INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Crowd-Sourced Design of Artificial Attentive Listeners

Catharine Oertel[1], *Patrik Jonell*[1], *Dimosthenis Kontogiorgos*[1], *Joseph Mendelson*[1], *Jonas Beskow*[1] , *Joakim Gustafson*[1]

[1]KTH Royal Institute of Technology, Sweden

catha@kth.se, pjjonell@kth.se, diko@kth.se, josephme@kth.se, beskow@kth.se, jocke@speech.kth.se

Abstract

Feedback generation is an important component of human-human communication. Humans can choose to signal support, understanding, agreement or also sceptiscism by means of feedback tokens. Many studies have focused on the timing of feedback behaviours. In the current study, however, we keep the timing constant and instead focus on the lexical form and prosody of feedback tokens as well as their sequential patterns.

For this we crowdsourced participant's feedback behaviour in identical interactional contexts in order to model a virtual agent that is able to provide feedback as an attentive/supportive as well as attentive/sceptical listener. The resulting models were realised in a robot which was evaluated by third-party observers.
Index Terms: multi-modal feedback tokens, human-robot interaction, crowd-sourcing

1. Introduction

In the last decade increasing efforts have been made in order to build artificial listeners [1], [2], [3]. An essential part of artificial listeners is their ability to convey the impression that they are listening and being *attentive* to the speaker. One way of doing this is to use short utterances to provide feedback. In the current paper, we use the term feedback to encompass both feedback tokens such as 'yeah', 'right', 'sure', 'okay', as well as backchannels such as 'mhm', and 'mh'. These tokens are quite short in duration and unobtrusive in their realisation, yet they may carry crucial information about the listeners reaction to the speaker's speech. In addition to indicating the listener's attention, they may also indicate the listeners feelings and his/her understanding of the speaker's intent. In short, feedback tokens, and their subset backchannels, support the conversation [4].

In order for an artificial listener to provide feedback in a human-like manner, two requirements need to be fulfilled: The generated feedback needs to be timed appropriately, and it needs to be realised with the correct lexical form and prosody.

What qualifies as acceptable feedback depends to a certain degree on the semantic content of the speaker's utterance but also to some degree on the mental state of the listener. A listener can be *supportive*, *neutral*, or *sceptical* towards the speaker's speech. These are only 3 of the many possible mental states a listener could assume; in this paper, however, we focus on the 3 states mentioned above.

Attentive listeners can be beneficial for many different application scenarios. One possible application scenario is in aiding the development of language skills in small children [5]. Another application lies in the role of a social companion for the elderly [6]. A third application scenario could be in the area of healthcare support [7].

Current dialogue systems often have random feedback generation - they focus on timing rather than on which feedback token to produce. In this study we test whether we can improve the human likeness of a virtual agent by instead using non-random, speaker-specific lexical choice generation, combined with condition-dependent prosodic variation gathered from crowd-sourced listeners. For this we pioneer a novel data-collection framework which enables us to collect variations of feedback token generation under an identical situational context. This approach enables us to quickly adapt the feedback behaviour of a conversational agent or robot and scale up to many interactional conditions while staying within a person-specific lexical style.

In order to build a model which captures the variation of situation-dependent lexical choice and prosodic realizations, we first test whether we can quantify differences within our crowd-sourced listeners. We hypothesise that the proportional distribution of lexical tokens differs significantly across situations, as do their prosodic realizations. Using the crowd-sourced data, and a unit selection feedback token database, we propose a model, implement it in a robot, and test whether people perceive the difference between an attentive/supportive and attentive/sceptical artificial listener.

2. Background

2.1. Feedback Tokens

Different studies have looked at how paralinguistic phenomena influence prosodic realizations of feedback tokens. [8] found that not-distracted listeners tended to speak more loudly and tended to have less variable energy level. They also found a relationship between the pitch variability and level of attentiveness. [9] found that feedback token were often multifunctional, including understanding, agreement, certainty, and negative surprise, and in [10], they also found that feedback tokens can express different degrees of engagement.

Situational appropriateness has been investigated in, for example [11], who found that using backchannels inappropriately may have negative consequences for the dialogue. These negative consequences manifest themselves in less frequent and less specific responses. In the context of neuropsychological interviewing, the choice of lexical backchannel items and their frequencies, as well as the prosodic contour, have been shown to relate to the perceived interviewee's performance [12]. Similarly, an effect on naturalness, empathy, and understanding has been found when considering dialogue context and form of backchannels [13].

Predicting morphological patterns of backchannels on the basis of the linguistic features of the preceding utterance has been investigated in [14].

http://dx.doi.org/10.21437/Interspeech.2017-926

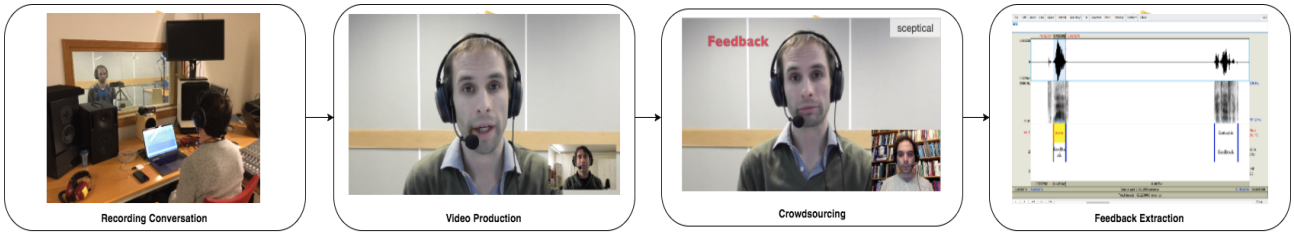

Figure 1: *Crowd-sourced feedback collection.*

2.2. Paper Contributions

In the current paper we investigate crowd-sourced design of feedback generation in an attentive artificial listening agent. Feedback generation has been primarily approached from 2 different angles: On one hand the focus has been on the understanding of human communication patterns, and on the other hand, the focus has been on the engineering of artificial agents and robots.

To our knowledge, studies focusing on the understanding of human communication patterns have looked at feedback tokens across *different* interactional contexts [15, 8, 9, 16, 17]. They have also quantified realizations of paralinguistic phenomena in feedback tokens such as interest, attentiveness, engagement, surprise, etc. [8, 9, 16, 17]. But to our knowledge, no study has investigated feedback token realizations within the *identical* interactional context. Moreover, to our knowledge no other study has explicitly investigated the realization of the paralinguistic phenomena of support and scepticism.

Meanwhile, studies which focused on the engineering of artificial agents and robots have emphasized the *timing* of feedback tokens [1, 18, 19] rather their lexical choice or prosodic realization. In this paper we would like to close the gap between human-centered analysis and agent-centric evaluation, focusing on lexical choice and prosodic realization in human-robot dialogue.

3. Data Collection

We collected 2 distinct corpora; one in-lab and one via crowd-sourcing, as detailed below.

3.1. In-lab

3.1.1. Audio-Visual Recordings

The first corpus comprises audio-visual recordings of one person giving a pitch for a job interview (the speaker), and a second person providing feedback tokens (the listener). Both speaker and listener were male native US English speakers. The speaker enacted 4 different job application scenarios: Pilot, Journalist, PhD Student, and Fashion Designer. Each of these job interviews was recorded under 3 different stances of the listener toward the job application: *supportive*, *sceptical*, and *neutral* (neither supportive nor sceptical). In order to ensure channel separation between the speaker and listener, we placed them in two isolated recording spaces, seated facing each other, where they had visual contact through a window (see Figure: 1). We placed the camera in such a manner as to convey the impression in the video that the speaker is addressing the viewer, while ensuring we did not obstruct the visual contact between speaker and listener. The video was then annotated for the listener's lexical form of feedback tokens as well as their temporal position within the interaction.

3.1.2. Unit Selection Database

We use the term 'Unit Selection Database' to describe the feedback tokens collected from the same listener during the Audio-Visual recordings detailed above. We carried out additional audio-only recordings to supplement this dataset. The database comprised a total of 536 feedback tokens.

3.2. Crowd-sourced

The second corpus of audio-visual recordings comprises video-recorded interactions between the in-lab corpus speaker (job applicant) and a crowd worker taking on the function of the listener. For this corpus collection we developed a web-based application where we presented the the job applicant's videos to the crowd workers, and recorded their audio-visdual data (figure 1). All crowd workers were requested to use a headset microphone, a web cam, and to perform the task in a quiet environment. Crowd workers who did not meet these conditions, or misunderstood the task, were discarded from the corpus.

We used three crowd-sourcing platforms: Prolific Academic [20], CrowdFlower [21], and Amazon Mechanical Turk [22]. We asked crowd workers to watch videos of the job applicant, and imagine they are interviewing him via video conferencing. After the crowd workers watched an initial tutorial video, we presented them with a set of 3 videos from one of four scenarios (e.g.Pilot), with the conditions (*supportive*, *sceptical*, and *neutral*). Their task was to provide feedback in the form of short utterances. We controlled for the timing of the feedback tokens by asking the crowd workers to only produce feedback tokens when prompted on-screen. To enable this, each feedback token from the in-lab corpus was annotated, and subsequently marked in the top left corner of the job applicant's video with an on screen countdown counter. The counter was initialized 3 seconds prior to the onset of the original listener's feedback ("3", "2", "1" "Feedback"). If 2 feedback tokens occurred less than 3 seconds apart from each other, the countdown was shortened accordingly.

In total we gathered interactions from 92 participants , for a total of 276 audio-visual recordings. In this paper, we use a random subset of 40 participants, and 120 total audio-visual recordings. All crowd workers were native English speakers (from USA and Canada).

4. Feedback Processing

4.1. Feature Extraction

We used OpenSmile[1] to extract the following features from the feedback tokens: *F0* mean and slope, *intensity* mean, slope and

[1] `http://audeering.com/technology/opensmile`

range, *jitter and shimmer* means, and *duration*. All features were normalised per speaker. For the subsequent analysis, we only used one-word feedback tokens.

4.2. Feedback Token Generation

As can be observed in Figure 2, token generation is a 3-step process. The first step consists of the generation of a 3 dimensional matrix. The x-axis represents a count of all the lexical tokens observed. The y-axis represents the various listeners and the z-axis represents the conditions. After the matrix has been created we begin the process of picking a lexical token and its prosodic realisation which best suits one of the three conditions: (*supportive*, *sceptical*, and *neutral*). We treat the lexical selection and prosodic realisation selection of the feedback token as two independent processes: For the selection of the lexical token we adapt the notion of different "listening styles". Listeners in our dataset appear to be very consistent in their feedback behaviour within themselves but not when comparing across listeners. Therefore, we decided to model the choice of lexical token according to an individual listener. We randomly pick a listener out of our matrix and perform weighted sampling according to the distribution of feedback tokens per condition. Concerning prosodic realisation, we use the multivariate distribution over all prosodic features and all listeners, within condition. We search our unit selection database for the most similar lexical from and prosodic realisation of the predicted feedback token. For similarity calculation we use a simple euclidean distance measure. If no lexical token in this condition can be found in the unit selection database, then the token with the next most similar prosodic realisation is selected.

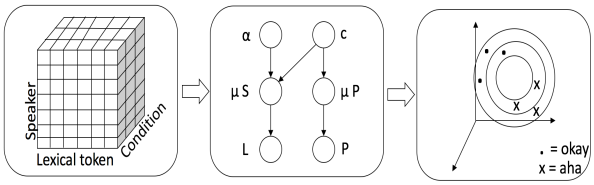

Figure 2: *Feedback token generation.* α*=All Speakers, S=Speaker, L=Lexical Token, C=Condition, P=Prosody*

5. Human-Robot Evaluation

In order to evaluate the performance of our model we generated the predicted feedback tokens in the robot-head "Furhat"[23] within a dialogue context. We ran a crowd-sourced third-party evaluation. We tested whether third-party observers can identify our model's realisations of a supportive versus a sceptical listener.

5.1. Psychological state recognition

For the recognition of the psychological states sceptical and supportive, we assumed the following approach. We provided 40 crowd-workers with side-by-side videos of the "Furhat" robot assuming the role of the listener on one side, and the original job applicant on the other. Each crowd worker compared 4 conditional video pairs (supportive/sceptical), in randomized order. Each video was approximately one minute long. We split the pool of third party observes in 2. Half were asked to indicate which robot listener appeared more supportive, and half

Figure 3: *Feedback generation evaluation, with side-by-side view of human job-applicant and robot.*

were asked to indicate which robot appeared more sceptical. In all cases, they were given the option to chose "I cannot tell the difference".

6. Results

6.1. Lexical Distributions

We wanted to investigate the relationship between the notion of supportive, sceptical, and neutral listening behaviour, and lexical tokens. A chi-square test revealed that the distributions of lexical tokens were different between supportive, sceptical and neutral. ($\chi^2(114, N = 658) = 277.049$; $p < 0.001$) For further detail please refer to Figure 4.

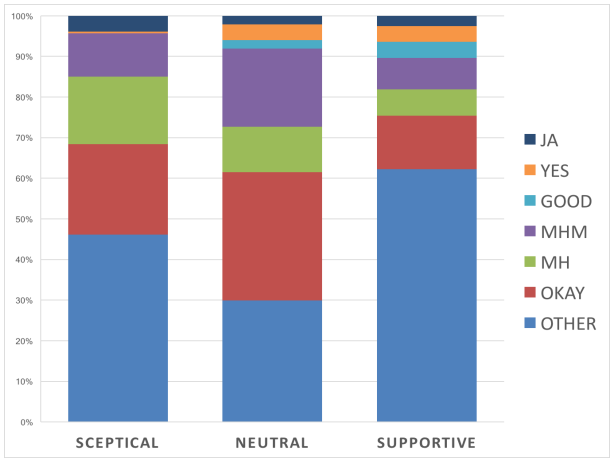

Figure 4: *Distribution of backchannels across conditions.*

6.2. Prosodic Realisations

We wanted to investigate how supportive, sceptical and neutral conditions are related to the prosodic features described in section 4.1. All these features were used as dependent variables in a MANOVA, which showed a general significant effect ($F(16, 1296) = 14.130; Wilk's\Lambda = 0.725; p < 0.001$). For further detail please refer to Table 1.

6.3. Human-Robot Evaluation

For the third-party observer recognition of psychological state we achieved a general accuracy of 63 %. Therefore, we outperform a 50% chance baseline by 13 %.

Table 1: *All features which showed significant differences in prosody between Supportive, Sceptical and Neutral for all one-word feedback utterances.*

Feature	Supportive	Sceptical	Neutral
	N = 204	N = 234	N = 220
F0 Mean	M = 0.44	M = -0.22	M = -0.21
	SD = 0.11	SD = 0.99	SD = 0.9
F0 Range	M = 100	M = 100	M = 100
	SD = 10	SD = 10	SD = 10
Intensity Mean	M = 0.42	M = -0.21	M = -0.44
	SD = 1.2	SD = 0.91	SD = 0.69
Intensity Slope	M = 0.11	M = -0.14	M = 0.10
	SD = 1.4	SD = 0.94	SD = 0.91
Intensity Range	M = 0.27	M = -0.34	M = -0.49
	SD = 1.11	SD = 0.83	SD = 0.64
Jitter Mean	M = 0.15	M = -0.08	M = -0.21
	SD = 1.12	SD = 1.11	SD = 0.88
Shimmer Mean	M = 0.06	M = -0.27	M = 0.04
	SD = 1.01	SD = 1.01	SD = 1.11
Duration	M = 0.45	M = 0.54	M = 0.43
	SD = 0.15	SD = 0.23	SD = 0.15

7. Discussion

We found that the distribution of lexical feedback tokens varies significantly across neutral, sceptical, and supportive conditions. It can be observed, for example, that 'Okay' occurs most frequently in the neutral condition, followed by sceptical, then supportive. While the feedback token 'mhm' is also the most frequent in the neutral condition, it is not noticeably different between sceptical and supportive. On the other hand, 'mh' is the most frequent in sceptical. 'Yes' occurs most frequently in supportive and neutral, and almost never in sceptical. 'Good' occurs nearly exclusively in supportive.

It has been found in previous studies that engagement, interest and surprise in feedback tokens are characterised by high or rising pitch cues[24, 25]. However, in another study, no difference of rise in final pitch slope could be observed for the perception of attentiveness[15]. In the current study, we also did not find a difference in pitch slope between supportive, sceptical or neutral feedback realisations. Differences in findings might be due to the different characteristics of the different corpora used, or the feature extraction methodology. Further difference could be explained by the fact that we fixed the timing of feedback token realisation which contrasts all previous studies. Although we did not find differences in pitch slope, we found differences in all other extracted prosodic features, including average pitch, across conditions. The fact that average pitch was higher in the supportive and neutral conditions is in line with findings of [24, 25]; it could be argued that listeners who want to convey their support are also perceived as more interested and engaged. Furthermore, our finding that in the affect-laden conditions feedback tokens are realized with a higher intensity in comparison to neutral, is in line with [8], and [15], who both found that attentive listeners tend to speak more loudly.

Moreover, our preliminary model for generating different psychological listener states in an attentive artificial listening robot outperforms the 50% chance baseline by 13 %.

While our preliminary model outperforms the chance baseline, it does not perform as well as we might have hoped for. A first possible explanation for this might be in the fact that we did not add an additional step in which the crowd-sourced feedback tokens were evaluated by another set of crowd-workers, or experts in terms of their suitability for expressing a given psychological state. However, while such an evaluation might have strengthened the perception test results, it might have also led to the selection of more stereotypical feedback token which might not necessarily the represent the variation present in human-human interaction. In future work we would like to investigate this trade-off further.

A second possible explanation might lie in the fact that we are neither modelling facial expressions nor head nods. For example, [26] found that fusing head nods and vocal backchannels aids the perception of attentiveness in virtual agents. In a future study we therefore would like to investigate whether perception test results might be improved if head nods and facial expressions are included as well.

8. Conclusion and Future Work

In this paper we investigated feedback generation in human-human as well as human-robot interaction. We showed that our novel crowd-sourcing data collection framework was valid, and could be used to easily scale up listening behaviour representing different psychological conditions. We found prosodic as well as lexical differences across conditions. Further, we implemented our models in a robot. Third party observers could distinguish between a sceptical versus a supportive feedback realisation in a "Furhat" robot. In a future study we would like to not only investigate the use and generation of audio-feedback tokens in different listening conditions and improve on the feedback generation algorithm but also investigate the use of head-nods, smiles and facial expressions in general. Moreover, we would like to investigate the use of different synthesis techniques. Currently, we are mapping from the model prediction to the closest lexical form and prosodic realisation in our unit selection database. We think that it should be possible to improve both on the selection algorithm for the most appropriate feedback form but also exchange the unit selection database for more flexible statistical-parametric synthesis.

9. Acknowledgements

The authors feel particularly thankful to Todd Shore for making the data recordings possible. The authors would also like to acknowledge the support from the Swedish Research Council Project InkSynt (2013-4935), the EU Horizon 2020 project BabyRobot (687831) and the Swedish Foundation for Strategic Research project EACare (RIT15-0107).

10. References

[1] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[2] R. Poppe, K. P. Truong, and D. Heylen, "Perceptual evaluation of backchannel strategies for artificial listeners," *Autonomous agents and multi-agent systems*, vol. 27, no. 2, pp. 235–253, 2013.

[3] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic, and C. Pelachaud, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.

[4] V. Yngve, "On getting a word in edgewise," in *Papers of the sixth regional meeting of the Chicago Linguistic Society*, Conference Proceedings.

[5] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal, "Telling stories to robots: The effect of backchanneling on a childs storytelling," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 100–108.

[6] C. Balaguer, A. Gimenez, A. Jardon, R. Cabas, and R. Correal, "Live experimentation of the service robot applications for elderly people care in home environments," *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2345–2350, 2005.

[7] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, Conference Proceedings.

[8] Z. Malisz, M. Wlodarczak, H. Buschmeier, S. Kopp, and P. Wagner, "Prosodic characteristic of feedback expressions in distracted and non-distracted listeners," in *Interspeech 2012*, 2012, Conference Proceedings, pp. 36–39.

[9] D. Neiberg, G. Salvi, and J. Gustafson, "Semi-supervised methods for exploring the acoustics of simple productive feedback," *Speech Communication*, vol. 55, no. 3, pp. 451–469, 2013.

[10] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in swedish," in *DiSS-LPSS*, 2010, Conference Proceedings, pp. 63–66.

[11] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology*, vol. 79, no. 6, pp. 941–952, 2000.

[12] G. Bailly, F. Elisei, A. Juphard, and O. Moreaud, "Quantitative analysis of backchannels uttered by an interviewer during neuropsychological tests," in *Interspeech*, 2016, Conference Proceedings.

[13] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. Ward, "Prediction and generation of backchannel form for attentive listening systems," in *Interspeech 2016*, 2016, Conference Proceedings, pp. 2890–2894.

[14] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. Ward, and T. Kawahara, "Analysis and prediction of morphological patterns of backchannels for attentive listening agents," in *International Workshop on Spocken Dialogue Systems*, 2016, Conference Proceedings.

[15] C. Oertel, J. Gustafson, and A. W. Black, "Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances," in *Proc. of Interspeech*, 2016, pp. 2915–2919.

[16] P. Satish and M. Schrder, "Annotating meaning of listener vocalizations for speech synthesis," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, Conference Proceedings, pp. 1–6.

[17] L. Catherine, "What do you mean, you're uncertain? the interpretation of cue words and rising intonation in dialogue," in *Interspeech 2010*, 2010, Conference Proceedings.

[18] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Computer Speech & Language*, vol. 28, no. 4, pp. 903–922, 2014.

[19] M. Johansson, T. Hori, G. Skantze, A. Höthker, and J. Gustafson, "Making turn-taking decisions for an active listening robot for memory training," in *International Conference on Social Robotics*. Springer, 2016, pp. 940–949.

[20] "Prolific Academic," https://www.prolific.ac, 2017, [Online; accessed 19-March-2017].

[21] "CrowdFlower," https://www.crowdflower.com, 2017, [Online; accessed 19-March-2017].

[22] "Amazon Mechanical Turk," https://www.mturk.com, 2017, [Online; accessed 19-March-2017].

[23] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive behavioural systems*. Springer, 2012, pp. 114–130.

[24] J. Liscombe, J. J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features." in *INTERSPEECH*, 2003.

[25] D. Neiberg and J. Gustafson, "Cues to perceived functions of acted and spontaneous feedback expressions," in *The Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Citeseer, 2012, pp. 53–56.

[26] C. Oertel, J. Lopes, Y. Yu, K. A. F. Mora, J. Gustafson, A. W. Black, and J.-M. Odobez, "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*. ACM, 2016, pp. 21–28.