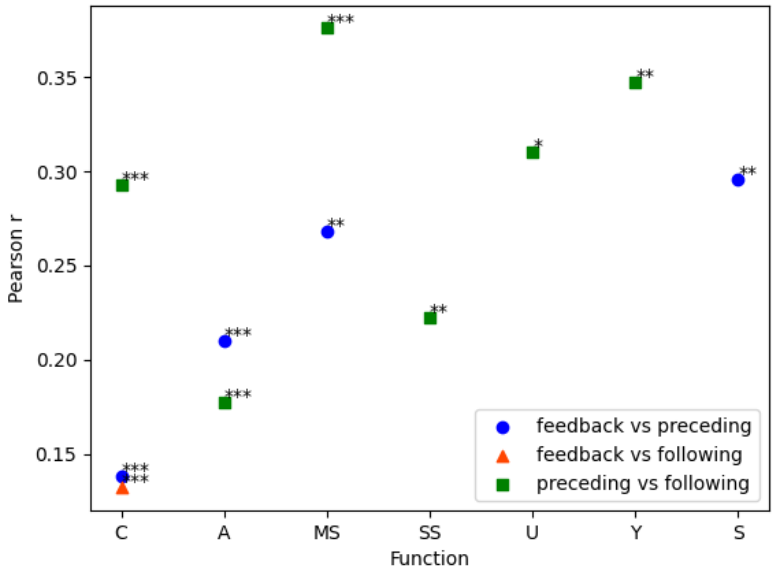


Figure 4.4. – Pearson's correlation coefficients between the mean intensity of feedback functions and preceding/following 500 ms of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

4.5.2.1. Feedback response vs. preceding utterance

We find significant evidence suggesting that the listener may align their intensity to that of the interlocutor's for feedback functions *(C) Continue* and *(A) Agree*. Although

there is a significant positive correlation between the mean intensity of the feedback function *(C) Continue* and the preceding utterance of the interlocutor, the correlation coefficient is fairly weak ($r = 0.138$, $p < .001$). Similarly, the correlation coefficient between the feedback function *(A) Agree* and the preceding utterance of the interlocutor is also significant but weak ($r = 0.21$, $p < .001$). For both functions, *(C) Continue* and *(A) Agree*, we can speculate that the listener does intend to take the floor and thus tends to 'tune-in' to the interlocutor's intensity.

Although R. Levitan, Beňuš, et al. (2015b) measured proximity, they observed that in terms of mean intensity, backchannels were more similar to the interlocutor's preceding utterance than smooth switches. Our results for the feedback function *(C) Continue*, are similar to the findings of Kawahara, Yamaguchi, Uesato, et al. (2015), Kawahara, Uesato, et al. (2015), and Ochi et al. (2024), who also measured synchrony in terms of power between Japanese backchannels and the preceding utterance of the interlocutor.

4.5.2.2. Feedback response vs. following utterance

We only find evidence suggesting that the interlocutor may align their intensity to the listener's feedback function *(C) Continue*, although the positive correlation is weak ($r = 0.132$, $p < .001$). In the case of feedback function *(C) Continue*, the alignment pattern observed is as follows: interlocutor A says something, listener B then produces a feedback function *(C) Continue* which aligns its intensity to that of interlocutor A, then interlocutor A continues speaking and aligns their intensity to the preceding feedback function *(C) Continue* of listener B.

In their comparison of mean intensity differences between backchannels and the interlocutor's preceding utterance, and between smooth switches, R. Levitan, Beňuš, et al. (2015b) found that the mean intensity was more similar between backchannels and the preceding utterance than it was in smooth switches. It is important to note that they measured proximity in their analysis.

4.5.2.3. Preceding utterance vs. following utterance

We find significant evidence suggesting that the intensity of the feedback functions *(C) Continue*, *(A) Agree*, and *(MS) Mild surprise* may be perceived as less obtrusive to the conversation. The correlation coefficient between the preceding and following utterance of the interlocutor for the function *(A) Agree* is positive and weak ($r = 0.177$, $p < .001$). For the correlation between the preceding and following utterance of the interlocutor for function *(C) Continue* is also weak, though it approaches moderate correlation ($r = 0.293$, $p < .001$). Meanwhile, the correlation between the preceding utterance and following utterance of the interlocutor for feedback function *(MS) Mild surprise* is positive and moderately correlated ($r = 0.376$, $p < .001$). These feedback functions are likely perceived as less obtrusive because the listener does not attempt to take the floor.

4.5.3. Pitch Slope

Unlike mean pitch and mean intensity, we did not find any significant or marginally significant results across all the feedback function categories. The results for all Pearson's correlations for pitch slope can be found in Section I.3 in Appendix I. The absence of significant results may be attributed to the method used to calculate pitch slope. For instance, Gravano and Hirschberg 2009 computed pitch slope by fitting least-squares linear regression models to the F_0 values; this way of computing pitch slope may have been better. In a study investigating alignment between the pitch slope of Swedish feedback responses and the preceding utterance of the interlocutor—where the pitch slope was calculated from the last syllables— Neiberg and Gustafson (2010) found that the there was evidence of alignment being low at the beginning of conversations, then increasing in the middle, and then decreasing towards the end. However, they concluded that detecting alignment in terms of pitch slope is challenging.

4.6. Discussion

Overall, we find more evidence of prosodic alignment in terms of mean intensity than in mean pitch. These results are similar to those observed in studies of feedback responses in Japanese (Kawahara, Yamaguchi, Uesato, et al. 2015; Kawahara, Uesato, et al. 2015; Ochi et al. 2024), where more prosodic alignment was observed in terms of power rather than pitch. Kawahara, Yamaguchi, Uesato, et al. (2015) and Kawahara, Uesato, et al. (2015) measured two types of alignment, proximity and synchrony, and both resulted in evidence of alignment in terms of power rather than pitch.

Specifically, for the feedback function *(C) Continue*, which is equivalent to what is referred as backchannels in previous studies, we do not observe prosodic alignment in terms of mean pitch. The differences in results compared to those found in previous U.S. English studies may be attributed to differences in the corpora used and the type of alignment measured. In this study, we measured synchrony, whereas previous U.S. English studies measured proximity. In order to make a direct comparison, future work should measure synchrony in the Columbia Games corpus and proximity in the Switchboard corpus.

Our results also indicate that some feedback functions are unobtrusive, but it does not necessarily imply prosodic alignment. For instance, although the feedback function *(Y) Yes* was found to be unobtrusive, there was no alignment with the preceding or following utterance of the interlocutor in terms of mean pitch. Unlike previous studies, which primarily focused on *(C) Continue*, we investigated prosodic alignment across additional feedback functions. However, we only found marginally significant results for *(U) Non-understanding*, *(S) Sympathy*, *(N) No*, *(MS) Mild Surprise*, and *(SS) Strong Surprise*. There were no significant or marginally significant results for *(D) Disagree* and *(DS) Disapproval*, suggesting that these two feedback functions may have been obtrusive to the conversation. Future studies should utilize the automatic annotations from our previous work (see Chapter 3) to determine if increasing the number of data points affects the significance of these results.

Unlike R. Levitan, Beňuš, et al. (2015b), we did not account for differences in latency due to limited data points for certain functions, which have been shown to negatively affect local alignment. Future studies should investigate the difference between the end time of the preceding utterance of the interlocutor and the start of the feedback, as well as the difference between the end time of the feedback and the start time of the following utterance for all feedback functions. Additionally, future research should also explore whether results vary with an increased window length for extracting prosodic features. This would help determine at which point local alignment can be most accurately measured.

In our analysis, we did not measure dynamic alignment, that is, taking into account the change of alignment across the whole conversation. It is possible that the way feedback functions align could vary from the beginning to the end of a conversation, as found in Neiberg and Gustafson (2010). However, it is important to note that the type of corpus can affect the results. As our findings suggest, there may be a difference between task-oriented speech and spontaneous conversational speech. Additionally, even within spontaneous conversational speech, social factors may influence alignment, such as whether the conversations occur between strangers, friends, or in a boss-employee context.

Lastly, in this study, we measured prosodic alignment using three features: mean pitch, mean intensity, and pitch slope. However, other acoustic features such as jitter, shimmer, and noise-harmonic ratio could also be explored. Similarly, other types of alignment, such as lexical alignment or phonetic alignment, warrant investigation. For example, do interlocutors use similar lexical forms or pronunciations as the conversation progresses?

4.7. Summary

In this chapter, we presented the results of our study investigating the prosodic alignment between feedback responses and the preceding and following utterance of the interlocutor, as well as between the preceding and following utterance of the interlocutor. We found evidence that interlocutors align in terms of intensity rather than pitch. This may because loudness may be more disruptive than pitch. We measured prosodic alignment across ten feedback functions: *(C) Continue*, *(U) Non-understanding*, *(A) Agree*, *(D) Disagree*, *(Y) Yes*, *(N) No*, *(S) Sympathy*, *(Ds) Disapproval*, *(MS) Mild Surprise* and *(SS) Strong Surprise*. However, our findings indicate that prosodic alignment does not occur across all these feedback functions. The lack of detected alignment may be due to our limited annotations.

Specifically, we find that listeners align their intensity for feedback functions *(C) Continue* and *(A) Agree* to the intensity of the preceding utterance of the interlocutor. For the feedback function *(C) Continue*, we observe a pattern of alignment in terms of intensity: when interlocutor A says something, listener B aligns the intensity of their feedback to that of interlocutor A, then interlocutor A continues speaking and aligns their intensity to the feedback of listener B. This finding can inform spoken dialogue

designers seeking to implement human-like feedback responses. Additionally, our results indicate that in terms of pitch and intensity, feedback functions *(C) Continue* and *(A) Agree* can be perceived as less obtrusive in conversation.

In the next chapter, we present our study evaluating the naturalness and appropriateness of the prosody of feedback responses synthesized using a Text-to-speech (TTS) system with the capability to control prosodic features. This TTS system could then be incorporated in a spoken dialogue system or conversational agent that can adapt the prosody of feedback responses to that of the user.

5. Prosodic Representation of feedback responses for synthesis

Summary

5.1. Introduction and Motivation 151
5.2. Synthesis methods 154
5.2.1. Signal processing approach 154
5.2.2. Text-to-speech approach 155
5.3. Experimental design 157
5.3.1. Participants 157
5.3.2. Stimuli 157
5.3.3. Participants' tasks 158
5.3.4. Statistical analysis 158
5.4. Results 158
5.4.1. Naturalness ratings 158
5.4.2. Appropriateness ratings 159
5.4.3. Perception of communicative functions 160
5.5. Discussion 161
5.6. Summary 162

5.1. Introduction and Motivation

This chapter is based on our publication *"Mhm... Yeah? Okay! Evaluating the Naturalness and Communicative Function of Synthesized Feedback Responses in Spoken Dialogue"* (Figueroa, Korte, et al. 2024), which was presented at the 25th Meeting of the Special Interest Group on Discourse and Dialogue (SigDial) 2024, in Kyoto, Japan.

The main goal of this work was to address our research question:

R5 *Can we synthesize feedback responses that sound natural and appropriate in context while conveying their intended communicative function?*

This study uses all components from the vocal feedback model. These two components are circled in Figure 5.1. Note that the prosodic features can be represented as discrete values such as word-level mean pitch, syllable-level mean pitch, phone-level

mean pitch, or they can be distributed representations such as prosody embeddings (Sigurgeirsson et al. 2023), or speech embeddings (Lin et al. 2023) or feedback embeddings (Qian et al. 2024).

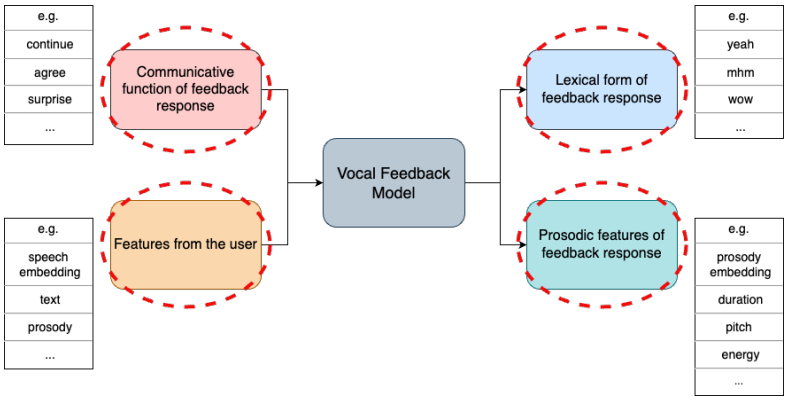


Figure 5.1. – Components from the vocal feedback model addressed by R5.

Incorporating feedback responses in spoken dialogue systems for conversational agents is an active research area (see Section 1.5 for a review). However, much of the existing work has concentrated on predicting the timing of feedback responses (Adiba, Homma, Bertero, et al. 2021; Ishii et al. 2021; Jinhan Wang et al. 2024). Recently, there has been growing interest in predicting the communicative functions of feedback responses (Boudin et al. 2024; Choi et al. 2024; Park, Liermann, Choi, Kim, et al. 2024). In contrast, fewer studies have focused on predicting the lexical forms (Fujie et al. 2005; Kawahara, Yamaguchi, Inoue, et al. 2016; Yamaguchi et al. 2016) or the prosodic features of feedback responses (Kawahara, Uesato, et al. 2015; Nath et al. 2022; Ochi et al. 2024).

Prosody, along with the lexical form, plays a crucial role in conveying the communicative function of feedback responses. In some instances, prosody may be the most significant factor; for example, the word “yeah” can express agreement, disagreement, or surprise depending on its prosodic realization (Figueroa, Adigwe, et al. 2022). Generating feedback responses with variable prosody is also vital as it prevents conversational agents from sounding monotonous. Moreover, manipulating prosody can help ensure that feedback responses are unobtrusive to the conversation (see Chapter 4). Additionally, modeling prosodic alignment requires controlling the prosody to enable the conversational agent to adapt its prosody to that of the user.

Previous studies have manipulated prosodic features using signal processing to explore their impact on the perceived communicative functions of synthesized feedback responses (Åsa Wallers et al. 2006; Stocksmeier et al. 2007; Chandler 2023). Unlike typical approaches, Pammi, Schröder, Charfuelan, et al. (2010) and Pammi and Schröder (2011) integrated feedback responses into a unit selection text-to-speech (TTS) synthesis system by treating entire feedback responses as units, rather than concatenating diphones or phones. However, the prosodic variability was obtained by performing

signal processing, which, as Pammi, Schröder, Charfuelan, et al. (2010) noted, can degrade audio quality.

Synthesizing feedback responses using TTS is relatively uncommon; we are aware of only a few studies. Oertel, Gustafson, et al. (2016) used statistical parametric speech synthesis to synthesize feedback responses. Fernandez et al. (2022) synthesized interjections within longer utterances, which included lexical forms shared with feedback responses. More recently, Mitsui et al. (2023) developed a TTS system that autonomously decides when and which feedback responses to produce, and how they should sound. Typically, feedback responses are not synthesized because TTS systems are generally trained on read speech, such as sentences from books, which excludes short feedback responses. Although there is growing interest in synthesizing conversational speech, feedback responses are often removed from the training data (Cooper 2019; Székely et al. 2019; O'Mahony, Lai, and King 2023; O'Mahony, Lai, and Székely 2024).

Aside from the perception studies that use signal processing to investigate how changes in prosody affect the perceived communicative functions of feedback responses, most studies that use TTS for generating feedback responses have not focused on whether the intended communicative function is conveyed or whether the prosody is contextually appropriate (if there is context). Evaluations of TTS systems typically focus on the naturalness of the synthesized speech (Kirkland et al. 2023). Furthermore, these evaluations are often conducted in isolation. However, there is emerging research that assesses synthesized speech in context (R. Clark et al. 2019; O'Mahony, Oplustil-Gallegos, et al. 2021; Lameris et al. 2023).

In this chapter, we explore the extent to which feedback responses can be synthesized using existing synthesis methods, so that they sound natural and prosodically appropriate given the context while conveying their intended communicative function. Our research question does not focus on predicting the prosodic features and lexical forms of feedback responses; rather, it investigates whether it is possible to synthesize them, given that we could make those predictions. For instance, as illustrated in Figure 5.2, we assume that the predicted lexical form and prosodic features of feedback responses are available and use these features as input for the TTS model to synthesize feedback responses.

Although Nath et al. (2022) did not conduct a subjective evaluation of their word-level prosodic predictions, they suggested that future work should perhaps predict the prosodic features of discourse markers at the frame-level. Perception studies that manipulate prosodic features of feedback responses typically adjust them at the word or syllable level. In this work, we explore the use of two levels of prosodic representations: the frame-level and the phone-level.

To investigate to what extent feedback responses can be synthesized that sound natural and prosodically appropriate, we re-synthesized feedback responses from the Switchboard corpus and transplanted the prosody of feedback responses by Switchboard "listeners" onto a voice talent's voice. We use two synthesis methods: a signal processing approach and a text-to-speech approach. In this chapter we report on the results of our listening tests, where participants evaluated the synthesized feedback re-

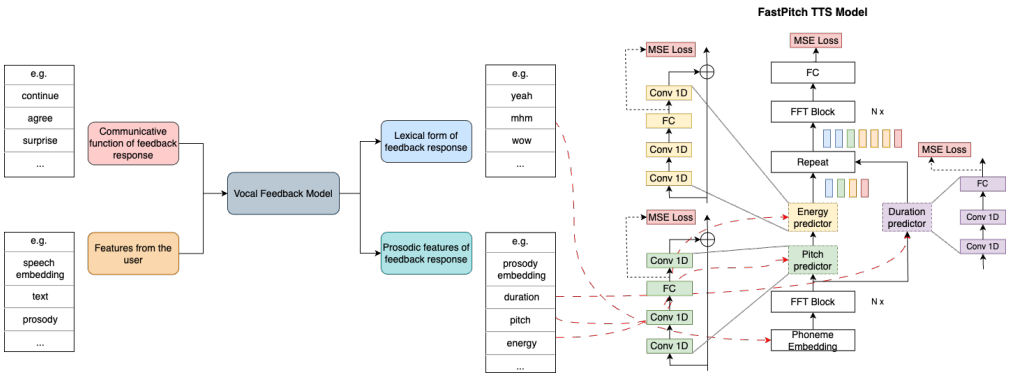


Figure 5.2. – Visualization of the lexical form and prosodic features of the feedback response serving as input to a text-to-speech system such as FastPitch.

sponses within their dialogue context. Participants rated the naturalness and prosodic appropriateness of these responses and identified their most likely communicative function. For comparison, they also evaluated the original feedback response and a re-synthesized monotone version created by flattening the pitch to remove intonation. To the best of our knowledge, this study is the first to evaluate the prosodic appropriateness of synthesized feedback responses in context.

5.2. Synthesis methods

To manipulate the prosodic features of feedback responses, we use two synthesis methods: a signal processing approach and a TTS approach. We transplant the prosody of feedback responses from "listeners" in the Switchboard corpus—referred to as our reference speakers—onto our target voice, a female voice talent.

5.2.1. Signal processing approach

One method of synthesis [1] involves using signal processing. The main advantage of this approach is the ability to manipulate a single audio recording (e.g., *yeah*) to create various renditions with different prosodic realizations. However, a significant disadvantage is that it can degrade audio quality and introduce artifacts. Using signal processing, the prosody of the original feedback response (as it appeared in the Switchboard conversation) is transplanted onto a *feedback template*, a single audio recording for each lexical form (e.g., *yeah, mhm*) of a feedback response recorded by the voice talent.

To transplant the prosody of the original feedback response onto the feedback template, we performed the following steps in this order:

1. Decomposing and recomposing the speech signal is often referred to as re-synthesis.

1. **Duration.** We first used the Montreal forced aligner (McAuliffe et al. 2017) to obtain the phone-level durations of the original feedback response, as well as the feedback template, and manually corrected them for alignment errors. We then used time-domain pitch-synchronous overlap-add (TD-PSOLA) to modify the phone durations of the feedback template to those of the original feedback response of the reference speaker.
2. **Pitch.** Second, we used the Python implementation (Dinh et al. 2019) of the WORLD vocoder (Morise et al. 2016) to extract the frame-level F_0 values from both the reference speaker's feedback response and the feedback template. We then z-score normalized the F_0 values of the reference speaker (Switchboard listener) per speaker, and then de-normalized these z-score values using the mean F_0 and standard deviation of the target speaker (voice talent). These de-normalized F_0 values were then used to replace those originally extracted from the target speaker, and the new F_0 values were used in the WORLD vocoder to resynthesize the audio of the target speaker.
3. **Intensity.** Finally, we transplanted the intensity contour of the reference speaker's feedback response onto the feedback template using the Praat Vocal Toolkit's (Corretge 2024) *Copy Intensity Contour* method.

Note that in this approach, duration is transplanted at the phone-level, while pitch and intensity are transplanted at the frame-level.

5.2.2. Text-to-speech approach

A second method of synthesis involves using a Text-to-speech (TTS) system. A TTS system takes as input text and outputs a speech waveform. Typically, a TTS system consists of two main components: an acoustic model that learns intermediate representations of speech, such as mel-spectrograms, and a vocoder that generates the waveform from these intermediate representations. The aim of a TTS system is to generate speech that has not been previously recorded by a voice talent, and this is the main advantage of using a TTS system. However, initial recordings are necessary to train a TTS system to synthesize the voice talent's voice. Another advantage of using a TTS system is that TTS systems are now able to produce speech that is close to human quality in terms of naturalness (Łańcucki 2021; K. Shen et al. 2023; Y. A. Li et al. 2024).

In this study, we use FastPitch 1.1 (Łańcucki 2021) for the acoustic model and HiFi-GAN Universal_V1 (Kong et al. 2020) for the vocoder model. While FastPitch is a deterministic model—meaning it generates the same prosodic realization for identical text input—it contains duration, pitch, and energy phone-level predictors that condition the acoustic features, enabling controllability of prosody. We specifically selected FastPitch to investigate whether phone-level prosodic representations could convey the intended communicative function of feedback responses. Note that FastPitch has previously been used to transplant prosody from one speaker to another (Van De Vreken et al. 2022).

To transplant the prosody of the reference speaker's original feedback response onto the synthesized feedback, we replaced the predicted prosodic features with the original ones during inference. These were the following replacements:

— **Duration.** We used the durations from the corrected phone-level alignments of the reference speaker's original feedback response to replace those predicted at inference from FastPitch's duration predictor (phone-level predictions) for the target speaker. Note the duration predictor predicts the number of frames per phone, therefore, the duration of the phones in seconds was converted to the number of frames per phone.

— **Pitch.** The F_0 values of the reference speaker's feedback response were extracted with Praat at the frame-level, averaged per phone, z-normalized and then de-normalized with the previously outlined procedure. These new phone-level pitch features were then used to replace the predicted pitch values from the pitch predictor during inference.

— **Energy.** We used FastPitch's energy extraction method to extract the energy values from the reference speaker's feedback response and then used these to replace those predicted from FastPitch's energy predictor at inference when synthesizing the feedback response of the target speaker.

In our initial attempt to determine the feasibility of synthesizing short feedback responses (i.e., single tokens such as *wow* rather than short utterances such as *oh wow*), we utilized a pre-trained TTS model—FastPitch and HiFi-Gan model—trained on the LJ-Speech Dataset (Ito et al. 2017). When synthesizing some short feedback responses from the Switchboard corpus, we observed that several sounded unintelligible. This likely occurred because the LJ-Speech Dataset contains longer utterances, and since the TTS system had not been exposed to isolated words during training, it struggled to synthesize them properly.

Since the pre-trained LJ TTS model could not synthesize feedback responses properly, we created a speech dataset using our voice talent for fine-tuning. Since we aimed to train a conversational voice and capture as much prosodic variation as possible, the voice talent recorded several types of speech for the TTS fine-tuning data:

— **Read speech.** 43 minutes were recorded from the CMU ARCTIC database (Kominek et al. 2004).

— **Role-Play acted speech.** 4 minutes were recorded from the Taskmaster-2 dataset (Byrne et al. 2019)

— **Conversational speech.** 34 minutes of speech were recorded from the voice talent during conversations with others via Zoom, capturing 981 feedback responses. These were dyadic conversations. In some sessions, the video was toggled between on and off: it was on for the first half and off for the second, or vice versa. In others, the video remained on throughout the entire conversation.

— **Feedback imitations.** The voice talent imitated 724 feedback responses from the Switchboard corpus, totaling 11 minutes of recording. This was necessary as the voice talent did not naturally use certain lexical forms (e.g., *golly*) during the

conversations[2].
All audio was recorded at 48 kHz and then downsampled to 22 050 Hz for training purposes.

The base acoustic model was trained on LJ-Speech for 500 epochs using phones as input, with batch size 16 and FastPitch's default learning rate scheduler. We fine-tuned this model on our target voice (the voice talent) for a further 500 epochs with the same hyperparameters as in pre-training, using a 97-3% train-validation split. To prevent the alignment loss from becoming unstable due to the unbalanced set between short and long utterances, we adjusted the target voice's data by tripling the longer utterances compared to the ratio of the feedback responses. We also fine-tuned a pre-trained HiFi-GAN Universal_V1 vocoder model on our target voice for 58,000 steps, using a batch size of 16, learning rate of $1e-5$, and the same train-validation split as for the acoustic model. All models were trained on a single NVIDIA T4 GPU.

5.3. Experimental design

5.3.1. Participants

We recruited and paid 86 native U.S. English speakers through Prolific (*Prolific* 2014): 48 females and 38 males within the age range of 24-73 years. All participants self-reported having no hearing impairments and self-reported wearing headphones or earphones during the study.

5.3.2. Stimuli

We hosted our online listening tests using Qualtrics (*Qualtrics* 2002). Participants listened to 12 distinct clips from the Switchboard conversations and were assigned to either set 1 or set 2; transcriptions of the clips can be found in Appendix J. Each clip featured a dialogue between one speaker and one listener. The listener's feedback responses could overlap with the speaker's talk or occur during the speaker's silence. Participants were presented with four conditions of the same set of Switchboard conversation clips, where the feedback responses were either:

— the original ones produced by the listener in the Switchboard corpus,
— the synthesized ones using signal processing, for the target speaker,
— the synthesized ones using TTS, for the target speaker,
— the flattened to a monotone pitch ones for the target speaker.

For the monotone feedback responses, the pitch was flattened using the Praat Vocal Toolkit's *Monotonize* method which flattens the pitch contour to the audio file's pitch median. This condition removes intonation information from the feedback response. Note that only the feedback responses were replaced; the Switchboard speaker's channel remained unchanged. All conditions were randomized, presented one by one, and the same conversation was never presented consecutively. Samples of the

2. Speakers may have their own preferred lexical forms when producing feedback responses

clips can be found at https://carolfigphd.github.io/SigDial2024_feedback_synthesis_samples/.

5.3.3. Participants' tasks

The participants had to complete three tasks for each feedback responses in the Switchboard clip:

— They were asked to assign a function from the 10 communicative functions of feedback responses from our previous work (Figueroa, Adigwe, et al. 2022): (C) Continue, (U) Non-Understanding, (A) Agree, (D) Disagree, (Y) Yes, (N) No, (S) Sympathy, (Ds) Disapproval, (MS) Mild Surprise and (SS) Strong Surprise.
— Participants were also asked to rate the naturalness of the feedback response on a Likert scale 1-5 where 1= Very Unnatural and 5= Very Natural.
— Participants were asked to rate the appropriateness of the prosody of the feedback response on a Likert scale 1-5 where 1= Very Inappropriate and 5= Very Appropriate.

Naturalness was defined as how human-like the feedback response was; participants were told beforehand that feedback responses were either machine-generated or human-generated. Appropriateness was defined as *"the way the listener says the feedback so that it conveys a meaning that makes sense in this context"*. Screenshots of the listening test interface can be found in Appendix K.

5.3.4. Statistical analysis

To analyze naturalness and appropriateness, we used a cumulative link mixed-model (CLMM) using the ordinal package v2023.12.4 (Christensen 2023) in R v4.3.2 (R Core Team 2023). The CLMM was fitted with the Laplace approximation, with a logit link and equidistant threshold. We fitted our data to a CLMM, where naturalness or appropriateness ratings were predicted by the synthesis method and we set the subject ID and stimuli ID (the feedback response ID) as random effects. The following formula was used for our condition model:

$$\text{clmm}(\textit{naturalness/appropriateness} \sim \textit{method} + (1|\textit{subjectID}) + (1|\textit{stimuliID}))$$

We used an ANOVA to compare our condition model to a null model:

$$\text{clmm}(\textit{naturalness/appropriateness} \sim (1|\textit{subjectID}) + (1|\textit{stimuliID}))$$

5.4. Results

5.4.1. Naturalness ratings

Figure 5.3 shows the distribution of the ratings for naturalness and mean μ and standard deviation σ for each condition: Monotone (Mon), text-to-speech (TTS), signal

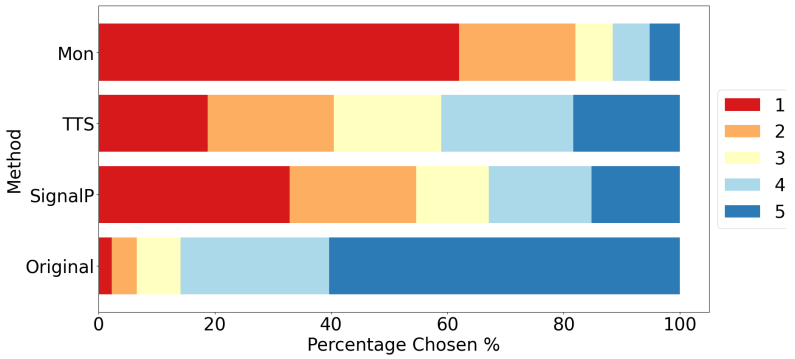


Figure 5.3. – Naturalness rating distribution per condition. Mon (μ=1.73, σ=1.15), TTS (μ=3.0, σ=1.39), SignalP (μ=2.61, σ=1.47), Original (μ=4.37, σ=0.96). 1= Very Unnatural, 5= Very Natural.

processing (SignalP), and the original Switchboard feedback response (Original). The results from our ANOVA comparison show that the synthesis method has significant impact on the model fit (AIC 20439, $p < .001$). We performed a post-hoc analysis pairwise comparisons using *emmeans* with a Bonferroni correction. Results showed that there were significant differences for all 6 pairwise comparisons ($p < .0001$): the feedback responses synthesized by the TTS were perceived as more natural than the feedback responses synthesized by signal processing and the monotone feedback responses. This was expected as signal processing degrades the audio, especially when there are larger differences in pitch. Also, as expected, the original Switchboard feedback responses were rated to be more natural than all conditions, yet not all feedback responses were rated as 5, despite having been produced by humans. Since naturalness was defined as human-likeness, we suspect participants also partially rated the audio quality.

5.4.2. Appropriateness ratings

Figure 5.4 shows the distribution of the rating for appropriateness and mean μ and standard deviation σ for each condition. The results from our ANOVA comparison show that the synthesis method has significant impact on the model fit (AIC 20215, $p < .001$). The post-hoc analysis showed that there were significant differences for almost all pairwise comparisons ($p < .0001$), except for the TTS and signal processing comparison, meaning both synthesis methods convey equally appropriate prosody for their context. Due to the prosodic information being removed in the monotone feedback responses, we observe that they are rated as more inappropriate than the other conditions. Despite asking separate questions for evaluating naturalness and prosody appropriateness, the relatively high score of the monotone appropriateness puts into question whether participants could disentangle naturalness and appropriateness.

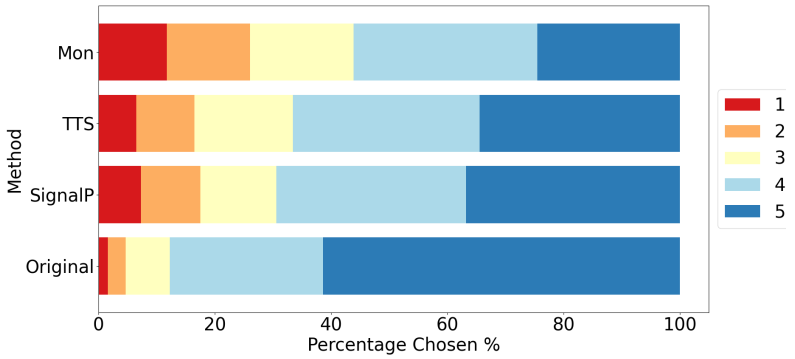


Figure 5.4. – Appropriateness rating distribution per condition. Mon (μ=3.43, σ=1.31), TTS (μ=3.78, σ=1.20), SignalP (μ=3.81, σ=1.24), Original (μ=4.43, σ=0.88). 1= Very Inappropriate and 5=Very Appropriate.

5.4.3. Perception of communicative functions

To evaluate whether the synthesis methods preserve the communicative function of the original Switchboard feedback response, we calculated Cohen's kappa coefficient scores of the participants' annotations of communicate function between their estimate of the original feedback vs. their estimate of the re-synthesized counterpart (see Table 5.1). Confusion matrices for each comparison can be found in Appendix L.

Comparison	**Cohen's kappa**
Original vs SignalP	0.79
Original vs TTS	0.71
Original vs Mon	0.74

Table 5.1. – Cohen's kappa coefficient scores per comparison of intra-annotator agreement.

The results show that the perceived communicative function is best preserved by the signal processing approach, which is expected since it transplants prosodic features pitch and intensity at the frame level. Although signal processing and TTS feedback responses convey equally appropriate prosody for their context, the feedback responses synthesized by the TTS approach were not as effective in preserving the communicative function, especially those containing attitudinal information, such as *(S) Sympathy*, *(MS) Mild Surprise* and *(SS) Strong Surprise*. For example, if the original communicative function of the Switchboard feedback response was *(S) Strong Surprise*, but the participants perceived the TTS feedback response as *(MS) Mild Surprise*, both functions and prosodic realizations were appropriate for the context but conveyed different communicative functions. In fact, the kappa for the TTS was even lower than the monotone condition, where no intonational (and thus very little prosodic)

information was preserved. Thus, the participants likely mainly relied on the lexical form and/or context in those conditions.

5.5. Discussion

As expected, feedback responses synthesized using TTS were rated as more natural than those synthesized using signal processing, as signal processing degrades the audio Pammi, Schröder, Charfuelan, et al. (2010). However, while both synthesis methods conveyed appropriate prosody for the given context, the signal processing method better preserved the intended communicative function than the text-to-speech method. This is likely because signal processing transplants pitch and intensity features at the frame level, compared to the phone level in TTS, allowing for more accurate preservation of the communicative function. Our results indicate that by manipulating the prosodic features of a TTS system, we can convey different communicative functions, though not always the intended one. These findings support Nath et al. (2022)'s recommendation that future work should attempt to predict the prosodic features of feedback responses at the frame-level.

Similar to previous perception studies (Åsa Wallers et al. 2006; Stocksmeier et al. 2007; Chandler 2023; Pammi, Schröder, Charfuelan, et al. 2010; Pammi and Schröder 2011) that manipulated prosodic features at the word or syllable level of feedback responses to investigate their effects on the perceived communicative function, our TTS system could be used to explore the significance of manipulating individual prosodic features such as energy, duration, and pitch in feedback responses.

One limitation of this study is that only 47 feedback responses were evaluated, which did not cover all possible lexical forms found in the Switchboard corpus. While we ensured representation of all ten communicative functions [3] identified in the Switchboard corpus (Figueroa, Adigwe, et al. 2022), only 23 of the 59 previously identified lexical forms were synthesized for the listening test. These included: *absolutely*, *ah*, *aww*, *exactly*, *hm*, *jeez*, *mhm*, *mm*, *no*, *oh*, *okay*, *oof*, *ooh*, *pardon*, *really*, *right*, *ugh*, *uh-huh*, *what*, *wow*, *yeah*, *yep*, *yes*.

A second limitation is the within-participant experimental design, meaning that participants were presented with the same clips for all conditions. Some participants provided feedback that this made the task somewhat repetitive. However, we chose this experimental design because we were interested in each participant's individual perception of the communicative functions, which can vary from person to person.

For example, if participant A perceived the communicative function of a feedback response in the original Switchboard clip as *(C) Continue*, while participant B perceived it as *(A) Agree*, and then both participants perceived the signal-processing feedback response as *(C) Continue*, it becomes difficult to determine whose perception of the original feedback response is "correct". Instead, our focus was on whether their perception changed or stayed the same. The within-participant design allowed us to treat

3. The communicative functions were perceived and annotated by us in Figueroa, Adigwe, et al. (2022), but participants did not always perceive them in the same way.

the original feedback responses from the Switchboard conversations as true labels for each participant.

Furthermore, our results suggest that participants may not have been able to disentangle naturalness of the audio and appropriateness of the prosody. This underscores the importance of how questions are framed. In our pre-pilot study, we did not define appropriateness as "the way the listener says the feedback so that it conveys a meaning that makes sense in this context", leading to varied interpretations of the term among participants. Despite our efforts to ask about the appropriateness of prosody without directly using the phrase "is the prosody appropriate" [4], our definition may still have been unclear. A better approach could have been to explain prosody more explicitly in the question, perhaps by asking about intonation, rhythm, and tone, which may have provided a more accurate assessment of prosodic appropriateness.

5.6. Summary

In this chapter, we presented the results from our listening tests, in which participants rated the appropriateness of the prosody, the naturalness of the synthesized feedback responses, and identified the communicative function while listening to them in context. Our findings showed that participants rated feedback responses synthesized by the TTS system as more natural than those produced by the signal processing approach. However, the TTS method struggled to convey the intended communicative function beyond the lexical form, while the signal processing method provided additional prosodic cues, likely due to its more fine-grained representation of prosody.

The implications of these finding are that, if we were to build a model that predicts the prosodic features of feedback responses, it may be beneficial to predict these features at the frame-level, as the frame-level signal processing approach better preserves the intended communicative function. For instance, such a prediction model could extend the work of Corkey et al. (2023), where an external predictor was trained to predict intonation.

4. We avoided this phrasing as participants might not be familiar with the term prosody.

Conclusion

Voice-based conversational agents such as Apple's Siri, Amazon's Alexa, and ChatGPT-4o are becoming ubiquitous in our daily lives. However, they often lack a crucial conversational characteristic integral to human-human communication: the use of short feedback responses such as "mhm", "aww", "wow". This thesis revolves around the central question, "What is needed for synthesizing feedback responses with appropriate prosody for its context?". Specifically, it addresses "how" these short feedback responses should sound—their prosodic realization and representation. In this thesis, we proposed components for a vocal feedback model that predicts both the lexical form and the prosodic features of feedback responses, based on the communicative function the conversational agent should convey and the context of the interlocutor. We argue that the communicative function, the lexical form, and the prosody of the feedback responses, along with contextual information from the user, are crucial aspects that need to be modeled to enable conversational agents to generate human-like feedback responses. These aspects motivate our research questions, which we summarize in the following section.

Summary and contributions

In this section, we review the research questions posed in our introduction and discuss the contributions of this thesis.

R1 *What are the various communicative functions of feedback responses in human-human conversations that a conversational agent may be required to convey?*

In Chapter 2, we introduced our coding scheme for annotating the communicative functions of feedback responses. This annotation scheme was developed through manual annotation and analysis of 1,627 feedback responses in the Switchboard corpus. We identified ten communicative functions in the Switchboard corpus: *(C) Continue*, *(U) Non-understanding*, *(A) Agree*, *(D) Disagree*, *(Y) Yes*, *(N) No*, *(S) Sympathy*, *(Ds) Disapproval*, *(MS) Mild Surprise*, and *(SS) Strong Surprise*. Unlike previous coding schemes that often categorized feedback responses containing attitudinal information into a single category, our scheme provides a more nuanced, fine-grained classification. We conducted our annotations from a perspective of generation (i.e., synthesis). We believe this granularity is essential for conversational agents to convey human-like feedback responses.

R2 *Per each communicative function found in human-human conversations what are the different distributions in terms of lexical forms and prosody?*

In Chapter 2, we also presented our analysis of lexical forms and prosodic features. From the 1,627 manual annotations of feedback responses, we identified 59 unique lexical forms, which can often be ambiguous as they may serve various communicative functions. These forms and their distribution across communicative functions are listed in Table 2.7. Identifying these lexical forms is crucial, as it enables a conversational agent to exhibit lexical variability and avoid sounding monotonous. Additionally, our prosodic analysis highlighted significant differences in duration, mean pitch, pitch slope, and pitch range among the communicative functions. These differences suggest that prosodic features could potentially distinguish between communicative functions.

R3 *Is it possible to automatically classify the communicative functions of feedback responses based on their lexical form, prosodic features, and preceding context?*

In Chapter 3, we presented the results of our experiments that explored combinations of lexical, prosodic, and contextual features for classifying ten communicative functions of feedback responses. Previous classification studies often categorized feedback responses with attitudinal information into a single category, or focused on classifying functions of a single lexical form. We used a Support Vector Machine (SVM) due to its ability to handle smaller datasets and experimented with using GPT-3 both as a classifier and as a means to encode lexical and contextual information for input into the SVM. Our findings indicate that GPT-3, when fine-tuned for this classification task, performed best but was not cost-effective. Our second-best model, at zero cost, also performed on par with human annotations using only lexical embeddings and SimCSE embeddings for contextual representation. Our results suggest that the prosodic features—duration, mean pitch, pitch slope, pitch range, and mean intensity—were either redundant given the contextual embedding or were inadequate representations of prosody. Using this second-best model to classify the remaining potential feedback responses in the Switchboard corpus, we obtained an additional 74,106 new feedback response annotations. We believe this classifier could be effectively applied to other U.S. English corpora.

R4 *Does prosodic alignment occur for different communicative functions of feedback responses?*

In Chapter 4, we presented the results of our study investigating prosodic alignment between feedback responses and the 500 ms preceding and following utterance of the interlocutor. Prosodic alignment refers to the phenomenon where interlocutors become similar in their prosodic features. Specifically, we measured prosodic alignment for the following prosodic features: mean pitch, mean intensity, and pitch slope. We used Pearson's correlation to measure a specific type of alignment, synchrony, which

measures the similarity in movement. This type of alignment had not been previously measured in other U.S. English studies and for a spontaneous conversational corpus. While previous studies (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Beňuš, et al. 2015b) have suggested that feedback responses are unobtrusive due to prosodic alignment, we found evidence of feedback responses being unobtrusive even without clear signs of alignment, possibly due to the limited data points from our manual annotations for certain communicative functions. Notably, we observed that listeners align their mean intensity to the preceding utterance of their interlocutor, particularly for feedback functions *(C) Continue* and *(A) Agree*. Specifically, for function *(C) Continue*, we observed alignment in both directions, in other words, the listener aligns their mean intensity to their interlocutor's preceding utterance, and the interlocutor aligns their mean intensity to the preceding listener's feedback response. This study provided insights into a human-human interaction phenomenon that should be modeled in generating feedback responses for conversational agents.

R5 *Can we synthesize feedback responses that sound natural and appropriate in context while conveying their intended communicative function?*

In Chapter 5, we presented results from our listening test, where participants assessed the appropriateness of prosody, the naturalness of audio, and identified the communicative function they perceived from synthesized short feedback responses within a given dialogue context. We used two synthesis methods: a signal processing approach and a text-to-speech (TTS) approach. To evaluate whether these methods could convey the intended communicative functions, we transplanted the prosody from listener feedback responses in the Switchboard corpus onto a voice talent's prosody. In our previous studies we used various word-level prosodic features as detailed in Table 5.1, this study, however, used frame-level and phone-level prosodic features. Both methods involved transplanting duration at the phone level; however, for signal processing, pitch and intensity were transplanted at the frame level, whereas for TTS, mean pitch and mean energy were transplanted at the phone level. Our findings indicated that while TTS produced the most natural-sounding feedback responses, both synthesis methods were deemed equally appropriate in terms of generating appropriate prosody. Nevertheless, the TTS approach fell short in conveying the intended communicative function beyond being able to synthesize the lexical form properly. In contrast, the signal processing method provided additional prosodic information, likely due to its more detailed prosodic representation. These results suggest that future research aiming to develop a vocal feedback model that predicts the prosodic features of feedback responses should consider predicting more granular, frame-level prosodic features.

To the best of our knowledge, this is the first study to evaluate the appropriateness of the prosody of synthesized feedback responses. We also demonstrated the feasibility of synthesizing single-token feedback responses. Modeling the prosody of these short feedback responses is crucial because it, along with the lexical form, conveys the communicative function of the feedback responses. Additionally, being able

to generate feedback responses with variable prosody would help a conversational agent avoid sounding monotonous. Controllability of the prosodic features is also crucial to modeling prosodic alignment, as this would enable a conversational agent to adapt its prosodic features to those of the user. Lastly, our dataset containing conversational speech and feedback responses is available upon request for those interested in modeling feedback responses [1].

Chapter	**Level**	**Prosodic features**
Chapter 2	word	duration, mean pitch, pitch range, pitch slope
Chapter 3	word	duration, mean pitch, pitch range, pitch slope, mean intensity
Chapter 4	word	mean pitch, pitch slope, mean intensity
Chapter 5	frame,	pitch, intensity
	phone	duration, mean pitch, mean energy

Table 5.1. – List of prosodic features and levels of representation used throughout this thesis.

Limitations

The research questions in this thesis have been deliberately scoped in several ways, leaving opportunities for further expansion in future work.

Language dependency

Feedback responses have been known to be language dependent (Tottie 1991; Ward and Tsukahara 2000). Therefore, our findings are specific to U.S. English and may not generalize to other variants of English and other spoken languages. This includes the list of lexical forms, their distributions across different communicative functions, and the results concerning their prosody.

Corpus dependency

Although we believe that our annotation scheme for communicative functions of feedback responses can be generalized to other languages, we do acknowledge that it is corpus dependent. For instance, if a conversational agent is designed to serve as a companion for the elderly, a corpus containing vocal feedback from conversations between caregivers and patients or between family and friends may be more suitable for creating more intimate interactions. The communicative functions in our scheme are limited to those found in the Switchboard corpus, however, we believe them to be representative for social talk.

1. https://carolfigphd.github.io/FiCa-speech-dataset/

Non-vocal feedback responses

In face-to-face conversations, vocal and visual feedback responses (such as facial expressions and head nods) can occur together or separately. This thesis focuses solely on modeling vocal feedback responses. Future work should, for example, compare vocal feedback in the Switchboard corpus versus those in multimodal corpora, such as the Candor corpus (Reece et al. 2023), to determine if there are differences. Generating vocal feedback responses is just one part of the puzzle in modeling feedback responses. Investigating what is needed for synthesizing feedback responses with appropriate prosody for their context is a necessary first step before modeling both visual and vocal feedback responses.

Limited to dyadic conversations

The Switchboard corpus only contains telephone conversations between two people. We do not investigate vocal feedback in multi-party conversations and whether there are differences. The only work we are aware of that models vocal feedback by third party listeners is the work by Lala, Inoue, et al. (2022).

Lack of timing and communicative function predictions

The proposed vocal feedback model does not predict the timing or the communicative function; it assumes these are provided. We assume that the designer has the communicative function in mind and the intention of the agent, which is then given to the model to convey. It is important to emphasize that the communicative function cannot be determined without knowing the agent's intention. For example, if the user asks their companion conversational agent, "Sally's new boyfriend was really nice, wasn't he?", both "uh-huh" indicating agreement and "no" indicating disagreement are plausible responses. The appropriate response depends on the context and the agent's intention, which may be determined by an external model. We also assume that an external model predicts when the response should be generated.

Future work

In this section we suggest directions for future work in modeling vocal feedback responses.

Further use of our Switchboard annotations

Our total count of 76,285 annotations of feedback functions, obtained through both manual and automatic annotations, provides a substantial dataset for more detailed prosodic analysis. For instance, with ample data points available, one could investigate the prosodic differences within a single lexical form. Specifically, we can analyze

whether the feedback response *yeah*, which can convey various communicative functions such as *(C) Continue*, *(A) Agree*, *(D) Disagree*, and *(MS) Mild Surprise*, exhibits distinct prosodic characteristics across these different functions.

The manual and automatic annotations could also be utilized to train a classifier with more balanced data, which could then be used to annotate communicative functions of feedback responses in another U.S. English corpus. For example, this could be applied to the Candor corpus, which consists of video calls. By comparing annotations from the Switchboard corpus and the Candor corpus, insights could be gained into whether there are differences in the frequency of vocal feedback responses between video and telephone conversations. Additionally, this comparison could reveal whether the communicative functions used in video and telephone conversations differ. Comparisons could also be made between feedback responses in task-oriented and in spontaneous conversational speech.

Previous studies that have predicted the communicative functions of feedback responses, such as those using the Switchboard corpus (Ortega, C.-Y. Li, et al. 2020; Choi et al. 2024; Park, Liermann, Choi, and Lee 2024), have used coarse-grained categories. These models predict only two broad communicative functions, *continuers* (e.g., *mhm*, *uh-huh*, *yeah*) and *assessments* (e.g., *aww*, *oh*, *wow*). All feedback responses containing attitudinal information are grouped into the single category of assessments. Our annotations could be used to train a model that predicts more fine-grained communicative functions, thus distinguishing between different types of attitudinal information. Such fine-grained predictions would enable conversational agents to generate more human-like feedback responses.

Our study investigating prosodic alignment between feedback responses and the preceding and following utterance of the interlocutor was limited by the small number of manual annotations. We recommend that future research utilize our automatic annotations to determine if the marginally significant results we observed become statistically significant with an increased number of data samples. The results from this new study should then be taken into account when modeling prosodic alignment in conversational agents.

Further exploration of alignment and feedback responses

In our prosodic alignment study, we only measured one type of alignment, *synchrony*, between the feedback responses and the preceding and following utterance of the interlocutor. We suggest that future work measure other types of alignment, such as *proximity*. Previous studies in U.S. English have only measured proximity between feedback responses and the preceding utterance of the interlocutor in the task-oriented Columbia Games Corpus (Heldner, Edlund, and Hirschberg 2010; R. Levitan, Gravano, et al. 2011; R. Levitan, Beňuš, et al. 2015b). For a fair comparison between the alignment types, *proximity* between the types of alignment, it would be beneficial to measure proximity in the Switchboard corpus and synchrony in the Columbia Games corpus. The outcomes of this comparison could inform which types of alignment should be modeled in conversational agents.

Other factors should be considered when measuring prosodic alignment between feedback responses and the preceding and following utterances of the interlocutor. For example, latency has been shown to negatively affect local alignment (R. Levitan, Beňuš, et al. 2015b). Thus, future studies should account for the durations between the end of the interlocutor's preceding utterance and the start of the feedback response, as well as the duration between the end of the feedback response and the start of the following utterance. Additionally, the window length used to extract prosodic features from the interlocutor's utterance should also be considered. Future studies might explore using longer time windows than the 500 ms window length we employed, potentially increasing it to better detect points where prosodic alignment can be detected.

Due to our method of automatically extracting clips with consecutive feedback responses in the Switchboard corpus, instead of annotating entire conversations from start to finish, we were unable to investigate dynamic alignment. Specifically, by removing the temporal context of the whole conversation, we could not measure changes in prosodic alignment as the conversation progressed. A comprehensive study could have revealed whether there is less prosodic alignment at the beginning of conversations and whether it increases as the conversation progresses. Additionally, other types of alignment, such as lexical alignment, could also be detected by considering the progression of the entire conversation. These observations could then be modeled in conversational agents.

Exploration of other prosodic representations

In this thesis, we have focused solely on discrete representations of prosody. However, the prosodic features of feedback responses are not limited to discrete values; they can also be represented by distributed representations such as prosody embeddings. Nevertheless, as noted by (Sigurgeirsson et al. 2023), prosody embeddings generally struggle to disentangle prosody information from speaker characteristics. Kakouros et al. (2023) probed BERT embeddings to understand what these embeddings learned about prosody. Similar investigations can be applied to speech embeddings like HuBERT (Hsu et al. 2021), which can be used to encode prosody and then condition the TTS to manipulate prosody (Hu et al. 2022). Additionally, there has been research on conditioning TTS with contextual information (Oplustil-Gallegos et al. 2021). We believe that both context and prosody embeddings could be used to condition the TTS, enabling the generation of prosodically varied feedback responses that are appropriate to the context. In fact, Qian et al. (2024) has developed feedback embeddings that can be clustered into different communicative function categories, which could be used to condition TTS and generate feedback responses that express specific communicative functions. Future research on predicting prosodic features of feedback responses should explore the use of distributed representations for more nuanced modeling.

Taking into account vocal and visual feedback responses

In this thesis, we have focused solely on vocal feedback responses due to our selection of the Switchboard corpus, which consists of telephone conversations. If the Candor corpus, which comprises video conversations, were annotated using our annotation scheme, these annotations could be used to identify the types of visual feedback responses that co-occur with vocal feedback responses. This analysis would provide insights into the visual feedback responses associated with vocal feedback that contains attitudinal information. However, researchers should be cautious about assuming that these modalities are interchangeable.

Most research on modeling vocal and visual feedback responses assumes that these modalities are interchangeable. Currently, there are studies predicting the timing of vocal and visual feedback responses jointly (Iwan de Kok et al. 2012b; Inden et al. 2013; Penteado et al. 2019; Blache et al. 2020). However, as noted by R. Poppe, Truong, and Heylen (2011), predicting the timing of vocal and visual feedback responses should be separate, as they may not be interchangeable. For example, while replacing an "uh-huh" with a nod might be perceived as appropriate, replacing a nod with an "uh-huh" may not be acceptable. Future work on predicting the timing of feedback responses should follow the direction of Jain et al. (2021), whose model predicts whether to generate vocal feedback, visual feedback, or both. We believe this distinction in modality will become more pronounced once conversational agents are able to generate feedback responses with attitudinal information; for instance, it would be inappropriate to replace a nod with an "ugh" to signal disapproval.

Bibliography

[Ach+23] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023) (cit. on p. 29).

[Adi+21] Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, et al. “Delay mitigation for backchannel prediction in spoken dialog system”. In: *Conversational Dialogue Systems for the Next Decade* (2021), pp. 129–143 (cit. on pp. 30, 32, 86, 152, 199, 202).

[AHM21] Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori 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 (cit. on pp. 90, 200, 203, 205).

[Al +12] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, et al. “Furhat: a back-projected human-like robot head for multiparty human-machine interaction”. In: *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. Springer. 2012, pp. 114–130 (cit. on p. 28).

[All+11] Melissa L Allen, Sarah Haywood, Gnanathusharan Rajendran, et al. “Evidence for syntactic alignment in children with autism”. In: *Developmental Science* 14.3 (2011), pp. 540–548 (cit. on p. 69).

[All+07] Jens Allwood, Loredana Cerrato, Kristiina Jokinen, et al. “The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena”. In: *Language Resources and Evaluation* 41.3 (2007), pp. 273–287 (cit. on pp. 45, 48, 107).

[ANA92] Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. “On the semantics and pragmatics of linguistic feedback”. In: *Journal of semantics* 9.1 (1992), pp. 1–26 (cit. on pp. 26, 42, 45–50, 52, 104–106).

[And+91] Anne H Anderson, Miles Bader, Ellen Gurman Bard, et al. “The HCRC map task corpus”. In: *Language and speech* 34.4 (1991), pp. 351–366 (cit. on p. 31).

[ABS22] Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. “Modeling feedback in interaction with conversational agents—a review”. In: *Frontiers in Computer Science* 4 (2022) (cit. on pp. 30, 84).

[Bar+13] Ellen Gurman Bard, Corine Astésano, Mariapaola D'Imperio, et al. "Aix MapTask: A new French resource for prosodic and discourse studies". In: *TRASP 2013* (2013), pp. 15–19 (cit. on p. 67).

[BCJ00] Janet B Bavelas, Linda Coates, and Trudy Johnson. "Listeners as co-narrators." In: *Journal of personality and social psychology* 79.6 (2000), p. 941 (cit. on pp. 42–46, 52, 108, 196–198).

[Beň14] Štefan Beňuš. "Social aspects of entrainment in spoken interaction". In: *Cognitive Computation* 6 (2014), pp. 802–813 (cit. on p. 140).

[Beň16] Štefan Beňuš. "The prosody of backchannels in Slovak". In: *Proc. Speech Prosody 2016*. 2016, pp. 415–419 (cit. on pp. 45, 50, 51, 55, 76, 80, 136, 139, 141, 145).

[BGH07] Štefan Beňuš, Agustín Gravano, and Julia Bell Hirschberg. "The prosody of backchannels in American English". In: 2007 (cit. on pp. 31, 45, 55, 102, 111, 114–116, 196–198).

[Ber+08] Roxane Bertrand, Philippe Blache, Robert Espesser, et al. "Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle". In: *Revue TAL: traitement automatique des langues* 49.3 (2008), pp–105 (cit. on p. 67).

[Ber+07] Roxane Bertrand, Gaëlle Ferré, Philippe Blache, et al. "Backchannels revisited from a multimodal perspective". In: *Auditory-visual Speech Processing*. 2007, pp. 1–5 (cit. on p. 83).

[Bev+10] Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, et al. "Multimodal backchannels for embodied conversational agents". In: *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer. 2010, pp. 194–200 (cit. on pp. 30, 84).

[Bla+20] Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, et al. "An integrated model for predicting backchannel feedbacks". In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 2020, pp. 1–3 (cit. on pp. 84, 170).

[BW24] Paul Boersma and David Weenink. *Praat: doing phonetics by computer*. `http://www.praat.org/`. 2024 (cit. on p. 97).

[Bou+21] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, et al. "A multimodal model for predicting conversational feedbacks". In: *International Conference on Text, Speech, and Dialogue*. Springer. 2021, pp. 537–549 (cit. on pp. 45, 52, 53, 90, 93, 200, 203, 205).

[Bou+24] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, et al. "A multimodal model for predicting feedback position and type during conversation". In: *Speech Communication* 159 (2024), p. 103066 (cit. on pp. 30, 52, 92, 93, 152, 200, 203, 207).

[BPC00] Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. "Syntactic co-ordination in dialogue". In: *Cognition* 75.2 (2000), B13–B25 (cit. on p. 69).

[Bro+20] Tom Brown, Benjamin Mann, Nick Ryder, et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901 (cit. on pp. 29, 124).

[Bub88] Wolfram Bublitz. "Supportive fellow-speakers and cooperative conversations". In: (1988) (cit. on pp. 42, 46).

[Bun99] Harry Bunt. "Dynamic interpretation and dialogue theory". In: *The structure of multimodal dialogue* 2 (1999), pp. 139–166 (cit. on p. 50).

[Bun09] Harry Bunt. "The DIT++ taxonomy for functional dialogue markup". In: *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*. 2009, pp. 13–24 (cit. on pp. 45, 50).

[Bus+11] Hendrik Buschmeier, Zofia Malisz, Marcin Włodarczak, et al. "Are You Sure You're Paying Attention?'-Uh-Huh'Communicating Understanding as a Marker of Attentiveness". In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011 (cit. on pp. 45, 49, 107).

[Byr+19] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, et al. "Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset". In: *Proc. EMNLP-IJCNLP*. 2019, pp. 4516–4525 (cit. on p. 156).

[Cam07] Nick Campbell. "Towards conversational speech synthesis; lessons learned from the expressive speech processing project." In: *SSW* 2207 (2007), pp. 22–27 (cit. on pp. 37, 94).

[Cas+99] Justine Cassell, Timothy Bickmore, Mark Billinghurst, et al. "Embodiment in conversational interfaces: Rea". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 1999, pp. 520–527 (cit. on p. 84).

[CCK03] Nicola Cathcart, Jean Carletta, and Ewan Klein. "A shallow model of backchannel continuers in spoken dialogue". In: *European ACL*. Citeseer. 2003, pp. 51–58 (cit. on pp. 86, 199, 202).

[Cha23] Ryan Lee Chandler. "Semantic-prosodic correlates of backchannel utterances in human-computer dialog". PhD thesis. University of Illinois at Urbana-Champaign, 2023 (cit. on pp. 37, 60, 61, 97, 152, 161).

[Cha+24] Yupeng Chang, Xu Wang, Jindong Wang, et al. "A survey on evaluation of large language models". In: *ACM Transactions on Intelligent Systems and Technology* 15.3 (2024), pp. 1–45 (cit. on p. 29).

[Cho+22] Eugene Cho, Nasim Motalebi, S Shyam Sundar, et al. "Alexa as an active listener: how backchanneling can elicit self-disclosure and promote user experience". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–23 (cit. on p. 96).

[CBK24] Yong-Seok Choi, Jeong-Uk Bang, and Seung Hi Kim. "Joint streaming model for backchannel prediction and automatic speech recognition". In: *ETRI Journal* (2024) (cit. on pp. 30, 32, 92, 152, 168, 200, 203, 207).

[Chr23] Rune H. B. Christensen. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4. 2023. URL: https://CRAN.R-project.org/package=ordinal (cit. on p. 158).

[CMW04] Christopher Cieri, David Miller, and Kevin Walker. "The Fisher corpus: A resource for the next generations of speech-to-text." In: *LREC*. Vol. 4. 2004, pp. 69–71 (cit. on p. 32).

[Cla+96] Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, et al. "The conversational use of reactive tokens in English, Japanese, and Mandarin". In: *Journal of pragmatics* 26.3 (1996), pp. 355–387 (cit. on pp. 42–44, 196–198).

[Cla94] Herbert H Clark. "Managing problems in speaking". In: *Speech communication* 15.3-4 (1994), pp. 243–250 (cit. on pp. 45, 47, 50, 106).

[Cla96] Herbert H Clark. *Using language*. Cambridge university press, 1996 (cit. on pp. 24, 25, 45, 105, 106).

[Cla+19a] Leigh Clark, Nadia Pantidi, Orla Cooney, et al. "What makes a good conversation? Challenges in designing truly conversational agents". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–12 (cit. on pp. 27–29).

[Cla+19b] Rob Clark, Hanna Silen, Tom Kenter, et al. "Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs". In: *10th ISCA Workshop on Speech Synthesis (SSW 10)*. 2019, pp. 99–104 (cit. on p. 153).

[Col17] Evan Hugh Coles-Harris. "Perspectives on the motivations for phonetic convergence". In: *Language and Linguistics Compass* 11.12 (2017), e12268 (cit. on p. 70).

[Coo19] Erica Cooper. *Text-to-speech synthesis using found data for low-resource languages*. Columbia University, 2019 (cit. on p. 153).

[COK23] Niamh Corkey, Johannah O'Mahony, and Simon King. "Intonation Control for Neural Text-to-Speech Synthesis with Polynomial Models of F0". In: *Proc. Interspeech 2023*. 2023, pp. 2014–2015 (cit. on p. 162).

[Cor24] Ramon Corretge. *Praat Vocal Toolkit*. https://www.praatvocaltoolkit.com. 2024 (cit. on p. 155).

[Cut10] Pino Cutrone. “The backchannel norms of native English speakers: A target for Japanese L2 English learners”. In: *language Studies Working papers* 2 (2010), pp. 28–37 (cit. on p. 31).

[DR08] Alfred Dielmann and Steve Renals. “Recognition of dialogue acts in multiparty meetings using a switching DBN”. In: *IEEE transactions on audio, speech, and language processing* 16.7 (2008), pp. 1303–1314 (cit. on p. 120).

[DKT19] Tuan Dinh, Alexander Kain, and Kris Tjaden. “Using a Manifold Vocoder for Spectral Voice and Style Conversion.” In: *INTERSPEECH*. 2019, pp. 1388–1392 (cit. on pp. 114, 155).

[DL67] Allen T Dittmann and Lynn G Llewellyn. “The phonemic clause as a unit of speech decoding.” In: *Journal of personality and social psychology* 6.3 (1967), p. 341 (cit. on pp. 41, 42, 197).

[DL68] Allen T Dittmann and Lynn G Llewellyn. “Relationship between vocalizations and head nods as listener responses.” In: *Journal of personality and social psychology* 9.1 (1968), p. 79 (cit. on pp. 41, 42).

[Dri+19] Hanna Drimalla, Niels Landwehr, Ursula Hess, et al. “From face to face: the contribution of facial mimicry to cognitive and emotional empathy”. In: *Cognition and Emotion* (2019) (cit. on p. 69).

[DH93a] Kent Drummond and Robert Hopper. “Back channels revisited: Acknowledgment tokens and speakership incipiency”. In: *Research on language and Social Interaction* 26.2 (1993), pp. 157–177 (cit. on pp. 42–44, 196–198).

[DH93b] Kent Drummond and Robert Hopper. “Some uses of yeah”. In: *Research on language and social interaction* 26.2 (1993), pp. 203–212 (cit. on pp. 42, 43, 46, 197, 198).

[Du +93] John W Du Bois, Stephan Schuetze-Coburn, Susanna Cumming, et al. “Outline of discourse transcription”. In: *Talking data*. Psychology Press, 1993, pp. 45–89 (cit. on p. 42).

[Dun72] Starkey Duncan. “Some signals and rules for taking speaking turns in conversations.” In: *Journal of personality and social psychology* 23.2 (1972), p. 283 (cit. on pp. 26, 41, 42, 196, 198).

[Dun74] Starkey Duncan. “On the structure of speaker–auditor interaction during speaking turns1”. In: *Language in society* 3.2 (1974), pp. 161–180 (cit. on pp. 41, 42).

[DN74] Starkey Duncan Jr and George Niederehe. “On signalling that it’s your turn to speak”. In: *Journal of experimental social psychology* 10.3 (1974), pp. 234–247 (cit. on pp. 41, 42, 196, 198).

[Dut+96] Thierry Dutoit, Vincent Pagel, Nicolas Pierret, et al. “The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes”. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Vol. 3. IEEE. 1996, pp. 1393–1396 (cit. on p. 96).

[Edl+10] Jens Edlund, Jonas Beskow, Kjell Elenius, et al. “Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture.” In: *LREC*. 2010, pp. 2992–2995 (cit. on p. 58).

[EHH09] Jens Edlund, Mattias Heldner, and Julia Hirschberg. “Pause and gap length in face-to-face interaction”. In: *Interspeech 2009*. 2009, pp. 2779–2782 (cit. on pp. 71, 74).

[EHP09] Jens Edlund, Mattias Heldner, and Antoine Pelcé. “Prosodic features of very short utterances in dialogue”. In: *Nordic Prosody-Proceedings of the Xth Conference*. Frankfurt am Main. 2009, pp. 57–68 (cit. on p. 121).

[Fer+22] Raul Fernandez, David Haws, Guy Lorberbom, et al. “Transplantation of Conversational Speaking Style with Interjections in Sequence-to-Sequence Speech Synthesis”. In: *Proc. Interspeech 2022*. 2022, pp. 5488–5492 (cit. on pp. 98, 153).

[Fig+22] Carol Figueroa, Adaeze Adigwe, Magalie Ochs, et al. “Annotation of Communicative Functions of Short Feedback Tokens in Switchboard”. In: *Proc. LREC*. June 2022, pp. 1849–1859 (cit. on pp. 38, 100, 103, 117, 122, 129, 140, 145, 152, 158, 161).

[FBS23] Carol Figueroa, Štefan Beňuš, and Gabriel Skantze. “Prosodic Alignment in Different Conversational Feedback Functions”. In: *20th International Congress of Phonetic Sciences (ICPhS)*. 2023 (cit. on pp. 38, 138).

[Fig+24] Carol Figueroa, Marcel de Korte, Magalie Ochs, et al. “Mhm... Yeah? Okay! Evaluating the Naturalness and Communicative Function of Synthesized Feedback Responses in Spoken Dialogue”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2024, pp. 544–553 (cit. on pp. 39, 151).

[FOS23] Carol Figueroa, Magalie Ochs, and Gabriel Skantze. “Classification of Feedback Functions in Spoken Dialog Using Large Language Models and Prosodic Features”. In: *27th Workshop on the Semantics and Pragmatics of Dialogue*. 2023 (cit. on pp. 38, 118, 210).

[Fis78] Pamela M Fishman. “Interaction: The Work Women Do”. In: *Social Problems* 25.4 (1978), pp. 397–406 (cit. on pp. 42, 46, 196–198).

[FC04] George Forman and Ira Cohen. “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*. Springer. 2004, pp. 161–172 (cit. on p. 128).

[FLP12] Heather Friedberg, Diane Litman, and Susannah B. F. Paletz. "Lexical entrainment and success in student engineering groups". In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. 2012, pp. 404–409 (cit. on p. 69).

[Fri52] Charles Carpenter Fries. *The structure of English: An introduction to the construction of English sentences*. Harcourt Brace World, 1952 (cit. on pp. 41, 42, 44, 46, 196–198).

[FFK04] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. "A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information". In: *Proc. ICARA Int. Conference on Autonomous Robots and Agents*. Citeseer. 2004, pp. 379–384 (cit. on pp. 88, 89, 94, 200, 202).

[FFK05] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. "Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system". In: *Proc. Interspeech 2005*. 2005, pp. 889–892 (cit. on pp. 88, 89, 94, 152, 200, 202).

[Fuj09] Donna T Fujimoto. "Listener responses in interaction: A case for abandoning the term, backchannel". In: (2009) (cit. on pp. 40, 42–44, 196–198).

[GYC21] Tianyu Gao, Xingcheng Yao, and Danqi 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 (cit. on p. 124).

[GCC91] Howard Giles, Nikolas Coupland, and Justine Coupland. "Accommodation theory: Communication, context, and consequence". In: *Contexts of accommodation: Developments in applied sociolinguistics* 1 (1991), pp. 1–68 (cit. on p. 69).

[Gil+87] Howard Giles, Anthony Mulac, James J Bradac, et al. "Speech accommodation theory: The first decade and beyond". In: *Communication yearbook 10*. Routledge, 1987, pp. 13–48 (cit. on p. 69).

[GO+07] Howard Giles, Tania Ogay, et al. "Communication accommodation theory". In: *Explaining communication: Contemporary theories and exemplars* (2007), pp. 293–310 (cit. on p. 69).

[GHM92] John J Godfrey, Edward C Holliman, and Jane 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 (cit. on pp. 32, 102).

[Goo86] Charles Goodwin. "Between and within: Alternative sequential treatments of continuers and assessments". In: *Human studies* 9 (1986), pp. 205–217 (cit. on pp. 42, 43, 46, 134, 197).

[GG87] Charles Goodwin and Marjorie Harness Goodwin. “Concurrent operations on talk: Notes on the interactive organization of assessments”. In: *IPrA papers in pragmatics* 1.1 (1987), pp. 1–54 (cit. on pp. 42, 196).

[Gra+07a] Agustín Gravano, Štefan Beňuš, Héctor Chávez, et al. “On the role of context and prosody in the interpretation of ‘okay’”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 2007, pp. 800–807 (cit. on pp. 31, 60, 197).

[Gra+07b] Agustín Gravano, Štefan Beňuš, Julia Hirschberg, et al. “Classification of discourse functions of affirmative words in spoken dialogue.” In: *Interspeech*. 2007, pp. 1613–1616 (cit. on pp. 31, 45, 49, 50, 52, 63–65, 136, 196–198).

[GH09] Agustín Gravano and Julia Hirschberg. “Backchannel-inviting cues in task-oriented dialogue”. In: *Proc. Interspeech 2009*. 2009, pp. 1019–1022 (cit. on pp. 31, 83, 140, 145, 148).

[GN10] Joakim Gustafson and Daniel Neiberg. “Prosodic cues to engagement in non-lexical response tokens in Swedish”. In: *DiSS-LPSS Joint Workshop 2010*. 2010 (cit. on pp. 56, 76, 80, 81, 139).

[Har+18] Kohei Hara, Koji Inoue, Katsuya Takanashi, et al. “Prediction of turn-taking using multitask learning with prediction of backchannels and fillers”. In: *Listener* 162 (2018), p. 364 (cit. on pp. 86, 199, 202).

[HPV04] Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. “Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals”. In: *Psychological science* 15.6 (2004), pp. 409–414 (cit. on p. 69).

[Hel+10] Mattias Heldner, Jens Edlund, Samer Al Moubayed, et al. “Very Short Utterances in Conversation”. In: *Proceedings in FONETIK 2010*. 2010 (cit. on p. 121).

[HEH10] Mattias Heldner, Jens Edlund, and Julia Hirschberg. “Pitch similarity in the vicinity of backchannels”. In: *Proc. Interspeech 2010*. 2010, pp. 3054–3057 (cit. on pp. 27, 36, 42, 68, 70, 75–78, 80–82, 139–142, 145, 165, 168).

[HHE13] Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. “Backchannel relevance spaces”. In: *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*. Peter Lang Publishing Group. 2013, pp. 137–146 (cit. on pp. 25, 87).

[Her84] John Heritage. “13. A change-of-state token and aspects of its sequential placement”. In: (1984) (cit. on pp. 42, 196–198).

[HA07] Dirk Heylen and Rieks op den Akker. “Computing backchannel distributions in multi-party conversations”. In: *Proceedings of the Workshop on Embodied Language Processing*. 2007, pp. 17–24 (cit. on pp. 61, 63, 128).

[Hir94] Lynette Hirschman. “Female–male differences in conversational interaction”. In: *Language in society* 23.3 (1994), pp. 427–442 (cit. on pp. 42, 196–198).

[Hja08] Anna Hjalmarsson. “Speaking without knowing what to say... or when to end”. In: *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. 2008, pp. 72–75 (cit. on p. 64).

[Hoc93] Beth Ann Hockey. “Prosody and the role of okay and uh-huh in discourse”. In: *Proceedings of the eastern states conference on linguistics*. Citeseer. 1993, pp. 128–136 (cit. on pp. 54, 55).

[Hoe+19] Rens Hoegen, Deepali Aneja, Daniel McDuff, et al. “An End-to-End Conversational Style Matching Agent”. In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 2019, pp. 111–118 (cit. on p. 140).

[HW11] Judith Holler and Katie Wilkin. “Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue”. In: *Journal of Nonverbal Behavior* 35 (2011), pp. 133–153 (cit. on p. 69).

[Hsu+21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al. “Hubert: Self-supervised speech representation learning by masked prediction of hidden units”. In: *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), pp. 3451–3460 (cit. on pp. 136, 169).

[Hu+22] Yushi Hu, Chunlei Zhang, Jiatong Shi, et al. “ProsodyBERT: Self-supervised prosody representation for style-controllable TTS”. In: (2022) (cit. on p. 169).

[HMG11] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. “Virtual Rapport 2.0”. In: *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer. 2011, pp. 68–79 (cit. on p. 30).

[HK16] Yuchi Huang and Saad Khan. “Mirroring facial expressions: Evidence from visual analysis of dyadic interactions”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 2016, pp. 225–228 (cit. on p. 69).

[HNP23] Lena Marie Huttner, Noël Nguyen, and Martin J Pickering. “Investigating the perception production link through perceptual adaptation and phonetic convergence”. In: *24th International Speech Communication Association, Interspeech 2023*. ISCA. 2023, pp. 3103–3107 (cit. on p. 69).

[Ind+13] Benjamin Inden, Zofia Malisz, Petra Wagner, et al. “Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent”. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. 2013, pp. 181–188 (cit. on pp. 42, 84, 170).

[Ino+20] Koji Inoue, Divesh Lala, Kenta Yamamoto, et al. "An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2020, pp. 118–127 (cit. on p. 82).

[Ish+21] Ryo Ishii, Xutong Ren, Michal Muszynski, et al. "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 (cit. on pp. 30, 86, 152, 199, 202).

[IJ17] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. 2017 (cit. on p. 156).

[JTD18] Yannick Jadoul, Bill Thompson, and Bart De Boer. "Introducing parselmouth: A python interface to praat". In: *Journal of Phonetics* 71 (2018), pp. 1–15 (cit. on pp. 123, 141).

[Jai+21] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, et al. "Exploring semi-supervised learning for predicting listener backchannels". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–12 (cit. on pp. 25, 85, 170).

[Jan+21] Jin Yea Jang, San Kim, Minyoung Jung, et al. "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 (cit. on pp. 30, 90, 200, 203, 205).

[Jef84] Gail Jefferson. "Notes on a systematic deployment of the acknowledgement tokens "yeah"; and "mm hm"". In: (1984) (cit. on pp. 42, 43, 196–198).

[Jur+98] Dan Jurafsky, Elizabeth Shriberg, Barbara Fox, et al. "Lexical, prosodic, and syntactic cues for dialog acts". In: *Discourse relations and discourse markers*. 1998 (cit. on pp. 32, 54, 55, 102, 103, 105, 111, 113–116, 136, 196–198).

[KO23] Sofoklis Kakouros and Johannah O'Mahony. "What does BERT learn about prosody?" In: *International Congress of Phonetic Sciences*. GUARANT International spol. s ro. 2023, pp. 1454–1458 (cit. on p. 169).

[Kaw+15a] Tatsuya Kawahara, Miki Uesato, Koichiro Yoshino, et al. "Toward adaptive generation of backchannels for attentive listening agents". In: *International workshop serien on spoken dialogue systems technology*. 2015, pp. 1–10 (cit. on pp. 76, 82, 95, 139, 141, 143, 147, 148, 152, 201, 203).

[Kaw+16] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, et al. "Prediction and Generation of Backchannel Form for Attentive Listening Systems". In: *Interspeech 2016*. 2016, pp. 2890–2894 (cit. on pp. 88, 89, 152, 200, 202).

[Kaw+15b] Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, et al. “Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening”. In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 392–395 (cit. on pp. 76, 81, 82, 139, 141, 143, 147, 148).

[Ken67] Adam Kendon. “Some functions of gaze-direction in social interaction”. In: *Acta psychologica* 26 (1967), pp. 22–63 (cit. on pp. 27, 41, 42, 44, 46, 196, 197).

[Kho+20] Pavel Kholiavin, Anna Mamushina, Daniil Kocharov, et al. “Automatic detection of backchannels in Russian dialogue speech”. In: *International Conference on Speech and Computer*. Springer. 2020, pp. 204–213 (cit. on p. 121).

[Kir+23] Ambika Kirkland, Shivam Mehta, Harm Lameris, et al. “Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation”. In: *12th ISCA Speech Synthesis Workshop (SSW2023)*. 2023, pp. 41–47 (cit. on p. 153).

[Kje09] Göran Kjellmer. “Where do we backchannel?: On the use of mm, mhm, uh huh and such like”. In: *International Journal of Corpus Linguistics* 14.1 (2009), pp. 81–112 (cit. on p. 134).

[Koi+98] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, et al. “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs”. In: *Language and speech* 41.3-4 (1998), pp. 295–321 (cit. on p. 83).

[KH12a] Iwan de Kok and Dirk Heylen. “A survey on evaluation metrics for backchannel prediction models”. In: *Interdisciplinary Workshop on Feedback Behaviors in Dialog, Stevenson, Washington, USA: Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. University of Texas. 2012, pp. 15–18 (cit. on pp. 84, 85, 87).

[KH12b] Iwan de Kok and Dirk Heylen. “Integrating backchannel prediction models into embodied conversational agents”. In: *International Conference on Intelligent Virtual Agents*. Springer. 2012, pp. 268–274 (cit. on pp. 84, 170).

[KB04] John Kominek and Alan W Black. “The CMU Arctic speech databases”. In: *Fifth ISCA Workshop on Speech Synthesis*. 2004 (cit. on p. 156).

[KKB20] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17022–17033 (cit. on p. 155).

[Kop+08] Stefan Kopp, Jens Allwood, Karl Grammer, et al. "Modeling embodied feedback with virtual humans". In: *Modeling Communication with Robots and Virtual Humans: Second ZiF Research Group International Workshop on Embodied Communication in Humans and Machines, Bielefeld, Germany, April 5-8, 2006, Revised Selected Papers*. Springer. 2008, pp. 18–37 (cit. on pp. 30, 45, 84, 94).

[KSG07] Stefan Kopp, Thorsten Stocksmeier, and Dafydd Gibbon. "Incremental multimodal feedback for conversational agents". In: *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*. Springer. 2007, pp. 139–146 (cit. on pp. 84, 94).

[Kou+13] Spyridon Kousidis, Zofia Malisz, Petra Wagner, et al. "Exploring annotation of head gesture forms in spontaneous human interaction". In: *Proceedings of the Tilburg Gesture Meeting (TiGeR 2013)*. 2013 (cit. on p. 110).

[Kra+05] Wessel Kraaij, Thomas Hain, Mike Lincoln, et al. "The AMI meeting corpus". In: *Proc. International Conference on Methods and Techniques in Behavioral Research*. 2005, pp. 1–4 (cit. on p. 61).

[Kra+77] Robert M Krauss, Connie M Garlock, Peter D Bricker, et al. "The role of audible and visible back-channel responses in interpersonal communication." In: *Journal of personality and social psychology* 35.7 (1977), p. 523 (cit. on pp. 42, 196–198).

[KW66] Robert M Krauss and Sidney Weinheimer. "Concurrent feedback, confirmation, and the encoding of referents in verbal communication." In: *Journal of personality and social psychology* 4.3 (1966), p. 343 (cit. on pp. 41, 42, 196, 197).

[Kru+22] Joanna Kruyt, Štefan Beňuš, Catherine Faget, et al. "Prosodic and lexical entrainment in adults with and without schizophrenia". In: *Proc. Speech Prosody 2022*. 2022, pp. 125–129 (cit. on p. 69).

[Kru+23] Joanna Kruyt, Dorina de Jong, Alessandro D'Ausilio, et al. "Measuring Prosodic Entrainment in Conversation: A Review and Comparison of Different Methods". In: *Journal of Speech, Language, and Hearing Research* 66.11 (2023), pp. 4280–4314 (cit. on p. 75).

[Lai08] Catherine Lai. "Prosodic cues for backchannels and short questions: really?" In: *Fourth Conference on Speech Prosody 2008*. 2008, pp. 413–416 (cit. on pp. 32, 36, 61, 62, 128).

[Lai09] Catherine Lai. "Perceiving surprise on cue words: Prosody and semantics interact on right and really". In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. 2009, pp. 1963–1966 (cit. on pp. 32, 59, 114–116).

[Lal+22] Divesh Lala, Koji Inoue, Tatsuya Kawahara, et al. “Backchannel Generation Model for a Third Party Listener Agent”. In: *Proceedings of the 10th International Conference on Human-Agent Interaction*. 2022, pp. 114–122 (cit. on pp. 30, 90, 93, 94, 167, 200, 203, 205).

[Lal+17] Divesh Lala, Pierrick Milhorat, Koji Inoue, et al. “Attentive listening system with backchanneling, response generation and flexible turn-taking”. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 2017, pp. 127–136 (cit. on pp. 85–87, 199, 202).

[Lam+23] Harm Lameris, Ambika Kirkland, Joakim Gustafson, et al. “Situating Speech Synthesis: Investigating Contextual Factors in the Evaluation of Conversational TTS”. In: *12th ISCA Speech Synthesis Workshop (SSW2023)*. 2023, pp. 69–74 (cit. on p. 153).

[Łań21] Adrian 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 (cit. on pp. 34, 155).

[Lev+15a] Rivka Levitan, Štefan Beňuš, Agustín Gravano, et al. “Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison”. In: *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*. 2015, pp. 325–334 (cit. on pp. 69, 72, 142).

[Lev+15b] Rivka Levitan, Štefan Beňuš, Agustín Gravano, et al. “Entrainment and turn-taking in human-human dialogue”. In: *2015 AAAI Spring Symposium Series*. 2015 (cit. on pp. 27, 36, 68, 70, 76, 79–82, 139, 142, 145, 147, 149, 165, 168, 169).

[Lev+16] Rivka Levitan, Beňuš, Štefan, Ramiro H Gálvez, et al. “Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar.” In: *Interspeech*. Vol. 16. San Francisco, CA. 2016, pp. 1166–1170 (cit. on p. 140).

[LGH11] Rivka Levitan, Agustín Gravano, and Julia Bell Hirschberg. “Entrainment in speech preceding backchannels”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 2011 (cit. on pp. 76–79, 82, 141, 142, 168).

[LH11] Rivka Levitan and Julia Hirschberg. “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions”. In: *Proc. Interspeech 2011*. 2011, pp. 3081–3084 (cit. on pp. 69, 72, 74).

[LXH18] Sarah Ita Levitan, Jessica Xiang, and Julia Hirschberg. “Acoustic-Prosodic and Lexical Entrainment in Deceptive Dialogue”. In: *Proc. Speech Prosody 2018*. 2018, pp. 532–536 (cit. on p. 69).

[Li15] Jamy Li. “The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents”. In: *International Journal of Human-Computer Studies* 77 (2015), pp. 23–37 (cit. on p. 27).

[Li+24] Yinghao Aaron Li, Xilin Jiang, Cong Han, et al. “StyleTTS-ZS: Efficient High-Quality Zero-Shot Text-to-Speech Synthesis with Distilled Time-Varying Style Diffusion”. In: *arXiv preprint arXiv:2409.10058* (2024) (cit. on p. 155).

[Lin+23] Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, et al. “On the utility of self-supervised models for prosody-related tasks”. In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2023, pp. 1104–1111 (cit. on pp. 136, 152).

[Liu06] Yang Liu. “Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus”. In: *Ninth International Conference on Spoken Language Processing*. 2006 (cit. on p. 120).

[Lou+12] Max M Louwerse, Rick Dale, Ellen G Bard, et al. “Behavior matching in multimodal communication is synchronized”. In: *Cognitive science* 36.8 (2012), pp. 1404–1426 (cit. on p. 69).

[LS16] Ewa Luger and Abigail Sellen. “" Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents”. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 5286–5297 (cit. on pp. 27, 29).

[Mal20] Bhavitvya Malik. *DialogTag*. 2020. URL: https://pypi.org/project/DialogTag (visited on 02/28/2023) (cit. on p. 124).

[Mal+12] Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Stefan Kopp, et al. “Prosodic characteristics of feedback expressions in distracted and non-distracted listeners”. In: *Proceedings of The Listening Talker. An Interdisciplinary Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions*. 2012, pp. 36–39 (cit. on p. 55).

[Mal+16] Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, et al. “The ALICO corpus: Analysing the active listener”. In: *Language resources and evaluation* 50.2 (2016), pp. 411–442 (cit. on pp. 45, 49, 106, 107).

[MB82] Daniel N Maltz and Ruth A Borker. “A cultural approach to male-female miscommunication”. In: *The matrix of language*. Routledge, 1982, pp. 81–98 (cit. on pp. 42, 46, 196, 198).

[May86] Senko K Maynard. “On back-channel behavior in Japanese and English casual conversation”. In: (1986) (cit. on pp. 42, 44, 46, 197, 198).

[McA+17] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, et al. “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi”. In: *Proc. Interspeech 2017*. 2017, pp. 498–502 (cit. on p. 155).

[McI06] Daniel N McIntosh. "Spontaneous facial mimicry, liking and emotional contagion". In: *Polish Psychological Bulletin* 37.1 (2006), p. 31 (cit. on p. 69).

[MSG14] Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. "Data-driven models for timing feedback responses in a Map Task dialogue system". In: *Computer Speech & Language* 28.4 (2014), pp. 903–922 (cit. on p. 42).

[MHS23] Kentaro Mitsui, Yukiya Hono, and Kei Sawada. "Towards human-like spoken dialogue generation between AI agents from written dialogue". In: *arXiv preprint arXiv:2310.01088* (2023) (cit. on pp. 99, 153).

[Mol+12] Lisette Mol, Emiel Krahmer, Alfons Maes, et al. "Adaptation in gesture: Converging hands or converging minds?" In: *Journal of Memory and Language* 66.1 (2012), pp. 249–264 (cit. on p. 69).

[MYO16] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE TRANSACTIONS on Information and Systems* 99.7 (2016), pp. 1877–1884 (cit. on pp. 114, 155).

[Mot+19] Nasim Motalebi, Eugene Cho, S Shyam Sundar, et al. "Can alexa be your therapist? how back-channeling transforms smart-speakers to be active listeners". In: *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*. 2019, pp. 309–313 (cit. on p. 96).

[Mue+15] Markus Mueller, David Leuschner, Lars Briem, et al. "Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques". In: *Human-Computer Interaction: Interaction Technologies: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II 17*. Springer. 2015, pp. 329–340 (cit. on pp. 85–87, 199, 202).

[Mül+22] Philipp Müller, Michael Dietz, Dominik Schiller, et al. "MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 7109–7114 (cit. on p. 121).

[NW22] Anindita Nath and Nigel Ward. "On the Predictability of the Prosody of Dialog Markers from the Prosody of the Local Context". In: *Speech Prosody 2022*. 2022, pp. 664–668 (cit. on pp. 94, 95, 152, 153, 161, 201, 203).

[NG10] Daniel Neiberg and Joakim Gustafson. "The prosody of Swedish conversational grunts". In: *Eleventh Annual Conference of the International Speech Communication Association*. 2010 (cit. on pp. 64, 66, 67, 148, 149).

[NG12a] Daniel Neiberg and Joakim Gustafson. "Cues to perceived functions of acted and spontaneous feedback expressions". In: *Feedback Behaviors in Dialog*. 2012 (cit. on pp. 46, 58, 59).

[NG12b] Daniel Neiberg and Joakim Gustafson. “Towards letting machines humming in the right way-prosodic analysis of six functions of short feedback tokens in english”. In: *Fonetik 2012* (2012) (cit. on p. 55).

[NSG13] Daniel Neiberg, Giampiero Salvi, and Joakim Gustafson. “Semi-supervised methods for exploring the acoustics of simple productive feedback”. In: *Speech Communication* 55.3 (2013), pp. 451–469 (cit. on pp. 26, 45, 48, 58, 102, 107).

[NGH08] Ani Nenkova, Agustın Gravano, and Julia Hirschberg. “High Frequency Word Entrainment in Spoken Dialogue”. In: *ACL-08: HLT* (2008), p. 169 (cit. on p. 140).

[OLK23] Johannah O’Mahony, Catherine Lai, and Simon King. “Synthesising turn-taking cues using natural conversational data”. In: *12th ISCA Speech Synthesis Workshop (SSW2023)*. 2023, pp. 75–80 (cit. on p. 153).

[OLS24] Johannah O’Mahony, Catherine Lai, and Éva Székely. “Well, what can you do with messy data? Exploring the prosody and pragmatic function of the discourse marker "well" with found data and speech synthesis”. In: *Interspeech 2024*. 2024, pp. 4084–4088 (cit. on p. 153).

[OMa+21] Johannah O’Mahony, Pilar Oplustil-Gallegos, Catherine Lai, et al. “Factors Affecting the Evaluation of Synthetic Speech in Context”. In: *11th ISCA Speech Synthesis Workshop (SSW 11)*. 2021, pp. 148–153. DOI: 10.21437/SSW.2021-26 (cit. on p. 153).

[Och+24] Keiko Ochi, Koji Inoue, Divesh Lala, et al. “Entrainment Analysis and Prosody Prediction of Subsequent Interlocutor’s Backchannels in Dialogue”. In: *Interspeech 2024*. 2024, pp. 462–466 (cit. on pp. 76, 82, 83, 95, 139, 143, 147, 148, 152, 201, 203).

[Oer+14] Catharine Oertel, Kenneth A Funes Mora, Samira Sheikhi, et al. “Who will get the grant? A multimodal corpus for the analysis of conversational behaviours in group interviews”. In: *Proceedings of the 2014 workshop on understanding and modeling multiparty, multimodal interactions*. 2014, pp. 27–32 (cit. on p. 98).

[OGB16] Catharine Oertel, Joakim Gustafson, and Alan W Black. “On data driven parametric backchannel synthesis for expressing attentiveness in conversational agents”. In: *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. 2016, pp. 43–47 (cit. on pp. 98, 153).

[Oer+17] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, et al. “Crowd-Sourced Design of Artificial Attentive Listeners”. In: *Interspeech 2017*. 2017, pp. 854–858 (cit. on pp. 37, 94).

[Oer+16] Catharine Oertel, José Lopes, Yu Yu, et al. “Towards building an attentive artificial listener: On the perception of attentiveness in audio-visual feedback tokens”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 21–28 (cit. on pp. 30, 84).

[OMF24] Tom Offrede, Christine Mooshammer, and Susanne Fuchs. “Breathing and Speech Adaptation: Do Speakers Adapt Toward a Confederate Talking Under Physical Effort?” In: *Journal of Speech, Language, and Hearing Research* (2024), pp. 1–17 (cit. on p. 69).

[Olm+21] Annie J Olmstead, Navin Viswanathan, Tiana Cowan, et al. “Phonetic adaptation in interlocutors with mismatched language backgrounds: A case for a phonetic synergy account”. In: *Journal of Phonetics* 87 (2021), p. 101054 (cit. on p. 69).

[Oni+23] Toshiki Onishi, Naoki Azuma, Shunichi Kinoshita, et al. “Prediction of Various Backchannel Utterances Based on Multimodal Information”. In: *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 2023, pp. 1–4 (cit. on pp. 30, 91, 200, 203, 206).

[OOK21] Pilar Oplustil-Gallegos, Johannah O’Mahony, and Simon King. “Comparing acoustic and textual representations of previous linguistic context for improving Text-to-Speech”. In: *11th ISCA Speech Synthesis Workshop (SSW 11)*. 2021, pp. 205–210 (cit. on p. 169).

[OLV20] Daniel Ortega, Chia-Yu Li, and Ngoc Thang 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 (cit. on pp. 30, 32, 90, 121, 168, 200, 203, 205).

[Ort+23] Daniel Ortega, Sarina Meyer, Antje Schweitzer, et al. “Modeling Speaker-Listener Interaction for Backchannel Prediction”. In: *arXiv preprint arXiv:2304.04472* (2023) (cit. on pp. 87, 91, 93, 200, 203, 206).

[OC21] Rachel Ostrand and Eleanor Chodroff. “It’s alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue”. In: *Journal of phonetics* 88 (2021), p. 101074 (cit. on p. 69).

[Pam11] Sathish Pammi. “Synthesis of listener vocalizations”. In: *Towards interactive speech synthesis (Ph. D. thesis). Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes, Saarbrücken, Germany* (2011) (cit. on p. 40).

[PS11] Sathish Pammi and Marc Schröder. “Evaluating the meaning of synthesized listener vocalizations”. In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011 (cit. on pp. 97, 98, 152, 161).

[PSC11] Sathish Pammi, Marc Schröder, and Marcela Charfuelan. "Multidimensional meaning annotation of listener vocalizations for synthesis". In: *Proceedings of the Workshop on Emotion and Computing (Berlin, Germany), Springer*. 2011 (cit. on p. 57).

[Pam+10] Sathish Pammi, Marc Schröder, Marcela Charfuelan, et al. "Synthesis of listener vocalisations with imposed intonation contours". In: *Proc. 7th ISCA Workshop on Speech Synthesis (SSW 7)*. 2010, pp. 240–245 (cit. on pp. 30, 97, 98, 152, 153, 161).

[Par06] Jennifer S Pardo. "On phonetic convergence during conversational interaction". In: *The Journal of the Acoustical Society of America* 119.4 (2006), pp. 2382–2393 (cit. on p. 69).

[Par+24a] Yo-Han Park, Wencke Liermann, Yong-Seok Choi, Seung Hi Kim, et al. "Backchannel prediction, based on who, when and what". In: *Interspeech 2024*. 2024, pp. 3570–3574 (cit. on pp. 92, 152, 200, 203, 207).

[Par+24b] Yo-Han Park, Wencke Liermann, Yong-Seok Choi, and Kong Joo Lee. "Improving Backchannel Prediction Leveraging Sequential and Attentive Context Awareness". In: *Findings of the Association for Computational Linguistics: EACL 2024*. 2024, pp. 1689–1694 (cit. on pp. 92, 168, 200, 203, 207).

[Ped+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on pp. 123, 128).

[Pen+19] Bruno Elias Penteado, Magalie Ochs, Roxane Bertrand, et al. "Evaluating temporal predictive features for virtual patients feedbacks". In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 2019, pp. 88–90 (cit. on pp. 84, 170).

[PG04] Martin J Pickering and Simon Garrod. "Toward a mechanistic psychology of dialogue". In: *Behavioral and brain sciences* 27.2 (2004), pp. 169–190 (cit. on pp. 69, 70).

[PG06] Martin J Pickering and Simon Garrod. "Alignment as the basis for successful communication". In: *Research on Language and Computation* 4 (2006), pp. 203–228 (cit. on p. 70).

[PTH11] Ronald Poppe, Khiet P Truong, and Dirk Heylen. "Backchannels: Quantity, type and timing matters". In: *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer. 2011, pp. 228–239 (cit. on pp. 85, 170).

[Pop+10] Ronald Poppe, Khiet P Truong, Dennis Reidsma, et al. "Backchannel strategies for artificial listeners". In: *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer. 2010, pp. 146–158 (cit. on p. 84).

[Por+18] Martin Porcheron, Joel E Fischer, Stuart Reeves, et al. "Voice interfaces in everyday life". In: *proceedings of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–12 (cit. on pp. 27–29).

[Pot+09] Emmanuel Pot, Jérôme Monceaux, Rodolphe Gelin, et al. "Choregraphe: a graphical tool for humanoid robot programming". In: *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2009, pp. 46–51 (cit. on p. 28).

[PGB16] Laurent Prévot, Jan Gorisch, and Roxane Bertrand. "A CUP of COFEE: A large collection of feedback utterances provided with communicative function annotations". In: (2016) (cit. on pp. 25, 45, 52, 107).

[PGM15] Laurent Prévot, Jan Gorish, and Sankar Mukherjee. "Annotation and classification of french feedback communicative functions". In: *The 29th Pacific Asia Conference on Language, Information and Computation*. 2015 (cit. on pp. 26, 36, 45, 50, 52, 67, 68, 102, 107, 121, 128).

[14] *Prolific*. https://www.prolific.com/. Oxford, Uk, 2014 (cit. on p. 157).

[QS24] Livia Qian and Gabriel Skantze. "Joint Learning of Context and Feedback Embeddings in Spoken Dialogue". In: *Interspeech 2024*. 2024, pp. 2955–2959 (cit. on pp. 94, 152, 169).

[02] *Qualtrics*. https://www.qualtrics.com/. Provo, Utah, 2002 (cit. on p. 157).

[R C23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/ (cit. on p. 158).

[RÖD20] Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. "Alignment in multimodal interaction: An integrative framework". In: *Cognitive science* 44.11 (2020), e12911 (cit. on p. 70).

[Ree+23] Andrew Reece, Gus Cooney, Peter Bull, et al. "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation". In: *Science Advances* 9.13 (2023), eadf3197 (cit. on pp. 33, 167).

[Rue+17] Robin Ruede, Markus Müller, Sebastian Stüker, et al. "Enhancing Backchannel Prediction Using Word Embeddings." In: *Interspeech*. 2017, pp. 879–883 (cit. on pp. 85–87, 199, 202).

[Rue+19] Robin Ruede, Markus Müller, Sebastian Stüker, et al. "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 (cit. on pp. 32, 86, 199, 202).

[SW06] Udit Sajjanhar and Nigel Ward. "Automatic labeling of back channels". In: (2006) (cit. on p. 121).

[San+11] Shotaro Sano, Shun Nishide, Hiroshi G Okuno, et al. "Predicting listener back-channels for human-agent interaction using neuro-dynamical model". In: *2011 IEEE/SICE International Symposium on System Integration (SII)*. IEEE. 2011, pp. 18–23 (cit. on pp. 86, 199, 202).

[SWG23] Simona Sbranna, Simon Wehrle, and Martine Grice. "The use of Backchannels and other very short utterances by Italian Learners of German". In: *Proceedings of XVIII AISV Conference "The Position of the Speaker in Interaction: Attitudes, Intentions, and Emotions in Verbal Communication*. 2023 (cit. on p. 31).

[Sch82] Emanuel A Schegloff. "Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences". In: *Analyzing discourse: Text and talk* 71 (1982), pp. 71–93 (cit. on pp. 41–43, 46, 107, 196–198).

[Sch+19] Steffen Schneider, Alexei Baevski, Ronan Collobert, et al. "wav2vec: Unsupervised pre-training for speech recognition". In: *arXiv preprint arXiv:1904.05862* (2019) (cit. on p. 136).

[Sch00] Marc Schröder. "Experimental study of affect bursts". In: *Proc. ITRW on Speech and Emotion*. 2000, pp. 132–137 (cit. on pp. 56, 57).

[Sch03] Marc Schröder. "Experimental study of affect bursts". In: *Speech communication* 40.1-2 (2003), pp. 99–116 (cit. on pp. 56, 57).

[SHP06] Marc Schröder, Dirk Heylen, and Isabella Poggi. "Perception of non-verbal emotional listener feedback". In: *Speech prosody 2006*. TUDpress. 2006, pp. 43–46 (cit. on pp. 56, 57).

[She+18] Jonathan Shen, Ruoming Pang, Ron J Weiss, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 4779–4783 (cit. on p. 98).

[She+23] Kai Shen, Zeqian Ju, Xu Tan, et al. "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers". In: *arXiv preprint arXiv:2304.09116* (2023) (cit. on p. 155).

[Shr+98] Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, et al. "Can prosody aid the automatic classification of dialog acts in conversational speech?" In: *Language and speech* 41.3-4 (1998), pp. 443–492 (cit. on p. 136).

[SK23] Atli Thor Sigurgeirsson and Simon King. "Do prosody transfer models transfer prosody*f*". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5 (cit. on pp. 152, 169).

[Sol+06] Thamar Solorio, Olac Fuentes, Nigel Ward, et al. "Prosodic feature generation for back-channel prediction." In: *INTERSPEECH*. Vol. 5. 2006, pp. 2398–2401 (cit. on pp. 86, 199, 202).

[Spa+24] Malin Spaniol, Simon Wehrle, Alicia Janz, et al. “The influence of conversational context on lexical and prosodic aspects of backchannels and gaze behaviour”. In: *Proc. Speech Prosody 2024*. 2024, pp. 607–611. DOI: 10.21437/SpeechProsody.2024-123 (cit. on p. 116).

[Ste87] Anna-Brita Stenström. “Carry-on signals in English conversation”. In: *Corpus linguistics and beyond*. Brill, 1987, pp. 87–119 (cit. on p. 42).

[SKG07] Thorsten Stocksmeier, Stefan Kopp, and Dafydd Gibbon. “Synthesis of prosodic attitudinal variants in German backchannel ja”. In: *Proc. Interspeech 2007*. 2007, pp. 1290–1293 (cit. on pp. 30, 37, 96, 97, 152, 161).

[Sto+00] Andreas Stolcke, Klaus Ries, Noah Coccaro, et al. “Dialogue act modeling for automatic tagging and recognition of conversational speech”. In: *Computational linguistics* 26.3 (2000), pp. 339–373 (cit. on p. 120).

[Stu98] Maria Stubbe. “Are you listening? Cultural influences on the use of supportive verbal feedback in conversation”. In: *Journal of Pragmatics* 29.3 (1998), pp. 257–289 (cit. on p. 134).

[Szé+19] Éva Székely, Gustav Eje Henter, Jonas Beskow, et al. “Spontaneous Conversational Speech Synthesis from Found Data”. In: *Interspeech 2019*. 2019, pp. 4435–4439 (cit. on p. 153).

[TBC98] Paul Taylor, Alan W Black, and Richard Caley. “The architecture of the Festival speech synthesis system”. In: (1998) (cit. on p. 98).

[Thó96] Kristinn Rúnar Thórisson. “Communicative humanoids: a computational model of psychosocial dialogue skills”. PhD thesis. Massachusetts Institute of Technology, 1996 (cit. on p. 84).

[Tot91] Gunnel Tottie. “Conversational style in British and American English: The case of backchannels”. In: *English corpus linguistics*. Routledge, 1991, pp. 254–271 (cit. on pp. 31, 42, 44–46, 166, 196–198).

[TH10] Khiet P Truong and Dirk Heylen. “Disambiguating the functions of conversational sounds with prosody: the case of ‘yeah’”. In: *Proc. Interspeech 2010*. 2010 (cit. on pp. 36, 61, 63, 64, 128).

[TPH10] Khiet P Truong, Ronald Walter Poppe, and Dirk Heylen. “A rule-based backchannel prediction model using pitch and pause information”. In: *11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. International Speech Communication Association (ISCA). 2010, pp. 3058–3061 (cit. on pp. 86, 199, 202).

[Tru+11] Khiet P Truong, Ronald Walter Poppe, IA de Kok, et al. “A multimodal analysis of vocal and visual backchannels in spontaneous dialogs”. In: *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*. International Speech Communication Association (ISCA). 2011, pp. 2973–2976 (cit. on p. 83).

[Uen04] Junko Ueno. “Gender differences in Japanese conversation”. In: *Intercultural Communication Studies* 13 (2004), pp. 85–100 (cit. on p. 134).

[VRL22] Emelie Van De Vreken, Korin Richmond, and Catherine Lai. “Voice Puppetry with FastPitch”. In: *Interspeech 2022*. 2022, pp. 5219–5220 (cit. on p. 155).

[Wal06] Åsa Wallers. *Minor sounds of major importance-prosodic manipulation of synthetic backchannels in Swedish*. 2006 (cit. on p. 37).

[WES06] Åsa Wallers, Jens Edlund, and Gabriel Skantze. “The effect of prosodic features on the interpretation of synthesised backchannels”. In: *Perception and Interactive Technologies: International Tutorial and Research Workshop, PIT 2006 Kloster Irsee, Germany, June 19-21, 2006. Proceedings*. Springer. 2006, pp. 183–187 (cit. on pp. 30, 96, 152, 161).

[Wan+24] Jinhan Wang, Long Chen, Aparna Khare, et al. “Turn-taking and backchannel prediction with acoustic and large language model fusion”. In: *arXiv preprint arXiv:2401.14717* (2024) (cit. on pp. 30, 32, 86, 152, 199, 202).

[War96] Nigel Ward. “Using prosodic clues to decide when to produce backchannel utterances”. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96*. Vol. 3. IEEE. 1996, pp. 1728–1731 (cit. on p. 83).

[War04] Nigel Ward. “Pragmatic functions of prosodic features in non-lexical utterances”. In: *Speech Prosody 2004, International Conference*. 2004 (cit. on p. 54).

[War06] Nigel Ward. “Non-lexical conversational sounds in American English”. In: *Pragmatics & Cognition* 14.1 (2006), pp. 129–182 (cit. on p. 54).

[War+07] Nigel Ward, Rafael Escalante, Yaffa Al Bayyari, et al. “Learning to show you’re listening”. In: *Computer Assisted Language Learning* 20.4 (2007), pp. 385–407 (cit. on p. 31).

[WT00] Nigel Ward and Wataru Tsukahara. “Prosodic features which cue backchannel responses in English and Japanese”. In: *Journal of pragmatics* 32.8 (2000), pp. 1177–1207 (cit. on pp. 30, 42–44, 46, 86, 166, 196–199, 202).

[Wei+21] Andreas Weise, Vered Silber-Varod, Anat Lerner, et al. “Talk to me with left, right, and angles”: Lexical entrainment in spoken Hebrew dialogue”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 292–299 (cit. on p. 69).

[Whi89] Sheida White. “Backchannels across cultures: A study of Americans and Japanese1”. In: *Language in society* 18.1 (1989), pp. 59–76 (cit. on pp. 42, 44, 196–198).

[Wie19] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. 2019 (cit. on p. 44).

[Wol08] James P Wolf. "The effects of backchannels on fluency in L2 oral task production". In: *System* 36.2 (2008), pp. 279–294 (cit. on p. 31).

[WB22] Camille J Wynn and Stephanie A Borrie. "Classifying conversational entrainment of speech behavior: An expanded framework and review". In: *Journal of Phonetics* 94 (2022), p. 101173 (cit. on pp. 27, 69, 72–75, 77, 79–82, 142).

[Yam+16] Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, et al. "Analysis and prediction of morphological patterns of backchannels for attentive listening agents". In: *Proc. 7th International Workshop on Spoken Dialogue Systems*. 2016, pp. 1–12 (cit. on pp. 88, 89, 152, 200, 202).

[Yam+21] Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, et al. "Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems". In: *Interspeech 2021*. 2021, pp. 246–250. DOI: 10.21437/Interspeech.2021-381 (cit. on p. 140).

[Yng70] Victor H Yngve. "On getting a word in edgewise". In: *Papers from the 6th Regional Meeting Chicago Linguistic Society*. 1970, pp. 567–578 (cit. on pp. 26, 41–43, 46, 197, 198).

[YM04] Richard F Young and Elizabeth R Miller. "Learning as changing participation: Discourse roles in ESL writing conferences". In: *The Modern Language Journal* 88.4 (2004), pp. 519–535 (cit. on p. 42).

[Zha+23] Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. "A survey of large language models". In: *arXiv preprint arXiv:2303.18223* (2023) (cit. on p. 29).

[Zim93] Don H Zimmerman. "Acknowledgment tokens and speakership incipiency revisited". In: *Research on language and social interaction* 26.2 (1993), pp. 179–194 (cit. on pp. 42, 43).

[ZW75] Don H Zimmermann and Candace West. "Sex roles, interruptions and silences in conversation". In: *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*. John Benjamins BV, 1975, pp. 211–236 (cit. on pp. 42, 46, 197, 198).

APPENDICES

A. Compilation lists of lexical forms of feedback responses in literature

Lexical Forms	Publications
absolutely	Jurafsky et al. 1998
ah	C. Goodwin and M. H. Goodwin 1987
alright	Beňuš et al. 2007, Gravano, Beňuš, Hirschberg, et al. 2007, Fujimoto 2009
exactly	Clancy et al. 1996, Jurafsky et al. 1998
fine	Fries 1952
gee	Clancy et al. 1996
good	Fries 1952
gosh	Kendon 1967
gotcha	Gravano, Beňuš, Hirschberg, et al. 2007
heavens	Kendon 1967
hell	Clancy et al. 1996
hm, hmm	White 1989, Hirschman 1994, Clancy et al. 1996, Ward and Tsukahara 2000, Fujimoto 2009
huh	Krauss and Weinheimer 1966, Fishman 1978, Clancy et al. 1996, Jurafsky et al. 1998, Gravano, Beňuš, Hirschberg, et al. 2007
huh-uh	Jurafsky et al. 1998
hunh	Tottie 1991
m, mm, mmm	Kendon 1967, Fishman 1978, Jefferson 1984, Tottie 1991, Fujimoto 2009
m-hm, mhm, mm hm, mm hmm, mm-hmm, mmhm	Duncan 1972, Duncan Jr et al. 1974, Krauss, Garlock, et al. 1977, Maltz et al. 1982, Schegloff 1982, Heritage 1984, Jefferson 1984, White 1989, Tottie 1991, Hirschman 1994, Drummond et al. 1993a, Clancy et al. 1996, Bavelas et al. 2000, Beňuš et al. 2007, Gravano, Beňuš, Hirschberg, et al. 2007, Fujimoto 2009
man	Clancy et al. 1996
mn-hm	Fujimoto 2009
no	Tottie 1991, Jurafsky et al. 1998

Table A.1. – Alphabetized list of feedback lexical forms found in literature ranging from A-N.

Lexical Forms	Publications
o.k., ok, okay	Krauss and Weinheimer 1966, Yngve 1970, Drummond et al. 1993a, Hirschman 1994, Clancy et al. 1996, Jurafsky et al. 1998, Ward and Tsukahara 2000, Beňuš et al. 2007, Gravano, Beňuš, Hirschberg, et al. 2007, Fujimoto 2009
oh	Fries 1952, Fishman 1978, Heritage 1984, C. Goodwin 1986, White 1989, Tottie 1991, Drummond et al. 1993a, Drummond et al. 1993b, Hirschman 1994, Clancy et al. 1996, Jurafsky et al. 1998, Fujimoto 2009,
really	White 1989, Tottie 1991, Clancy et al. 1996, Jurafsky et al. 1998, Fujimoto 2009
right	Schegloff 1982, White 1989, Tottie 1991, Hirschman 1994, Jurafsky et al. 1998, Ward and Tsukahara 2000, Beňuš et al. 2007, Gravano, Beňuš, Chávez, et al. 2007, Fujimoto 2009
shit	Clancy et al. 1996
so	Fries 1952
sure	Krauss, Garlock, et al. 1977, Tottie 1991, Clancy et al. 1996, Jurafsky et al. 1998
surely	Kendon 1967
uh	Jurafsky et al. 1998
uh huh, uh-huh, uhhuh	Yngve 1970, Zimmermann et al. 1975, Schegloff 1982, Jefferson 1984, Maynard 1986, C. Goodwin 1986, White 1989, Tottie 1991, Drummond et al. 1993a, Drummond et al. 1993b, Clancy et al. 1996, Jurafsky et al. 1998, Ward and Tsukahara 2000, Bavelas et al. 2000,Beňuš et al. 2007 , Gravano, Beňuš, Hirschberg, et al. 2007, Fujimoto 2009
unh hunh	Fries 1952
uhh-hunh	Fujimoto 2009
um, umm	Zimmermann et al. 1975, Fishman 1978, Jurafsky et al. 1998, Fujimoto 2009
um hm, um hmm, um-hmm	Dittmann et al. 1967, Zimmermann et al. 1975, Drummond et al. 1993b

Table A.2. – Alphabetized list of feedback lexical forms found in literature ranging from O-U.

Lexical Forms	Publications
well	Hirschman 1994
wonderful	Schegloff 1982
wow	Schegloff 1982
yeah	Fries 1952, Duncan 1972, Duncan Jr et al. 1974, Zimmermann et al. 1975, Fishman 1978, Schegloff 1982, Jefferson 1984, White 1989, Tottie 1991, Drummond et al. 1993a, Drummond et al. 1993b, Hirschman 1994, Clancy et al. 1996, Jurafsky et al. 1998, Bavelas et al. 2000, Beňuš et al. 2007, Ward and Tsukahara 2000, Fujimoto 2009, Gravano, Beňuš, Hirschberg, et al. 2007
yeh	Fujimoto 2009
yep	Beňuš et al. 2007, Gravano, Beňuš, Hirschberg, et al. 2007
yes	Fries 1952, Yngve 1970, Krauss, Garlock, et al. 1977, Maltz et al. 1982 , Heritage 1984, Jefferson 1984, Maynard 1986, White 1989, Tottie 1991, Jurafsky et al. 1998, Beňuš et al. 2007, Gravano, Beňuš, Hirschberg, et al. 2007,
yup	Beňuš et al. 2007, Gravano, Beňuš, Hirschberg, et al. 2007

Table A.3. – Alphabetized list of feedback lexical forms found in literature ranging from W-Z.

B. Language and Corpus type in modeling feedback responses

Publication	Language(s)	Corpus type(s)
Ward and Tsukahara 2000	U.S. English, Japanese	Conversational
Cathcart et al. 2003	Scottish English	Task-oriented
Solorio et al. 2006	Iraqi English	Conversational
Truong, R. W. Poppe, and Heylen 2010	Dutch	Conversational
Sano et al. 2011	Japanese	Explanations of posters
Mueller et al. 2015	U.S English*	Conversational
Lala, Milhorat, et al. 2017	Japanese	Counseling conversations
Ruede et al. 2017	U.S English*	Conversational
Hara et al. 2018	Japanese	Job interviews, conversational
Ruede et al. 2019	U.S. English*	Conversational
Adiba, Homma, Bertero, et al. 2021	U.S. English*	Conversational
Ishii et al. 2021	Japanese	Conversational
Jinhan Wang et al. 2024	U.S. English*	Conversational

Table B.1. – List of language and corpus type used in publications predicting the timing of feedback responses. * denotes the Switchboard corpus.

Publication	Language(s)	Corpus type(s)
Fujie et al. 2004, Fujie et al. 2005	Japanese	Conversational
Kawahara, Yamaguchi, Inoue, et al. 2016	Japanese	Counseling Conversations
Yamaguchi et al. 2016	Japanese	Counseling Conversations

Table B.2. – List of language and corpus type used in publications predicting the lexical form of feedback responses.

Publication	Language(s)	Corpus type(s)
Ortega, C.-Y. Li, et al. 2020	U.S. English*	Conversational
Adiba, Homma, and Miyoshi 2021	Japanese	Conversational
Boudin et al. 2021	French	Conversational
Jang et al. 2021	Korean	Counseling Conversations
Lala, Inoue, et al. 2022	Japanese	Radio show with added third party listener
Onishi et al. 2023	Japanese	Retelling animation
Ortega, Meyer, et al. 2023	German,	Conversational
	U.S. English*	Conversational
Boudin et al. 2024	French	Conversational
Choi et al. 2024	U.S. English*	Conversational
Park, Liermann, Choi, and Lee 2024	Korean, U.S. English*	Korean counseling conversations, U.S. English conversational speech
Park, Liermann, Choi, Kim, et al. 2024	Korean	Counseling Conversations

Table B.3. – List of language and corpus type used in publications predicting the communicative function of feedback responses. * denotes the Switchboard corpus.

Publication	**Language(s)**	**Corpus type(s)**
Kawahara, Uesato, et al. 2015	Japanese	Counseling conversations
Nath et al. 2022	U.S. English*	Conversational speech
Ochi et al. 2024	Japanese	Conversational speech, where participants talked to the humanoid robot ERICA in a wizard-of-oz setup.

Table B.4. – List of language and corpus type used in publications predicting the prosody of feedback responses. * denotes the Switchboard corpus.

C. Summary of evaluation metrics used for modeling feedback responses

Publication	Evaluation metric
Ward and Tsukahara 2000	Accuracy, coverage, figure of merit
Cathcart et al. 2003	Precision, Recall, F-score
Solorio et al. 2006	Precision, Recall, F-score
Truong, R. W. Poppe, and Heylen 2010	Precision, Recall, F-score
Sano et al. 2011	Precision, Recall, F-score
Mueller et al. 2015	Precision, Recall, F-score, subjective
Lala, Milhorat, et al. 2017	Precision, Recall, F-score, AUC, subjective
Ruede et al. 2017	Precision, Recall, F-score, subjective
Hara et al. 2018	Precision, Recall, F-score
Ruede et al. 2019	Precision, Recall, F-score
Adiba, Homma, Bertero, et al. 2021	Precision, Recall, F-score
Ishii et al. 2021	F-score
Jinhan Wang et al. 2024	AUC

Table C.1. – List of publications predicting the timing of feedback responses and the type of evaluation metrics reported.

Publication	Evaluation metric
Fujie et al. 2004, Fujie et al. 2005	subjective
Kawahara, Yamaguchi, Inoue, et al. 2016	Precision, Recall, F-score
Yamaguchi et al. 2016	Precision, Recall, F-score

Table C.2. – List of publications predicting the lexical form of feedback responses and the type of evaluation metrics reported.

Publication	Evaluation metric
Ortega, C.-Y. Li, et al. 2020	Accuracy
Adiba, Homma, and Miyoshi 2021	F-score
Boudin et al. 2021	Accuracy, Kappa
Jang et al. 2021	F-score
Lala, Inoue, et al. 2022	Precision, Recall, F-score, subjective
Onishi et al. 2023	Precision, Recall, F-score
Ortega, Meyer, et al. 2023	Accuracy, F-score
Boudin et al. 2024	Precision, Recall, F-score
Choi et al. 2024	F-score
Park, Liermann, Choi, and Lee 2024	F-score
Park, Liermann, Choi, Kim, et al. 2024	F-score

Table C.3. – List of publications predicting the communicative function of feedback responses and the type of evaluation metrics reported.

Publication	Evaluation metric
Kawahara, Uesato, et al. 2015	Mean Square Error
Nath et al. 2022	Root Mean Square Error
Ochi et al. 2024	Mean Absolute Error

Table C.4. – List of publications predicting the prosody of feedback responses and the type of evaluation metrics reported.

D. Communicative functions and definitions in predicting the functions of feedback responses

Publication	Prediction categories and definitions
Ortega, C.-Y. Li, et al. 2020	*BC- Continuer*: exhibit acknowledgment and understanding of what the speaker said, as well as encourage speaker to continue. *BC- Assessment*: are contextualized responses, so that the listener reacts to what the speaker uttered without intending to take the floor. *No-BC*
Adiba, Homma, and Miyoshi 2021	*Responsive interjections*: express acceptance, at various levels, of an other's utterance. *Expressive interjections*: are used when the speaker expresses a notice, surprise, disappointment, admiration, etc. elicited by an other's utterance or situation. *non-BC*
Boudin et al. 2021*	*Generic*:displaying attention to the speaker. *Specific*: expressing responses to the content of the speaker's production.
Jang et al. 2021	*Continuer*: short 'yey', 'ney' etc. or its competition expression of agreement *Understanding*: long 'yey' or long 'ney' *Empathetic Response*: exclamation or laughter *No BC*
Lala, Inoue, et al. 2022*	*Responsives*: show acceptance of another utterance. *Expressives*: expresses admiration, surprise, disappointment etc. elicited by another utterance. *Shared laughs*: laughing with another participant.

Table D.1. – List of categories predicted and their definitions in publications from 2020-2022 predicting the communicative function of feedback responses. * denotes there was a separate model predicting whether there should or should not be a feedback response.

Publication	Prediction categories and definitions
Onishi et al. 2023	*Neutral word*: utterances that do not include opinions or emotions. *Positive word*: utterances such as “unun”, “sousou”, “soreii”, “naruhodo” and “tashikani” in Japanese. *Non-positive word*: negative utterances such as “ūn", “hūn" and “iya" in Japanese. *Emotional word*: other short interjections such as “sugo”, “hē" and “wao" in Japanese. *Anticipation*: utterances anticipating the speaker’s topic. *Confirmation*: confirmation to prompt explanation. *Repetition of speaker’s utterance*: repetition of the speaker’s remarks. *Summary of speaker’s utterance*: summary of the speaker’s remarks. *Other*: utterances that do not belong to any of the other types.
Ortega, Meyer, et al. 2023	*Yeah*: expresses that the listener aligns themself with what the speaker just uttered and claims to have understood the message. *Uh-huh*: classic continuer response and commonly described as example of passive recipiency that apprises the speaker of carrying on. *No-BC*

Table D.2. – List of categories predicted and their definitions in publications from 2023 predicting the communicative function of feedback responses.

Publication	Prediction categories and definitions
Boudin et al. 2024*	*Generic*: items such as mh or nods which express interest and understanding. *Specific*: more complex verbal utterances, tone of voice, eyebrows movements, etc.) and involve an evaluative function. *Positive*: positive stance of the main speaker's speech. *Negative*: negative stance of the main speaker's speech. *Given*: information status. *New*: information status.
Choi et al. 2024	*Continuer*: utterances such as "uh-huh", "um-hum" and their variations. *Assessment*: unique utterances used as BCs. *Non-BC*
Park, Liermann, Choi, and Lee 2024	(Korean) *Continuer*: generic backchannels that signal a listener's undivided attention, ultimately encouraging the speaker to continue speaking. (Korean) *Understanding*: signals that the speaker has been understood. (Korean) *Empathetic*: actively expresses the listener's emotions and thoughts related to the speaker's utterance. (U.S. English) *Continuer*: generic form, similar to "Uh-huh". (U.S. English) *Assessment*: specific form. NoBC
Park, Liermann, Choi, Kim, et al. 2024	*Continuer*: a brief vocalic cue or repetition; signalling continued attention to the speaker. *Understanding*: covers longer cues that indicate that a speaker's statement has been understood. *Empathetic*: involve actively mirroring the listener's emotions, such as admiration, surprise, or agreement.

Table D.3. – List of categories predicted and their definitions in publications from 2024 predicting the communicative function of feedback responses. * denotes there was a separate model predicting whether there should or should not be a feedback response.

E. Tukey pair-wise tests

Function	Duration (s)	Tukey, p < 0.05
(S) Sympathy	0.580	> A, D, N, MS, C, A/C, Y, U
(SS) Strong Surprise	0.511	> A, D, N, MS, C, A/C, Y, U
(Ds) Disapproval	0.511	> A, D, N, MS, C, A/C, Y, U
(A) Agree	0.389	> C, A/C, U
(D) Disagree	0.360	
(N) No	0.357	
(MS) Mild Surprise	0.341	
(C) Continue	0.329	
A/C	0.320	
(Y) Yes	0.313	
(U) Non-understanding	0.288	

Table E.1. – Tukey pair-wise test for duration. Empty cells indicate that the duration of the function was not significantly greater than other functions.

Function	Mean Pitch (SD)	Tukey, p < 0.05
(U) Non-understanding	0.865	> D, SS, N, Y, C, Ds, A, A/C, S
(MS) Mild Surprise	0.514	> SS, N, Y, C, Ds, A, A/C, S
(D) Disagree	0.240	> C, A, A/C
(SS) Strong Surprise	0.065	> C, A, A/C
(N) No	0.043	
(Y) Yes	-0.110	
(C) Continue	-0.130	
(Ds) Disapproval	-0.160	
(A) Agree	-0.181	
A/C	-0.192	
(S) Sympathy	-0.226	

Table E.2. – Tukey pair-wise test for mean pitch. Empty cells indicate that the mean pitch of the function was not significantly greater than other functions.

Function	**Pitch Slope (SD)**	**Tukey, p < 0.05**
(U) Non-understanding	0.919	> C, A/C, Y, Ds, A, N, SS, S, D, MS
(C) Continue	-0.092	> A, SS, D, MS
A/C	-0.172	> SS, D, MS
(Y) Yes	-0.211	> MS
(Ds) Disapproval	-0.303	> MS
(A) Agree	-0.307	> MS
(N) No	-0.314	> MS
(SS) Strong Surprise	-0.483	> MS
(S) Sympathy	-0.572	
(D) Disagree	-0.582	
(MS) Mild Surprise	-0.817	

Table E.3. – Tukey pair-wise test for pitch slope. Empty cells indicate that the pitch slope of the function was not significantly greater than other functions.

Function	**Pitch Range (SD)**	**Tukey, p < 0.05**
(MS) Mild Surprise	2.675	> A, C, A/C, N
(U) Non-understanding	2.460	> A, C, A/C, N
(SS) Strong Surprise	2.403	> A, C, A/C, N
(Ds) Disapproval	2.283	> C, A/C
(S) Sympathy	2.061	
(D) Disagree	1.989	
(A) Agree	1.795	
(Y) Yes	1.768	
(C) Continue	1.674	
A/C	1.595	
(N) No	1.594	

Table E.4. – Tukey pair-wise test for pitch range. Empty cells indicate that the pitch range of the function was not significantly greater than other functions.

F. Top 30 part-of-speech bigrams

bigrams
VBD+VBN
NNP+VBD
PDT+DT
DT+UH
IN+NNP
WRB+JJ
VBG+DT
RB+VBZ
WP+IN
VB+VBD
NN+WDT
IN+UH
IN+CD
PRP+MD
DT+NNPS
VBP+NNP
IN+VBP
VB+TO
UH+UH
WDT+PRP
UH+WRB
PRP+VBG
RBS+RBS
PRP+VBN
NN+CC
JJ+VBP
VBD+RP
UH+HYPH
WP+VBZ
UH+VBP

Table F.1. – Top 30 part-of-speech bigrams used in Figueroa, Ochs, et al. (2023) as context features.

G. DialogTag examples

Examples were taken from https://pypi.org/project/DialogTag/#:~:text=Dialogue%20act%20classification%20is%20the,classification%20on%20the%20Switchboard%20corpus.

Dialog Act	**Example**
Statement-non-opinion	Me, I'm in the legal department.
Acknowledge (Backchannel)	Uh-huh.
Statement-opinion	I think it's great
Agree/Accept	That's exactly it.
Appreciation	I can imagine.
Yes-No-Question	Do you have to have any special training?
Yes answers	Yes.
Conventional-closing	Well, it's been nice talking to you.
Uninterpretable	But, uh, yeah
Wh-Question	Well, how old are you?
No answers	No.
Response Acknowledgement	Oh, okay.
Hedge	I don't know if I'm making any sense or not.
Declarative Yes-No-Question	So you can afford to get a house?
Other	Well give me a break, you know.
Backchannel in question form	Is that right?
Quotation	You can't be pregnant and have cats
Summarize/reformulate	Oh, you mean you switched schools for the kids.
Affirmative non-yes answers	It is.
Action-directive	Why don't you go first
Collaborative Completion	Who aren't contributing.
Repeat-phrase	Oh, fajitas
Open-Question	How about you?
Rhetorical-Questions	Who would steal a newspaper?
Hold before answer/agreement	I'm drawing a blank.
Negative non-no answers	Uh, not a whole lot.
Signal-non-understanding	Excuse me?
Conventional-opening	How are you?
Or-Clause	or is it more of a company?
Dispreferred answers	Well, not so much that.
3rd-party-talk	My goodness, Diane, get down from there.
Offers, Options Commits	I'll have to check that out
Self-talk	What's the word I'm looking for
Downplayer	That's all right.
Maybe/Accept-part	Something like that
Tag-Question	Right?
Declarative Wh-Question	You are what kind of buff?
Apology	I'm sorry.
Thanking	Hey thanks a lot

Table G.1. – Examples of the 39 dialog acts available from DialogTag.

H. Confusion matrices for feedback function classification

H.1. Lexical feature

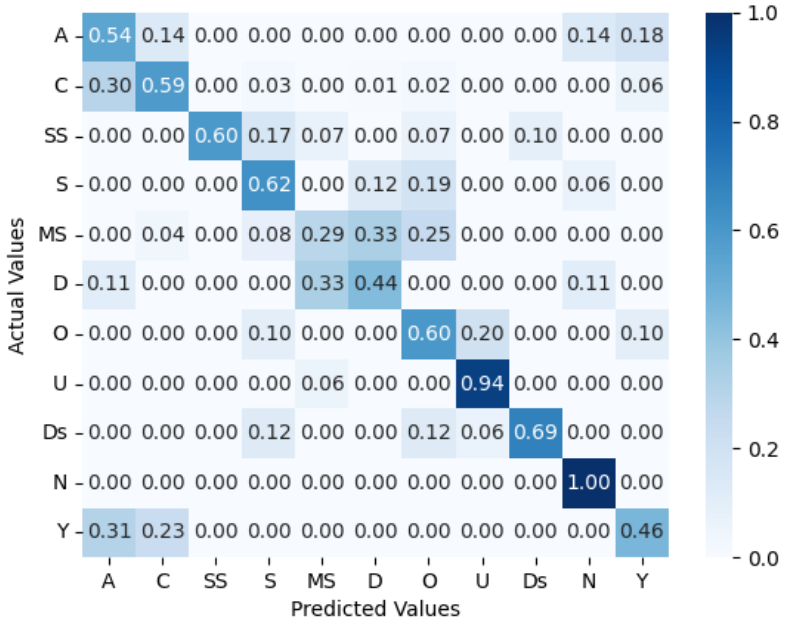


Figure H.1. – Confusion matrix for Model 1. Features = lexical.

H.2. Prosodic features

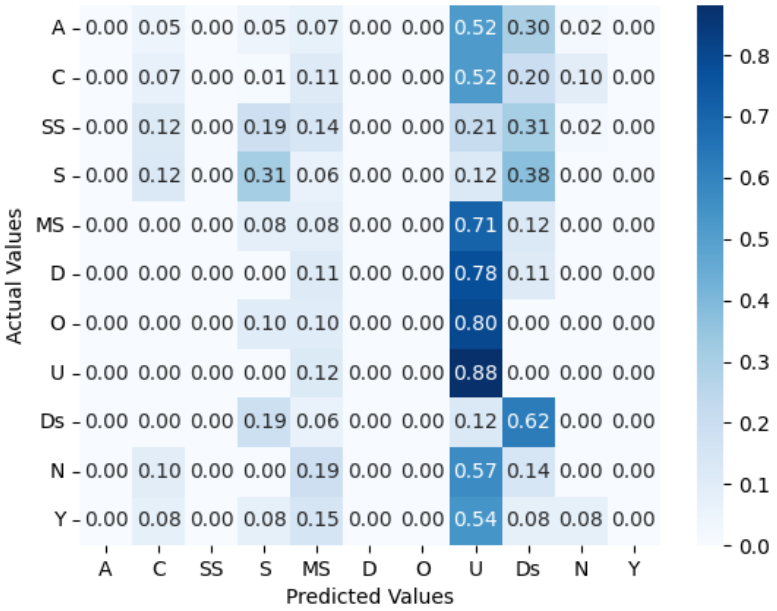


Figure H.2. – Confusion matrix for Model 2. Features = duration.

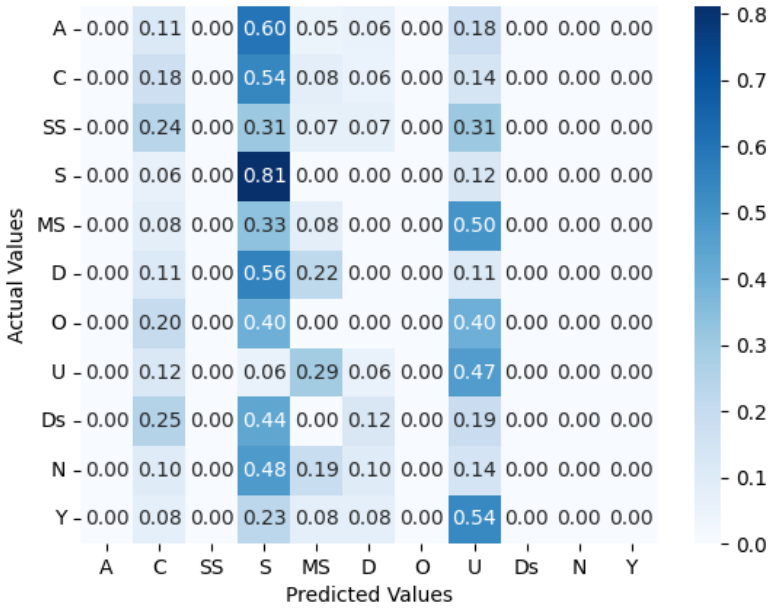


Figure H.3. – Confusion matrix for Model 3. Features = mean pitch.

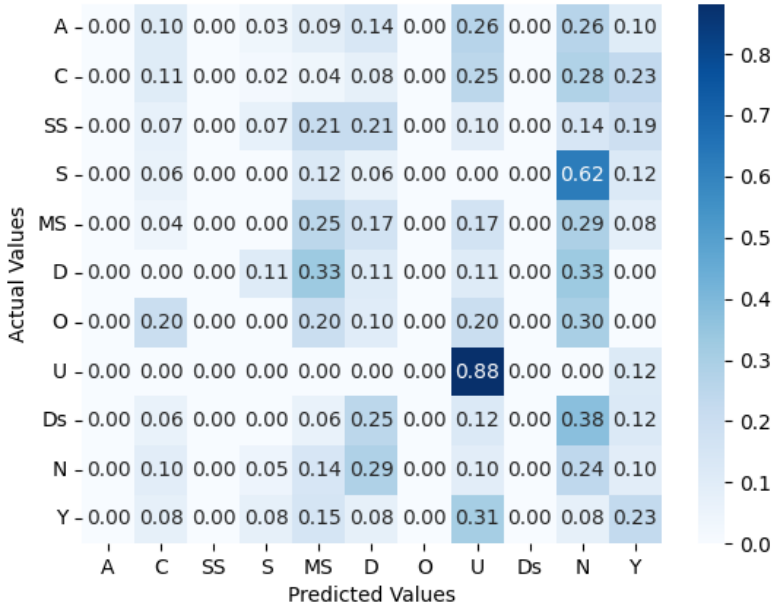


Figure H.4. – Confusion matrix for Model 4. Features = pitch slope.

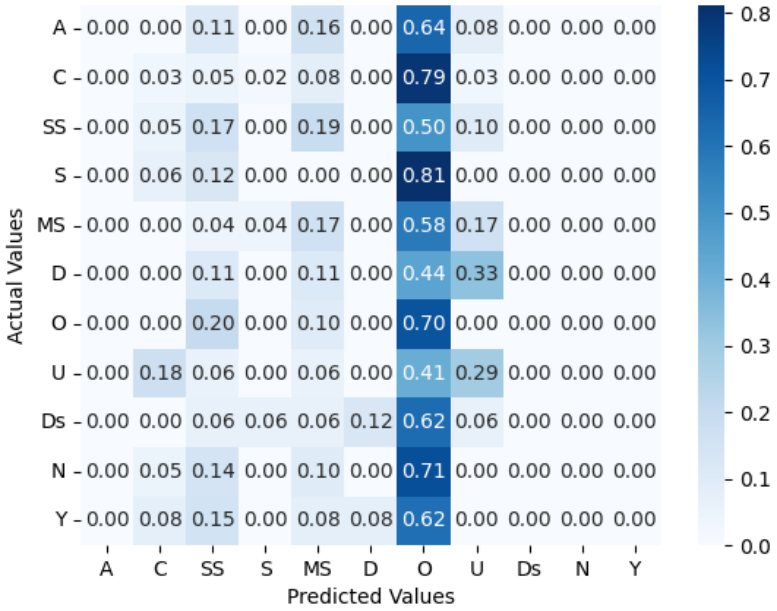


Figure H.5. – Confusion matrix for Model 5. Features = pitch range.

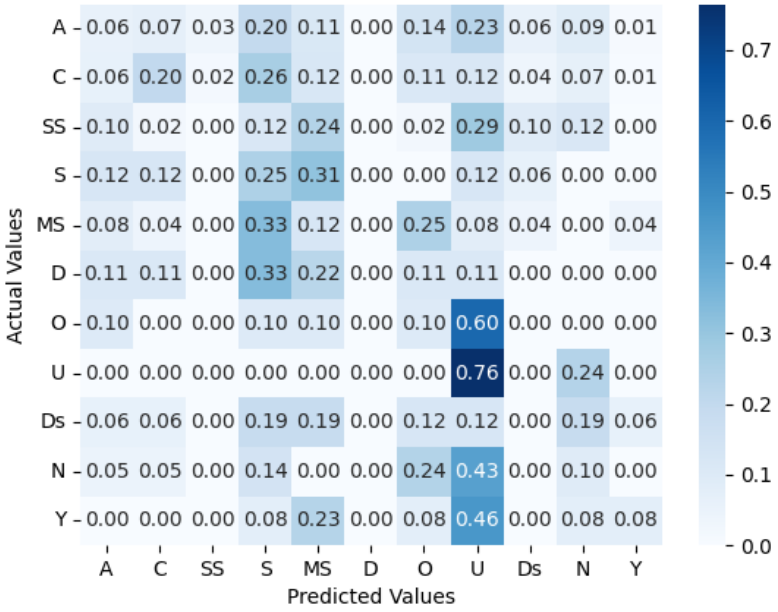


Figure H.6. – Confusion matrix for Model 6. Features = mean intensity.

H.3. Context features

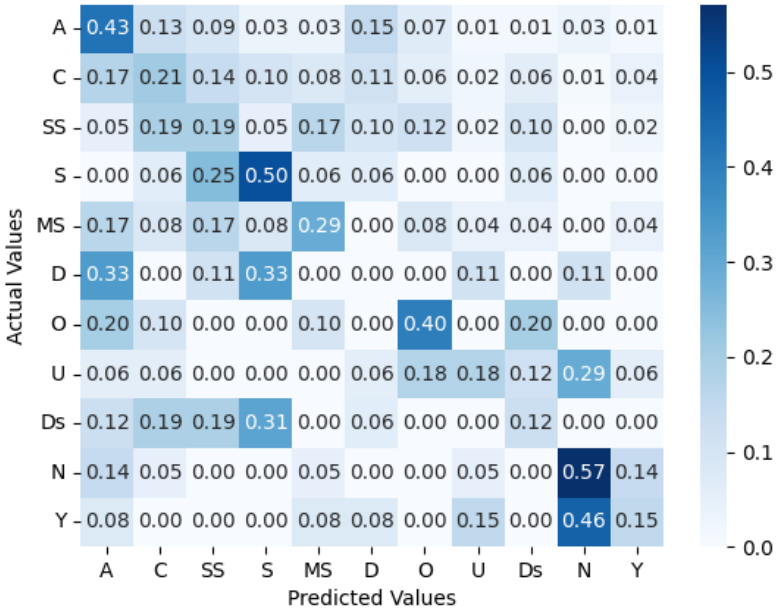


Figure H.7. – Confusion matrix for Model 7. Features = SimCSE embeddings.

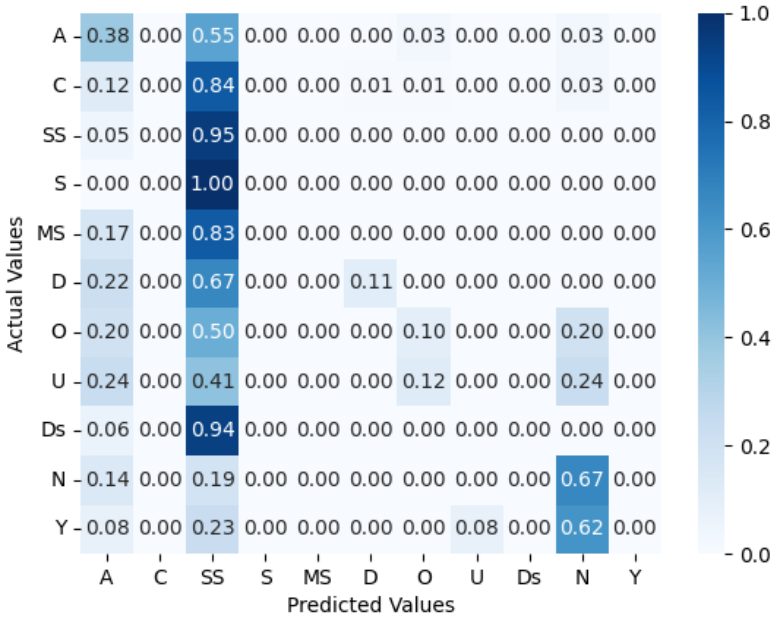


Figure H.8. – Confusion matrix for Model 8. Features = dialog act.

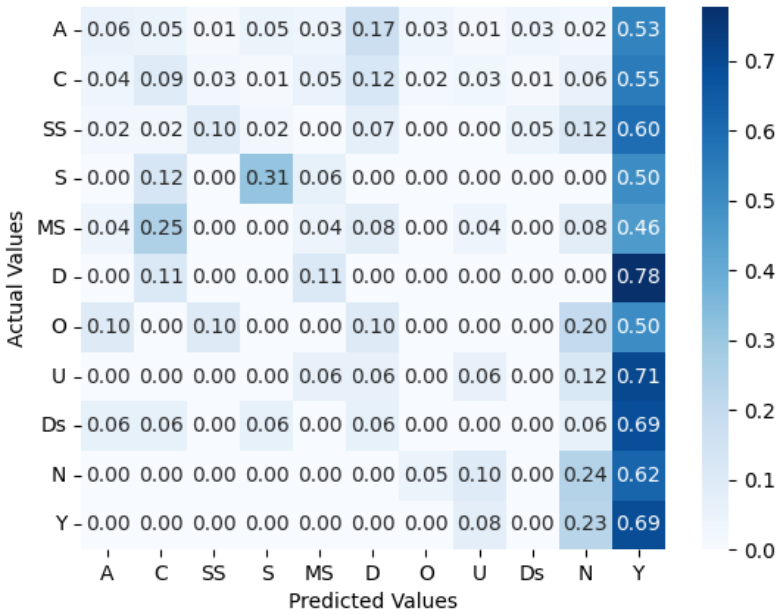


Figure H.9. – Confusion matrix for Model 9. Features = part-of-speech tags.

H.4. GPT-3

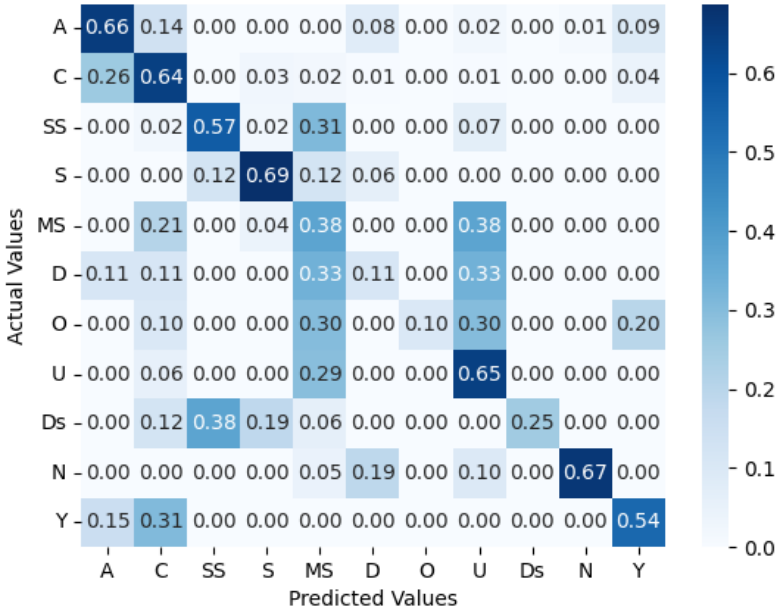


Figure H.10. – Confusion matrix for Model 10. GPT-3 used as the main classifier with zero-shot approach.

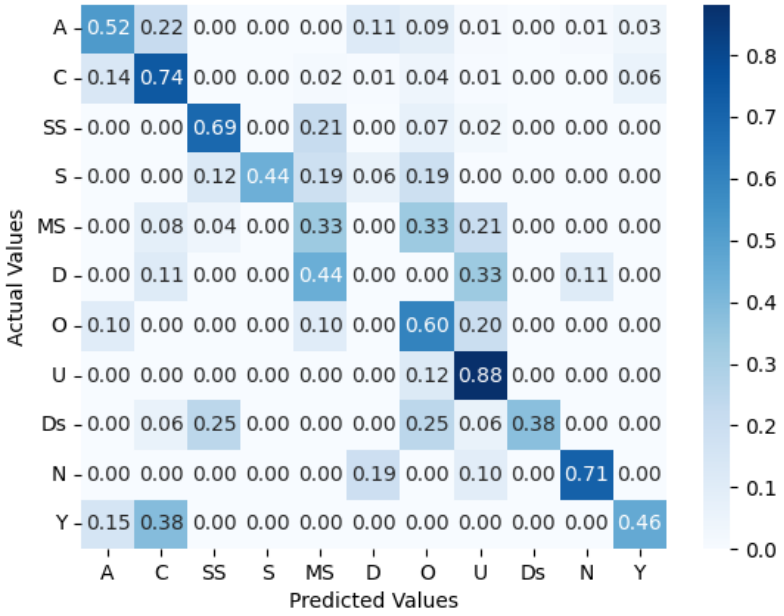


Figure H.11. – Confusion matrix for Model 11. GPT-3 used as the main classifier with few-shot approach.

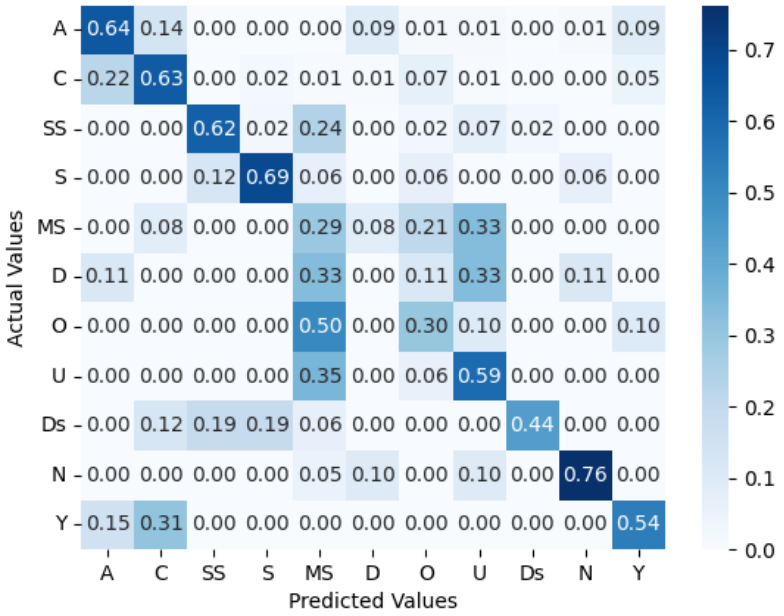


Figure H.12. – Confusion matrix for Model 13. Features = probability distributions from GPT-3 using zero-shot approach.

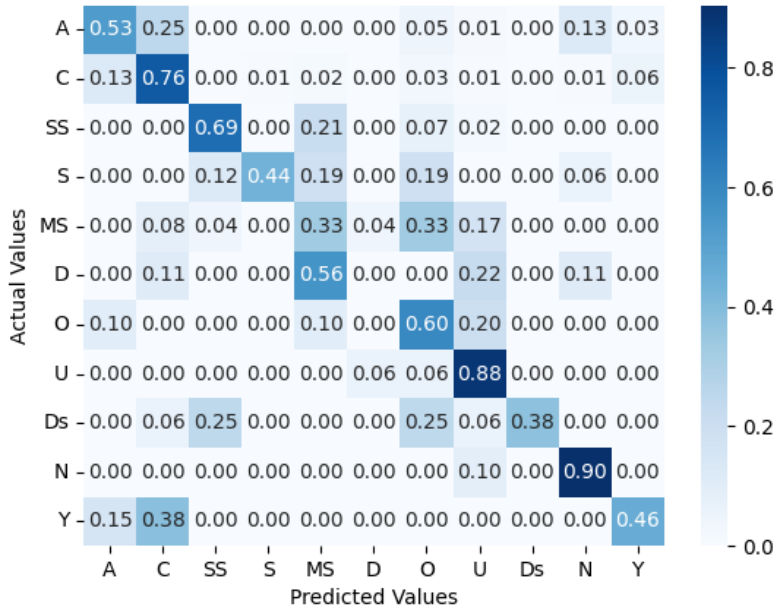


Figure H.13. – Confusion matrix for Model 14. Features = probability distributions from GPT-3 using few-shot approach.

H.5. Combinations

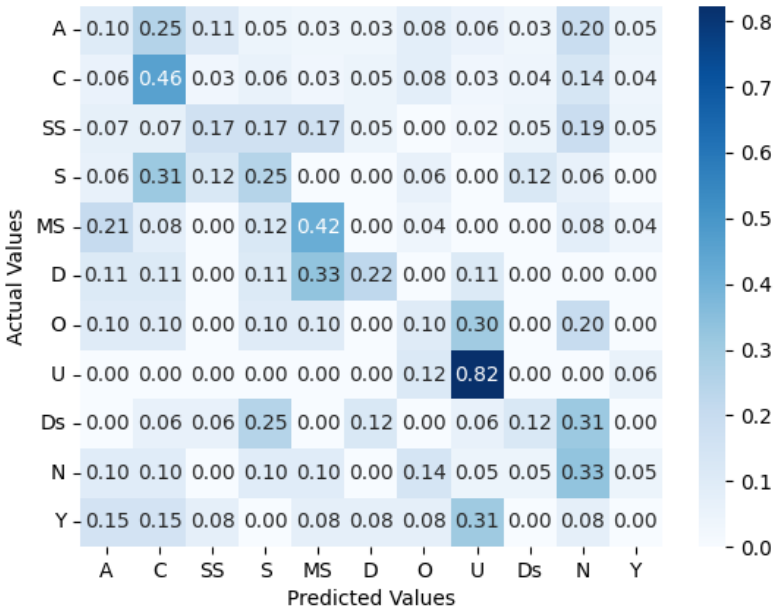


Figure H.14. – Confusion matrix for Model 15. Features = duration, mean pitch, pitch slope, pitch range, mean intensity.

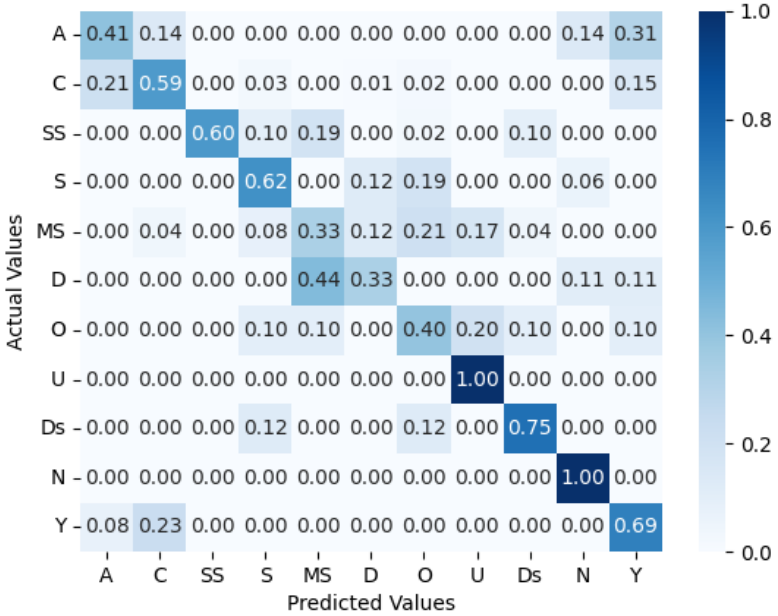


Figure H.15. – Confusion matrix for Model 16. Features = lexical, duration, mean pitch, pitch slope, pitch range, mean intensity.

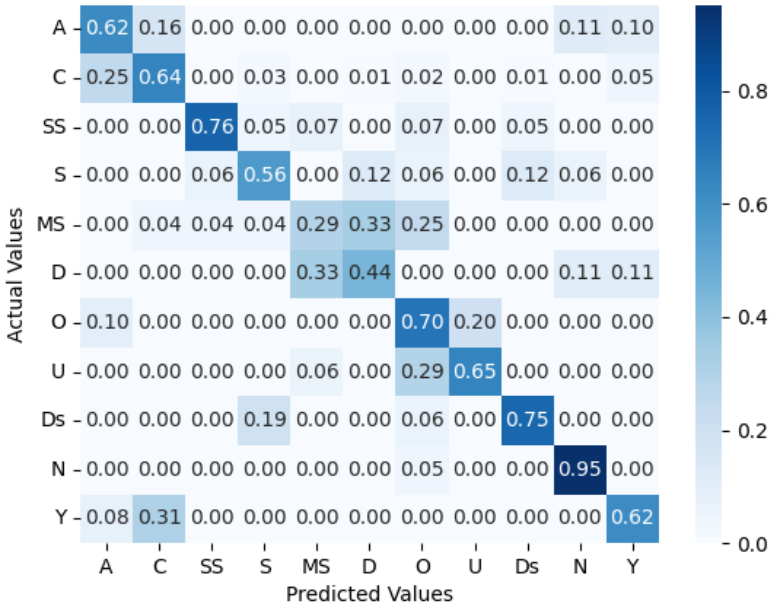


Figure H.16. – Confusion matrix for Model 17. Features = lexical, probability distributions from GPT-3 using zero-shot approach.

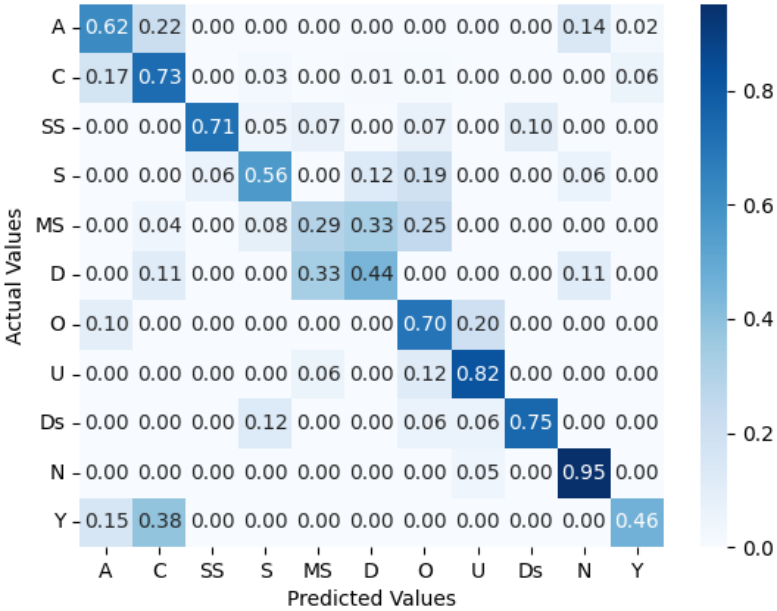


Figure H.17. – Confusion matrix for Model 18. Features = lexical, probability distributions from GPT-3 using few-shot approach.

I. Pearson's correlation coefficients

I.1. Mean Pitch Results

Function	r	p value	n	Significance
(C) Continue	0.002	0.943	1004	
(U) Non-understanding	0.348	0.006	61	**
(A) Agree	0.043	0.379	413	
(D) Disagree	-0.126	0.42	43	
(Y) Yes	0.217	0.105	57	
(N) No	-0.04	0.685	107	
(S) Sympathy	0.23	0.036	83	*
(Ds) Disapproval	-0.043	0.732	65	
(MS) Mild Surprise	0.105	0.301	99	
(SS) Strong Surprise	-0.083	0.259	186	

Table I.1. – Pearson's correlation coefficients between the mean pitch of feedback responses and preceding 500 ms utterance of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Function	**r**	**p value**	**n**	**Significance**
(C) Continue	0.039	0.211	1004	
(U) Non-understanding	0.234	0.069	61	
(A) Agree	0.107	0.03	413	*
(D) Disagree	-0.038	0.807	43	
(Y) Yes	0.238	0.074	57	
(N) No	0.234	0.015	107	*
(S) Sympathy	0.319	0.003	83	**
(Ds) Disapproval	0.049	0.701	65	
(MS) Mild Surprise	-0.075	0.462	99	
(SS) Strong Surprise	0.109	0.137	186	

Table I.2. – Pearson's correlation coefficients between the mean pitch of feedback responses and following 500 ms utterance of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Function	**r**	**p value**	**n**	**Significance**
(C) Continue	0.165	$p < .001$	1004	***
(U) Non-understanding	0.244	0.059	61	
(A) Agree	0.209	$p < .001$	413	***
(D) Disagree	0.219	0.159	43	
(Y) Yes	0.53	$p < .001$	57	***
(N) No	-0.014	0.884	107	
(S) Sympathy	0.161	0.146	83	
(Ds) Disapproval	0.064	0.613	65	
(MS) Mild Surprise	0.17	0.093	99	
(SS) Strong Surprise	0.108	0.143	186	

Table I.3. – Pearson's correlation coefficients between the mean pitch of the interlocutor's preceding and following 500 ms utterance. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

I.2. Mean Intensity

Function	r	p value	n	Significance
(C) Continue	0.138	$p < .001$	1004	***
(U) Non-understanding	0.172	0.185	61	
(A) Agree	0.21	$p < .001$	413	***
(D) Disagree	0.206	0.185	43	
(Y) Yes	0.107	0.43	57	
(N) No	0.168	0.084	107	
(S) Sympathy	0.296	0.007	83	**
(Ds) Disapproval	0.03	0.81	65	
(MS) Mild Surprise	0.268	0.007	99	**
(SS) Strong Surprise	0.139	0.059	186	

Table I.4. – Pearson's correlation coefficients between the mean intensity of feedback responses and preceding 500 ms utterance of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Function	r	p value	n	Significance
(C) Continue	0.132	$p < .001$	1004	***
(U) Non-understanding	0.158	0.224	61	
(A) Agree	0.003	0.952	413	
(D) Disagree	-0.274	0.075	43	
(Y) Yes	-0.042	0.757	57	
(N) No	-0.012	0.902	107	
(S) Sympathy	0.049	0.658	83	
(Ds) Disapproval	0.093	0.46	65	
(MS) Mild Surprise	0.154	0.129	99	
(SS) Strong Surprise	0.142	0.053	186	

Table I.5. – Pearson's correlation coefficients between the mean intensity of feedback responses and following 500 ms utterance of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Function	r	p value	n	Significance
(C) Continue	0.293	$p < .001$	1004	***
(U) Non-understanding	0.31	0.015	61	*
(A) Agree	0.177	$p < .001$	413	***
(D) Disagree	-0.037	0.814	43	
(Y) Yes	0.347	0.008	57	**
(N) No	0.04	0.681	107	
(S) Sympathy	0.168	0.128	83	
(Ds) Disapproval	0.187	0.136	65	
(MS) Mild Surprise	0.376	$p < .001$	99	***
(SS) Strong Surprise	0.222	0.002	186	**

Table I.6. – Pearson's correlation coefficients between the mean intensity of the interlocutor's preceding and following 500 ms utterance. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

I.3. Pitch Slope

Function	r	p value	n	Significance
(C) Continue	-0.023	0.464	1004	
(U) Non-understanding	0.248	0.054	61	
(A) Agree	-0.052	0.292	413	
(D) Disagree	-0.234	0.131	43	
(Y) Yes	0.041	0.762	57	
(N) No	0.08	0.411	107	
(S) Sympathy	-0.158	0.153	83	
(Ds) Disapproval	-0.113	0.371	65	
(MS) Mild Surprise	0.001	0.99	99	
(SS) Strong Surprise	0.007	0.928	186	

Table I.7. – Pearson's correlation coefficients between the pitch slope of feedback responses and preceding 500 ms utterance of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Function	r	p value	n	Significance
(C) Continue	-0.006	0.849	1004	
(U) Non-understanding	-0.1	0.444	61	
(A) Agree	0.022	0.66	413	
(D) Disagree	-0.103	0.513	43	
(Y) Yes	0.027	0.843	57	
(N) No	-0.158	0.105	107	
(S) Sympathy	-0.047	0.676	83	
(Ds) Disapproval	-0.038	0.764	65	
(MS) Mild Surprise	-0.089	0.382	99	
(SS) Strong Surprise	-0.024	0.744	186	

Table I.8. – Pearson's correlation coefficients between the pitch slope of feedback responses and following 500 ms utterance of the interlocutor. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Function	r	p value	n	Significance
(C) Continue	0.031	0.328	1004	
(U) Non-understanding	0.04	0.762	61	
(A) Agree	-0.017	0.735	413	
(D) Disagree	0.028	0.857	43	
(Y) Yes	-0.054	0.69	57	
(N) No	-0.011	0.912	107	
(S) Sympathy	-0.144	0.195	83	
(Ds) Disapproval	-0.109	0.386	65	
(MS) Mild Surprise	0.046	0.651	99	
(SS) Strong Surprise	-0.109	0.137	186	

Table I.9. – Pearson's correlation coefficients between the pitch slope of the interlocutor's preceding and following 500 ms utterance. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

J. Switchboard clip transcriptions

J.1. Clip transcriptions

The transcriptions of the 12 clips that the participants where asked to listen to. The short *feedback responses* by the listener are in bold within brackets.

J.1.1. Set 1

Clip 1 : Yeah yes um on the other hand you know I I had a similar had a similar health plan and uh one of my kids was in a car accident and and [**mm**] I had wound up having to pay for you know a bunch of doctor visits and stuff out of my pocket because of you know no no insurance policy happened to cover it which is

Clip 2: um but you know they're building the baseball stadium and they've got land set aside for a football stadium if they ever get a NFL team [**hm**] so it's um real easy access from from south of Baltimore like um you know like the airport or more importantly for the Orioles from Washington DC [**yeah**] because the Orioles say they get twenty percent of their population i mean uh their attendance from uh DC

Clip 3: well I mean just for me the mortgage to to get a mortgage on my house I mean they invest investigated me personally to the point where I was insulted [**yeah**] and I was putting $40,000 down on a $160,000 house [**yeah**] I mean I would have though goh we're happy to do it just sign here you know [**really**] I mean they had forty thousand dollars in

Clip 4: actually we um met some people that were in the naval base down there [**okay**] and uh they didn't particularly like living down there because it was very foreign very different the the people they they didn't treat them nice [**okay**] they you know um so I think there I what I learned from them there was a lot of resentment towards the Americans so and it was like they were they're Puerto Rican and were Americans [**right**] so that's why they're so um emotional about statehood yet like you say it's they can't really support themselves

Clip 5: but um they put up a nice fence so we still have a lot of privacy and we grow a lot of food [**uh-huh**] uh I enjoy it um the gardens are kind of old you have to step down in them now that the [**uh-huh**] we've tilled them so much but they're still we we my sister uses plenty of fertilizer I don't know if that's a good thing or a bad

thing

Clip 6: have you ever got to go back [**no no**]

Clip 7: there's a there's a race in Australia with solar powered cars [**ah**] and Ford and General

Clip 8: mhm and you know grace type waste that you mentioned we see often highlighted in the military and the defense department [**absolutely**] but it's uh I'm sure it's widespread to every agency

Clip 9: they have a a new waterfront uh marina in Philadelphia it isn't as developed as uh Water Side in Norfolk or the Baltimore uh waterfront but uh the marina is only about uh two or three blocks from the historic district [**oh**] so that's quite uh [**yeah**] handy for our our youngsters we can take them up and show them Independence Hall and the Liberty Bell and uh [**yeah**]

Clip 10: I know it I know it [**oh wow**] and it's almost like talking about the checkless society and and all of that and you know there was talk in fact my brother uh was with IBM from 1954 until about three years ago so we really had a family history of talking about development of uh of equipment [**wow**]

Clip 11: with alcohol [**pardon**] they do it with alcohol [**yeah**]

Clip 12: mercury on it or something [**ugh**] and uh to keep the because the corn gets treated to keep uh insect pests away [**uh-huh**] so so if you go in and you dig into the pheasant yeah you can get mercury poising but uh so there's sort of some risks to that actually uh let me think gun control

J.1.2. Set 2

Clip 1: oh they do have on site care [**no**]

Clip 2: uh we're trying to get my mother's go you know trying to get my mother's family going because my grandmother just died [**aww**] so if like uh well she's been dead a year now and before anybody else dies

Clip 3: yeah they fill the court there the jails up and suddenly let them go and they're back on the street we had this murder up here um Art Shawcross up in Rochester killed nineteen prostitutes up here [**oof**] and he was let out on parole from up in uh um Watertown

Clip 4: well I don't know what our next trip will be I guess our next well I know what my next trip I'm going to be a grandmother in July [**ooh**]

Clip 5: think realistically you know you can have your college loans delayed now because I had them delayed because I'm back in graduate school at thirty years old [**yes**] um I've had them delayed because I'm back in graduate school and on that form it says if your joining the Peace Corps you can have them delayed [**hm**] uh and I thought that was you know very interesting and I I would have thought of that earlier I probably would have done you know just like is that is this is that the Mormon church [**yes**] that does that

Clip 6: the front yard [**mhm**] and uh so when we left you know the back yard had um the saint um I think it was Saint Augustine that we had um it it had held onto a small portion but primarily once the weeds start in the back yeah we were just re you know resigned to well the only way we were going to fix this one is if you know if you plow it all under and [**mhm**] put everything back on top of it again [**hm**] but I don't know that's the bad thing there is that we spent so much money or you would spend so much money trying to keep a a large lawn alive the the only thing I didn't like about lawns and we were sitting there wondering there must be a better way to landscape so that you don't have to spend so much money trying to keep the lawn

Clip 7: having a kid is rough isn't it [**what**] from what I hear having a kid is rough

Clip 8: so you do not have any place that has a mop board off or a [**no**] a piece uh we have a friend who uh rents homes redoes homes and rents them [**uh-huh**] and he never quite has finished any one of the houses that he's done I mean there

Clip 9: yeah I think if there's any major piece of advice I'd give is to find a way of getting an education that doesn't incur that kind of debt [**yeah**] it't not i mean remember seeing an article one time about you know if the average person who spent that much money going to college just took the same amount of money and put it in a a in an investment fund they'd be considerably wealthier than they would be from the job they'd get after college [**exactly**] so it's it's really kind of crazy

Clip 10: I've never heard that that's very nice oh so I'm all for the metric system and converting over and I think I guess my feeling is the way to do it is to just start giving weights you know have a very brief transition period and then just start giving weights and kilometers or distance in kilometers and weights in kilograms and everything like that and uh just have people start using it rather than having people constantly trying to convert remember getting a package of something that said one pound this a package of dates mind you it's was presumable something you weigh fairly precisely it said one 1 pound and then in parenthesis it said 554.6 grams [**right right**] and as near as I could tell seeing that was basically anti-metric propaganda cause anyone who would say well look I can either buy 1 point of something at 464.6

grams which of course they couldn't weigh it out accurately anyway um every time I see something like that I think well that's that's an anti-metric argument [**yep**]

Clip 11: so um well I'll tell you my situation is that I have an elderly grandmother that we did just recently put in a nursing home and um her son which is my father is also elderly and this is one of the reasons why she had to go to the nursing home is that she was literally driving him nuts in his later years now my father's almost 80 and my grandmother's almost 97 [**jeez**] so um it's strange because it it so hit so close to home but um um my father's an only child and really me and my sister are the only ones that will deal with my grandmother she had many sisters and a couple of them took care of her and then one her last sister died and it was probably 7 or 8 months after that she had to go in a nursing home because I was pretty much giving up my life my sister was and plus she was driving my father crazy she went through three housekeepers live-n housekeepers so she's kind of a cranky to get along with there's nothing physically wrong with her except she's very very old but her personality is is very grating I mean I hope I don't get like that when I get old [**yes**]

Clip 12: oh yeah I I think it's a wonderful thing to do and there's a lot I think there's a lot more I guess another possible solution is since taxpayers aren't going to start paying more money for this and and other budgets aren't going to be cut to pay for it [**no**] um more of the volunteer network service because everyone gains from it [**mhm**] would be would might be really useful um and if it's you know uh just people helping people I think make makes the community so much happier

K. Screenshots of listening test

Note each clip was presented one by one to the participants. The instructions, the clip, the description of the communicative functions, along with the questions asking the participants to assign the communicative function, rate the naturalness and rate the appropriateness were presented in a single page.

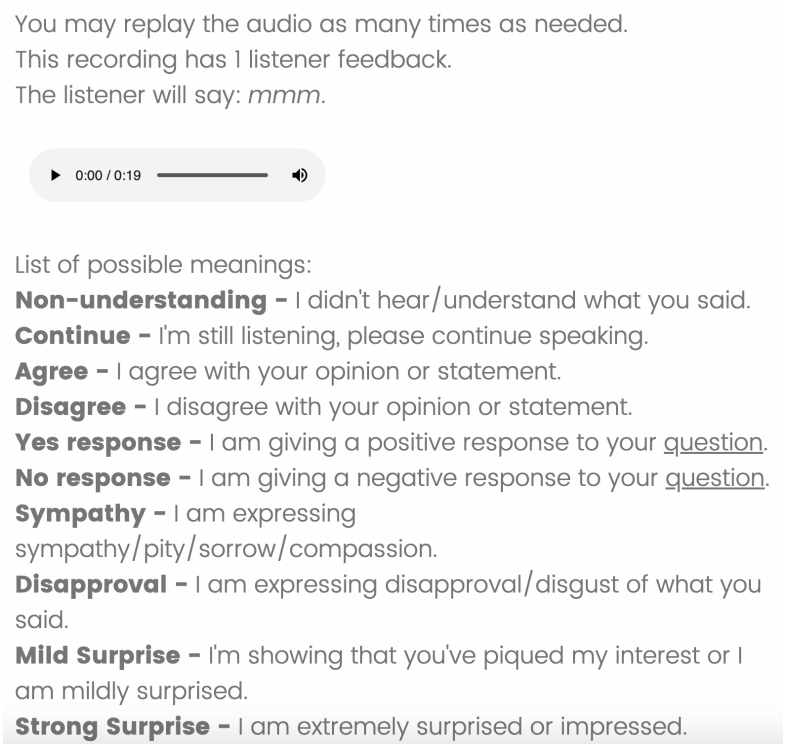


Figure K.1. – Screenshot of the clip presentation and the list of communicative functions with their corresponding descriptions.



Figure K.2. – Screenshot of the question asking the participant to select the communicative function.

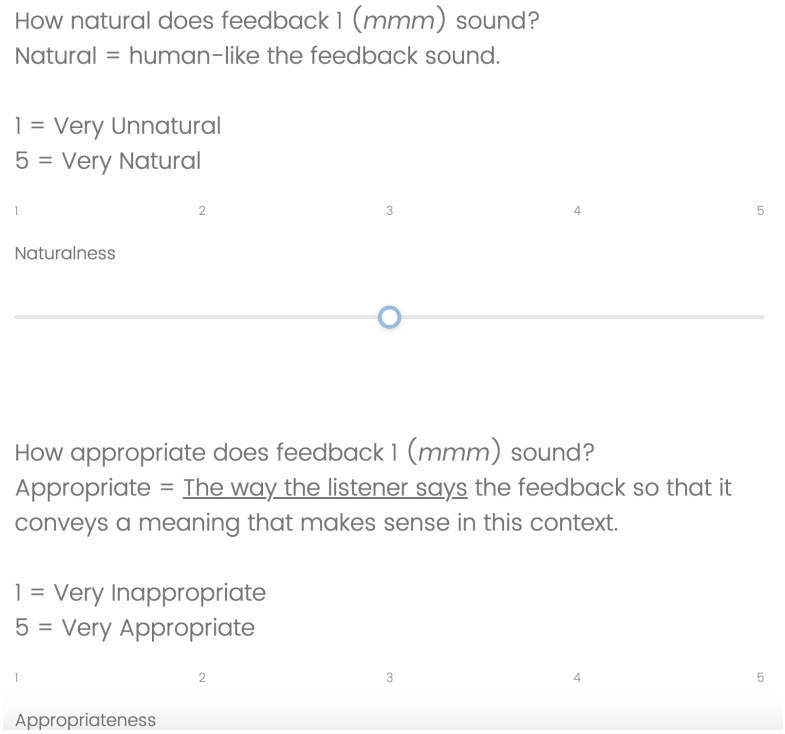


Figure K.3. – Screenshot of the question asking the participant to rate the naturalness of the feedback response on a Likert scale.

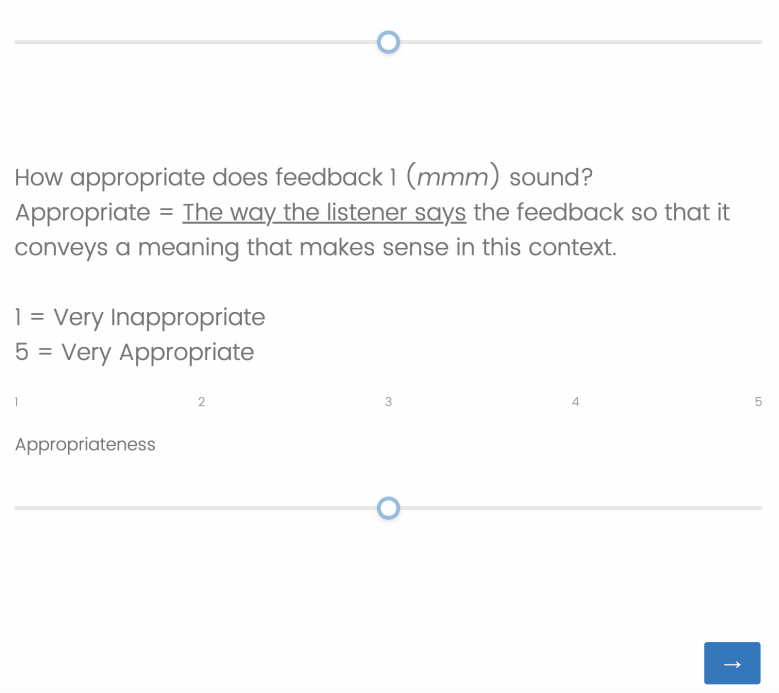


Figure K.4. – Screenshot of the question asking the participant to rate the appropriateness of the prosody of the feedback response on a Likert scale.

L. Confusion Matrices of intra-annotator agreement

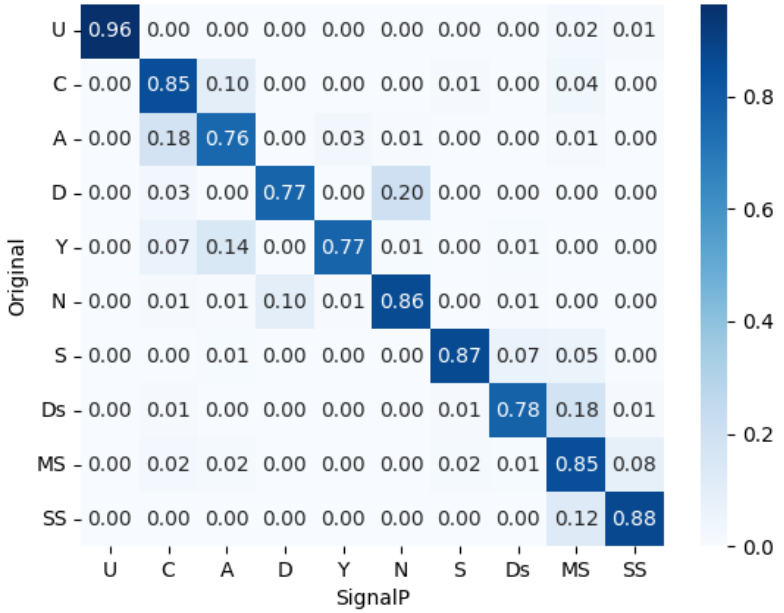


Figure L.1. – Participants' perception of the original Switchboard feedback (Original) compared to feedback synthesized by signal processing approach (SignalP).

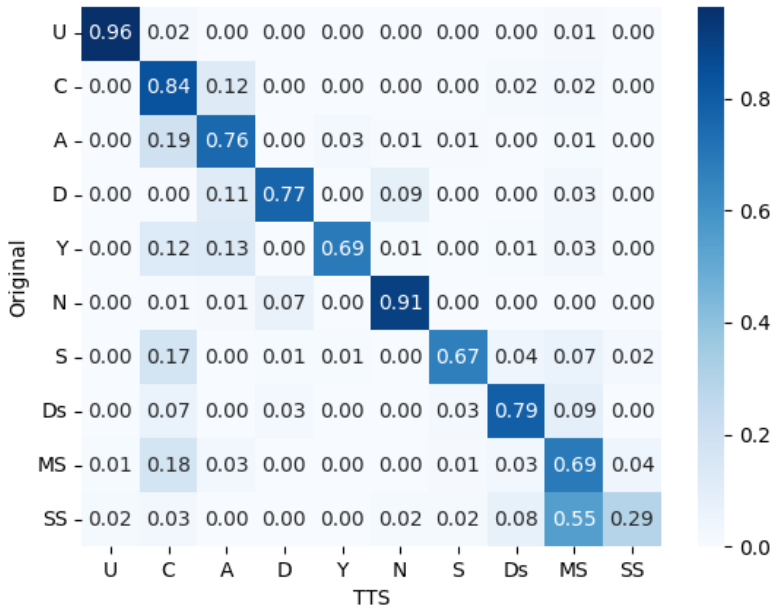


Figure L.2. – Participants' perception of the original Switchboard feedback (Original) compared to feedback synthesized by TTS approach (TTS).

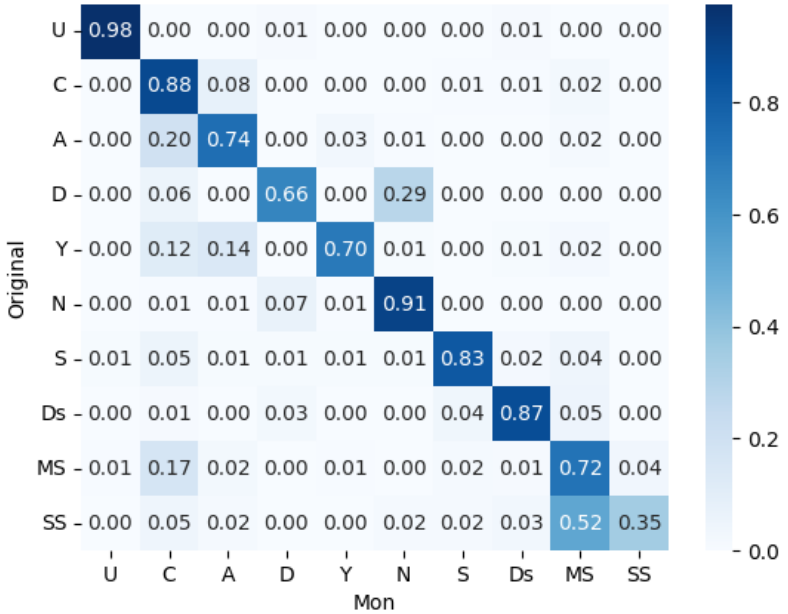


Figure L.3. – Participants' perception of the original Switchboard feedback (Original) compared to the monotone (Mon) feedback.