# Mhm... Yeah? Okay! Evaluating the Naturalness and Communicative Function of Synthesized Feedback Responses in Spoken Dialogue

**Carol Figueroa[1, 2], Marcel de Korte[3], Magalie Ochs[1], Gabriel Skantze[2,4],**

[1]Aix-Marseille University, [2]Furhat Robotics, [3]Constructor Technology,
[4]KTH Royal Institute of Technology

carol.figueroa@etu.univ-amu.fr, marcel.korte@constructor.tech,
magalie.ochs@lis-lab.fr, skantze@kth.se

## Abstract

To create conversational systems with human-like listener behavior, generating short feedback responses (e.g., "mhm", "ah", "wow") appropriate for their context is crucial. These responses convey their communicative function through their lexical form and their prosodic realization. In this paper, we transplant the prosody of feedback responses from human-human U.S. English telephone conversations to a target speaker using two synthesis techniques (TTS and signal processing). Our evaluation focuses on perceived naturalness, contextual appropriateness and preservation of communicative function. Results indicate TTS-generated feedback were perceived as more natural than signal-processing-based feedback, with no significant difference in appropriateness. However, the TTS did not consistently convey the communicative function of the original feedback.

## 1 Introduction

In dyadic human-human conversations, interlocutors often take turns listening and speaking. However, while one interlocutor speaks, the listener doesn't remain silent; instead, they give short feedback responses like "uh-huh", "yeah" and "wow". Although these responses are known by different names (e.g., *backchannels* (Yngve, 1970), *continuers* (Schegloff, 1982), *assessments* (Goodwin, 1986)), we follow Allwood et al. (1992) in adopting the term *feedback*, since it encompasses the many communicative functions of these short responses. Feedback responses are crucial for smooth turn-taking and establishing common ground, i.e., people's mutual knowledge or beliefs (Clark, 1996). If the listener hasn't understood or heard what was said, they might say "huh?", "sorry?", or "what?", prompting the speaker to clarify. Other responses, such as "mhm", can be used to unobtrusively signal the speaker to continue. The communicative functions of feedback are conveyed through both their lexical form and prosody, with prosody sometimes being the most important. For example, "yeah" can express agreement, disagreement or surprise depending on its prosodic realization.

Incorporating feedback in spoken dialogue systems for conversational agents is an active research area (Axelsson et al., 2022). Many studies have focused on predicting the timing of backchannels (Adiba et al., 2021a,b; Wang et al., 2024), while others have focused on predicting their communicative function (Boudin et al., 2021; Lala et al., 2022; Choi et al., 2024).

Previous studies have used signal processing to manipulate prosodic features to understand how these affect the perceived communicative functions of synthesized feedback (Åsa Wallers, 2006; Stocksmeier et al., 2007; Chandler, 2023). Short feedback responses have been incorporated into unit selection text-to-speech (TTS) synthesis systems by treating entire responses as units rather than concatenating diphones or phones (Campbell, 2007; Pammi et al., 2010). Further, Oertel et al. (2016) used statistical parametric speech synthesis for feedback responses. Recently, Mitsui et al. (2023) introduced a TTS system that can synthesize feedback without transcriptions.

Despite these efforts, there has been little focus on predicting the prosodic features of feedback or evaluating their contextual appropriateness. Nath and Ward (2022) predicted prosodic features of discourse markers, which are lexically similar to many feedback responses, at the token level, but suggested future work should focus on the frame level. When it comes to evaluation, most studies (on TTS in general) have primarily focused on whether the speech sounds natural, and less on whether the intended communicative function is conveyed.

In this paper, we investigate to what extent feedback responses can be synthesized, using existing synthesis methods, so that they sound natu-

*Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 544–553
September 18–20, 2024. ©2024 Association for Computational Linguistics

ral and appropriate in their context, while at the same time conveying their intended communicative function. Thus, our research question is not how to predict the prosodic and lexical features of feedback responses, but whether it is possible to synthesize them, given that we could make those predictions. To investigate this, we re-synthesize feedback responses in human-human U.S. English telephone conversations by transplanting their original prosody. We use two synthesis methods: (1) signal processing and (2) text-to-speech, which both have different advantages. Signal processing allows for more fine-grained control of prosody (compared to the TTS used here) but can degrade audio quality and introduce artifacts, while TTS tends to sound more natural. In our listening tests, we let participants listen to these synthesized feedback responses in their dialogue context, and ask participants to rate their naturalness and appropriateness, as well as to assign the most likely communicative function. For comparison, we also let them rate the original feedback responses, as well as a re-synthesized monotone version, where signal processing is used to flatten the pitch and thus to remove intonation. To the best of our knowledge, this is the first work evaluating the appropriateness of the prosody of synthesized feedback responses in context.

## 2 Method

To manipulate the prosodic features of feedback responses, we use two synthesis methods: a signal processing and a TTS approach. We transplant the prosody of feedback responses from "listeners" in the U.S. English Switchboard corpus (Godfrey et al., 1992) – referred to as our reference speakers – onto our target voice, a female voice talent.

### 2.1 Signal processing

Using signal processing, the prosody of the original feedback response (as it appeared in the Switchboard conversation) is transplanted to a *feedback template*, which is recorded from the voice talent. Thus, we recorded one feedback template per lexical form (e.g., "yeah", "mhm"). To transplant the prosody of the original feedback onto the template, we first used the Montreal forced aligner (McAuliffe et al., 2017) to obtain the phone-level durations of the original feedback 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. 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 of the target speaker. We z-score normalized the Switchboard speaker's $F_0$ values per speaker and then denormalized these z-score values using the voice talent's mean $F_0$ and standard deviation, after which we re-synthesized the audio with the new $F_0$ values. Finally, we transplanted the intensity contour of the original feedback to the feedback template using the Praat Vocal Toolkit (Corretge, 2024).

### 2.2 Text-to-speech

For TTS, we use FastPitch 1.1 (Łańcucki, 2021) for the acoustic model and HiFiGAN (Kong et al., 2020) as the vocoder model. Although FastPitch is a deterministic model, i.e., it generates the same prosodic realization for the same 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.

To transplant the prosody of the original feedback onto the synthesized feedback, we replaced the predicted prosodic features with the original ones during inference. We used the durations from the phone-level alignments of the original feedback. The $F_0$ values were extracted with Praat at the frame-level, averaged per phone, z-normalized and then de-normalized with the previously outlined procedure. We used the energy extraction method from FastPitch to extract energy values of the original feedback.

Most TTS voices are trained on read speech and therefore exclude short feedback responses. Since we aimed to train a conversational voice and capture as much prosodic variation as possible, for the TTS training data, the voice talent recorded different types of speech: 1) "read speach", 43 minutes were recorded from the CMU ARCTIC database (Kominek and Black, 2004); 2) "role-play acted speech", 4 minutes were recorded from the Taskmaster-2 dataset (Byrne et al., 2019); 3) "feedback imitations" 724 feedback responses were imitated from Switchboard amounting to 11 minutes; 4) "conversational speech", 34 minutes of speech were recorded from the voice talent while chatting with people. 981 instances of feedback were captured. All audio was recorded at 48 kHz and then downsampled to 22 050 Hz for training.

The base acoustic model was trained on LJSpeech (Ito and Johnson, 2017) 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 for a further 500 epochs with the same hyperparameters as in pre-training, using a 97-3% train-validation split. We also fine-tuned a pre-trained HiFiGAN universal vocoder on our target voice for 58000 steps, using a batch size of 16, learning rate of $1e-5$, and the same train-validation split as for the acoustic model.

## 3 Experimental design

### 3.1 Participants

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

### 3.2 Stimuli

We used Qualtrics (qua, 2002) to host our online listening tests. Participants listened to 12 distinct clips of Switchboard conversations and were assigned to either set 1 or set 2 (see Appendix A). Each clip featured one speaker and one listener, with the listener producing feedback responses that could either 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 conversations, where the feedback responses were either: 1) the original ones, 2) synthesized by signal processing, 3) synthesized by TTS, or 4) flattened to a monotone pitch. Note that only the feedback responses were replaced, not the Switchboard speaker channel. All conditions were randomized, presented one by one, and the same conversation was never presented consecutively. Samples of the clips can be found at `https://carolfigphd.github.io/SigDial2024_feedback_synthesis_samples/`.

### 3.3 Participants' tasks

Participants were asked to assign a function from the 10 communicative functions of feedback in Figueroa et al. (2022): Non-understanding (U), Continue (C), Agree (A), Disagree (D), Yes response (Y), No response (N), Sympathy (S), Disapproval (Ds), Mild Surprise (MS), and Strong Surprise (SS). Participants were also asked to rate the naturalness and appropriateness of the prosody of the feedback responses on a Likert scale 1-5 where (1=Very Unnatural, 5=Very Natural) and (1=Very Inappropriate, 5= Very Appropriate). Naturalness was defined as how human-like the feedback response was; participants were told beforehand that feedback responses were either machine- 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 B.

### 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 ID) as random effects. The following formula was used for our condition model: $\mathrm{clmm}(\textit{naturalness/appropriateness} \sim \textit{method} + (1|\textit{subjectID}) + (1|\textit{stimuliID}))$. We used an ANOVA to compare our condition model to a null model $\mathrm{clmm}(\textit{naturalness/appropriateness} \sim (1|\textit{subjectID}) + (1|\textit{stimuliID}))$.

## 4 Results and discussion

### 4.1 Naturalness

Figure 1 shows the distribution of the ratings for naturalness and mean $\mu$ and standard deviation $\sigma$ for each condition: Monotone (Mon), text-to-speech (TTS), signal 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 synthesized by the TTS was perceived as more natural than the feedback synthesized by signal processing and the monotone feedback. This was expected as signal processing degrades the au-

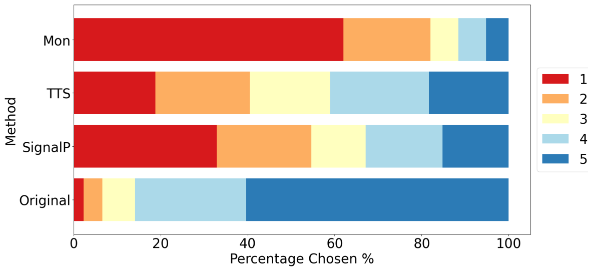


Figure 1: Naturalness rating distribution per condition. Mon ($\mu$=1.73, $\sigma$=1.15), TTS ($\mu$=3.0, $\sigma$=1.39), SignalP ($\mu$=2.61, $\sigma$=1.47), Original ($\mu$=4.37, $\sigma$=0.96). 1= Very Unnatural, 5= Very Natural.

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

Table 1: Cohen's kappa coefficient scores per comparison of intra-annotator agreement.

dio. Also, as expected, the original Switchboard feedback was rated to be more natural than all conditions, yet not all feedback 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.

### 4.2 Appropriateness

Figure 2 shows the distribution of the rating for appropriateness and mean $\mu$ and standard deviation $\sigma$ 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, 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 make it uncertain whether participants could disentangle naturalness and appropriateness.

### 4.3 Perception of communicative functions

To evaluate whether the synthesis methods preserve the communicative function of the original Switchboard feedback, 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 resynthesized counterpart (see Table 1). Confusion matrices for each comparison can be found in Appendix C.

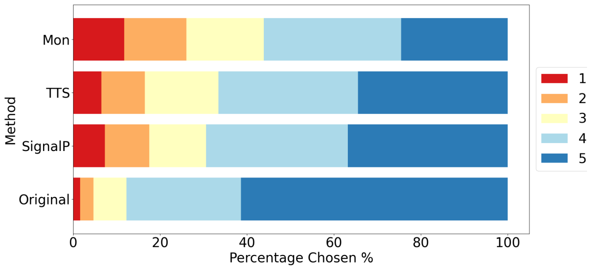


Figure 2: Appropriateness rating distribution per condition. Mon ($\mu$=3.43, $\sigma$=1.31), TTS ($\mu$=3.78, $\sigma$=1.20), SignalP ($\mu$=3.81, $\sigma$=1.24), Original ($\mu$=4.43, $\sigma$=0.88). 1= Very Inappropriate and 5=Very Appropriate.

The results show that the perceived communicative function is best preserved by the signal processing approach, which is expected since it transplants prosody at the frame level. Although signal processing and TTS feedback convey equally appropriate prosody for their context, the feedback synthesized by the TTS approach was not good for 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 was Strong Surprise, but the participants perceived the TTS feedback as Mild Surprise, both functions and prosodic realizations are appropriate for the context but are 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 is preserved. Thus, the participants likely mainly relied on the lexical form in those conditions.

## 5 Conclusion and future work

This paper investigated to what extent existing synthesis methods (signal processing and TTS) can produce feedback that sound natural and appropriate, while at the same time conveying the various communicative functions of feedback responses. We found that the TTS produced the most natural sounding feedback, but that both synthesis methods produced feedback that were deemed to be equally appropriate, given the context. However, we find that the TTS method fails to convey the intended communicative function, beyond the lexical form, while the signal processing method does provide additional prosodic information, most likely due to the more fine-grained prosodic control.

The implication of these findings are that, if we were to build a model that predicts the prosodic features of feedback, it may be beneficial to predict these features at the frame-level because the frame-level signal processing best preserves the intended communicative function. Such a prediction model could for example extend Corkey et al. (2023), in which an external predictor was trained to predict intonation.

## 6 Limitations

One limitation of this study is that only 47 feedback responses were evaluated, which did not cover all the possible lexical forms found in Switchboard. A second limitation is the within-participant experimental design; meaning that participants were presented with the same clips for all conditions. However, we chose this experimental design because we were interested in each individual participant's perception of the communicative functions, which can vary from person to person. The within-participant design allowed us to treat the original feedback responses from the Switchboard conversations as true labels. Furthermore, our results highlight that a better explanation of prosody to participants may help obtain more precise appropriateness of prosody ratings. For example, defining prosody as a combination of intonation, rhythm and tone may be a better way to ask about appropriateness of prosody.

## Acknowledgments

This work was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 859588. We would like to thank Esther Klabbers, Morgan Bell, and Louis Simeonidis for their participation in the pre-pilot listening test and for their valuable suggestions for modifications to the pilot test.

## References

2002. Qualtrics. https://www.qualtrics.com/.

2014. Prolific. https://www.prolific.com/.

Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, Takashi Sumiyoshi, and Kenji Nagamatsu. 2021a. Delay mitigation for backchannel prediction in spoken dialog system. *Conversational Dialogue Systems for the Next Decade*, pages 129–143.

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

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

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

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A multimodal model for predicting conversational feedbacks. In *International Conference on Text, Speech, and Dialogue*, pages 537–549. Springer.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proc. EMNLP-IJCNLP*, pages 4516–4525.

Nick Campbell. 2007. Towards conversational speech synthesis; lessons learned from the expressive speech processing project. *SSW*, 2207:22–27.

Ryan Lee Chandler. 2023. *Semantic-prosodic correlates of backchannel utterances in human-computer dialog*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Yong-Seok Choi, Jeong-Uk Bang, and Seung Hi Kim. 2024. Joint streaming model for backchannel prediction and automatic speech recognition. *ETRI Journal*.

Rune H. B. Christensen. 2023. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Niamh Corkey, Johannah O'Mahony, and Simon King. 2023. Intonation control for neural text-to-speech synthesis with polynomial models of F0. In *Proc. Interspeech 2023*, pages 2014–2015.

Ramon Corretge. 2024. Praat Vocal Toolkit. https://www.praatvocaltoolkit.com.

Tuan Dinh, Alexander Kain, and Kris Tjaden. 2019. Using a manifold vocoder for spectral voice and style conversion. In *Proc. Interspeech 2019*, pages 1388–1392.

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, and Gabriel Skantze. 2022. Annotation of communicative functions of short feedback tokens in Switchboard. In *Proc. LREC*, pages 1849–1859.

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

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9(2-3):205–217.

Keith Ito and Linda Johnson. 2017. The LJ speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

John Kominek and Alan W Black. 2004. The CMU Arctic speech databases. In *Fifth ISCA Workshop on Speech Synthesis*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

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

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Kentaro Mitsui, Yukiya Hono, and Kei Sawada. 2023. Towards human-like spoken dialogue generation between ai agents from written dialogue. *Preprint*, arXiv:2310.01088.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884.

Anindita Nath and Nigel Ward. 2022. On the predictability of the prosody of dialog markers from the prosody of the local context. In *Proc. Speech Prosody*, pages 664–668.

Catharine Oertel, Joakim Gustafson, and Alan W Black. 2016. 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*, pages 43–47.

Sathish Pammi, Marc Schröder, Marcela Charfuelan, Oytun Türk, and Ingmar Steiner. 2010. Synthesis of listener vocalisations with imposed intonation contours. In *Proc. 7th ISCA Workshop on Speech Synthesis (SSW 7)*, pages 240–245.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Thorsten Stocksmeier, Stefan Kopp, and Dafydd Gibbon. 2007. Synthesis of prosodic attitudinal variants in German backchannel ja. In *Proc. Interspeech 2007*, pages 1290–1293.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. *arXiv preprint arXiv:2401.14717*.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the 6th Regional Meeting Chicago Linguistic Society*, pages 567–578.

Åsa Wallers. 2006. Minor sounds of major importance-prosodic manipulation of synthetic backchannels in Swedish.

## A 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.

**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

**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

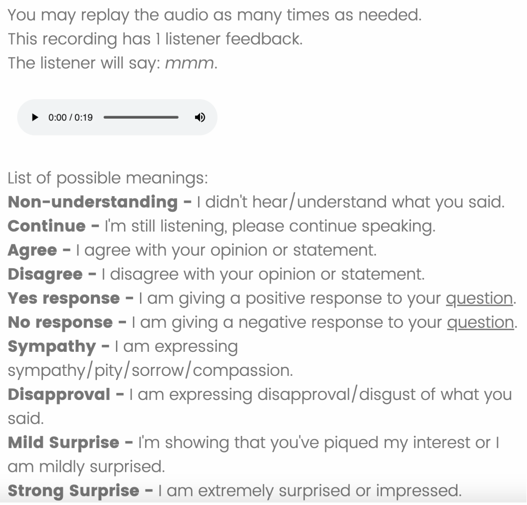
Clip 10: I've never heard that that's very nice oh so I'm all for the metric system and converting over and I think I guess my feeling is the way to do it is to just start giving weights you know have a very brief transition period and then just start giving weights and kilometers or distance in kilometers 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

## B Screenshots of listening test

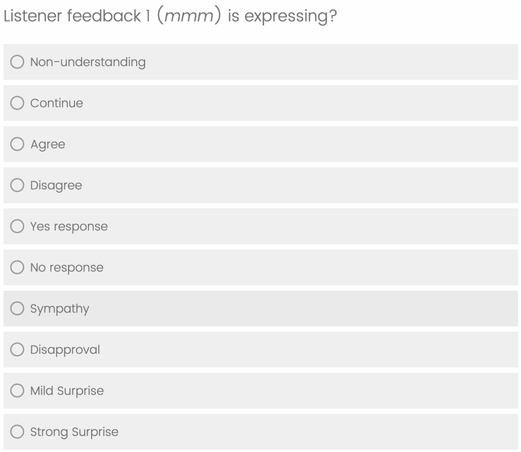
Note each clip was presented one by one to the participants and the questions about the communicative function, naturalness rating, appropriateness rating were presented in a single page.



You may replay the audio as many times as needed.
This recording has 1 listener feedback.
The listener will say: *mmm*.

List of possible meanings:
**Non-understanding -** I didn't hear/understand what you said.
**Continue -** I'm still listening, please continue speaking.
**Agree -** I agree with your opinion or statement.
**Disagree -** I disagree with your opinion or statement.
**Yes response -** I am giving a positive response to your question.
**No response -** I am giving a negative response to your question.
**Sympathy -** I am expressing sympathy/pity/sorrow/compassion.
**Disapproval -** I am expressing disapproval/disgust of what you said.
**Mild Surprise -** I'm showing that you've piqued my interest or I am mildly surprised.
**Strong Surprise -** I am extremely surprised or impressed.

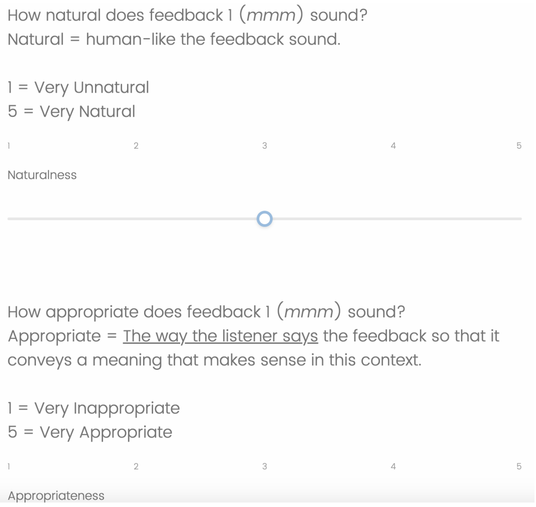
Figure 3: Screenshot of clip presentation and communicative function list.



Listener feedback 1 (*mmm*) is expressing?

○ Non-understanding
○ Continue
○ Agree
○ Disagree
○ Yes response
○ No response
○ Sympathy
○ Disapproval
○ Mild Surprise
○ Strong Surprise

Figure 4: Screenshot of question asking for communicative function.



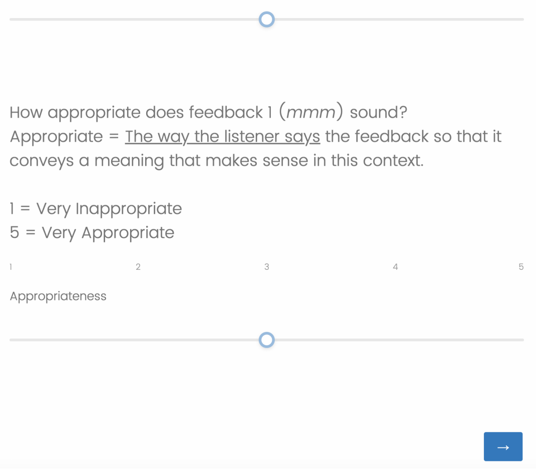
How natural does feedback 1 (*mmm*) sound?
Natural = human-like the feedback sound.

1 = Very Unnatural
5 = Very Natural

How appropriate does feedback 1 (*mmm*) sound?
Appropriate = The way the listener says the feedback so that it conveys a meaning that makes sense in this context.

1 = Very Inappropriate
5 = Very Appropriate

Figure 5: Screenshot of question asking for naturalness rating.



How appropriate does feedback 1 (*mmm*) sound?
Appropriate = The way the listener says the feedback so that it conveys a meaning that makes sense in this context.

1 = Very Inappropriate
5 = Very Appropriate

Figure 6: Screenshot of question asking for appropriateness rating.

## C Confusion Matrices of intra-annotator agreement

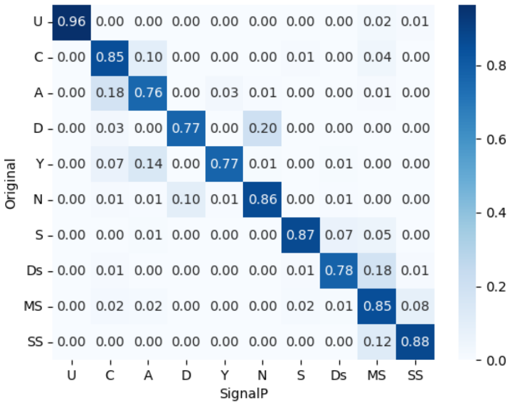


Figure 7: Participants' perception of the original Switchboard feedback (Original) compared to feedback synthesized by signal processing approach (SignalP).

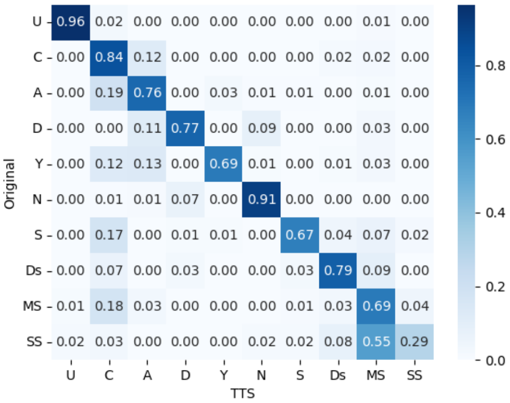


Figure 8: Participants' perception of the original Switchboard feedback (Original) compared to feedback synthesized by TTS approach.

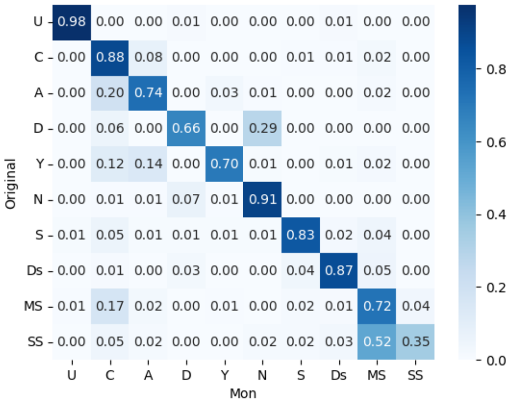


Figure 9: Participants' perception of the original Switchboard feedback (Original) compared to the monotone (Mon) feedback.