Radboud University Nijmegen


Artificial Intelligence

# Effects of emotion on the preceived performance of backchannel strategies

*Author:*
Wouter Eijlander

*Supervisor:*
Khiet Truong

June 2015

# Table of Contents

## Abstract

**The effect of conversational emotion on the perceived performance of models of backchannel behaviour is evaluated. These are behavioural models which are used to appropriately provide feedback to a speaker in a non-obtrusive manner to display attention. The effects of anger and happiness on three models of behaviour are investigated: one proposed by researchers previously, one random, and one to copy human behaviour. The backchannel behaviour was performed by a robot. In a user experiment subjects were shown video recordings of this behaviour during conversation and were asked to rate the appropriateness of the behaviour. Significant effects of both differences in emotion and model on the perceived performance of backchannel behaviour have been found, but an interaction effect was absent.**

# 1 Introduction

In 1950, Alan Turing proposed a test to assess artificial intelligence, which we now refer to as the Turing Test [1]. In this test, the conversational behaviour of a computer and a human would be compared by a subject who does not know the true nature of either party. If a computer could convince 70% of the subjects of its humanity within five minutes of conversation, it would be deemed 'intelligent'. However, such a conversation would happen through a chatbox-like interaction, with no actual face-to-face value. Thus, an algorithm that passes the Turing Test does not necessarily hold convincing conversation in all modalities. Such an algorithm uploaded to a robot would still only be convincing in the modality it has been created for: generating human-like conversation in a text-based modality. A robot, however, requires many more convincing qualities than just the capability to formulate sentences: It would need to be correct in all its timing, movement and prosodic features, as well as provide feedback to its conversational partner.

Such feedback, more commonly referred to as 'backchannels' (BC), a term first introduced by Yngve [2] is commonly seen as short utterances indicating attention by the listener in a conversation [3] without actually interrupting the speaker. BC behaviour can consist of vocal utterances, both verbal and non-verbal (i.e. "mm-hm" and "yeah" respectively), but can also be characterized by a movement, such as a nod of the head. While humans seem to exhibit this behaviour automatically, robots need to use algorithms to determine when and how to use such cues to indicate that they are still partaking in the interaction. Researchers have proposed and tested many such algorithms in order to find out how positively people perceive the behaviours generated by them. The technical backgrounds of these models vary widely, ranging from relatively simple rule-based systems such as the Ward and Tsukahara model [3] to regression-based models such as proposed by Terell and Mutlu [4]. The performances of these models differ widely as well.

Comparative research between such models has been done on several occasions. An interesting example is an article by Poppe et al. [5], who compare the widely known model by Ward & Tsukahara to a range of other models. They found that the model by Ward & Tsukahara was outperformed by nearly all other models. This clear distinction between performances of a much referred-to model and, among others, a strategy that performs essentially semi-randomly seemed striking. These results were based on research performed with the Semaine Corpus [6], which features lengthy video material of human conversation with several different prevalent emotions. The results found by Poppe et al., however, did not take into account the possibility of an effect of the prevalent emotion on model performance. This lead me to hypothesize that conversational emotion may affect the way backchannel behaviour is perceived. I theorized that this may have caused the Ward & Tsukahara model to be perceived so negatively as compared to others, (especially a random model,) while it has been shown to be a better predictor for human backchannel behaviour than a random model. It is here that a clear distinction must be made between prediction accuracy and perceived performance: *Prediction accuracy* denotes how accurately a model

can predict human backchannel behaviour [3,4,7]. *Perceived performance*, however, refers to how positively people think about the behaviour dictated by backchannel models [5].

In this thesis I will attempt to identify whether prevalence of a specific emotion may indeed have an adverse or beneficial effect on a selection of these models over others. More specifically, I will answer the following questions:

- *Does conversational emotion affect the performance of backchannel models?*
- *Can conversational emotion cause models to have a higher perceived performance than others that score higher under different conversational emotion?*

By answering these questions I hope to explain the seeming contradiction that a model with higher prediction accuracy than others may still have a lower perceived performance on average. It may also serve to improve understanding of the circumstantial performance of backchannel models, and contribute to the field of human-robot interaction.

In the following sections I will highlight some relevant research into backchannel models as well as conversational emotion. Section 3 will explain the BC models used during the experiment, the creation of the used stimuli, and the overall procedure of the experiment. Section 4 will present the experiment results and statistical analysis thereof. Finally, section 5 will discuss the results presented earlier and the thesis will be concluded.

# 2 Previous Work

## 2.1 Backchannel research

Previous research into the models of BC behaviour are usually aimed at finding out the general effectiveness of these algorithms. Some measure this 'performance' by subject ratings, asking them to grade the behaviour based on how human-like, suitable, or simply how pleasant it seems [5]. Others focus rather on the mimicking of human backchannel behaviour, and measure the performance of their model by how well it can predict human backchannels.

For example, Ward and Tsukahara [3] proposed a model that is by now quite well-known among researchers in the field. This model was made to provide a prediction of listener backchannels by analyzing continuity and pitch of the presented speech. The model was then tested for prediction accuracy in English as well as Japanese conversations. It was found to perform better than random in both languages. A notable difference in accuracy between languages was found, with Japanese conversations resulting in higher accuracy levels for both the tested model and the random model (34% and 24% respectively, as compared to 18% and 13% for English).

A much more intricate prediction model was proposed by Morency et al., who utilized a probabilistic model trained with a combination of Hidden Markov Models and Conditional Random Fields [8] to predict listener backchannels in real-time [8]. They then compared the performance of their model to that of Ward & Tsukahara and a random strategy, correlating the timing of predicted backchannels to the timing of actual listener backchannels. They found that, although the model by Ward & Tsukahara outperformed the random strategy, their own model was a significantly better predictor for listener backchannels.

Poppe et al. [5] proposed several models based on speech continuity and pitch, and compared these models to a random strategy and a listener-copying strategy based on how fitting human subjects felt the BC timings were. Among these models were a random strategy, a copy strategy that dictates the same behaviour as a documented human listener exhibits, and three of their own models, based on gaze direction, on pitch level and pause duration, and on pitch level, pause duration and gaze direction. They found that, among the tested models, the Ward & Tsukahara model was rated much lower than the others, including the random strategy. These results may

suggest that a model that has been found to provide more accurate BC timings than a random model does not necessarily cause the behaviour to be viewed more positively.

Shinya et al. [9] went a step further and researched the workings of a backchannel model based partly on linguistic content of the spoken utterances. They report an interesting finding that, where the properly applied content of the backchannel is largely dependant on the content of the spoken utterance, the timing of the backchannel is based rather on the style of the spoken utterance. Both aspects of backchannel behaviour still depend in part on what they call the *dialogue strategy*. However, this finding suggests that while the timing of a backchannel is important in research into the model, the actual behaviour performed during the backchannel is a variable that requires another factor to be taken into account. This introduces a divide between backchannel *timing* and *content*, suggesting that both require dedicated research into the proper use of backchannels in HCI.

### 2.2 Conversational emotion

It has been shown on several occasions that emotion can be reliably derived from speech. Banse and Scherer [10] have found that different emotions show significantly different acoustic features during speech. Cowie et al. [11] provide similar results, but also mention a certain variance between speakers. They noticed that each speaker had their own stable state, noting amplitude deviations from this state , with effect sizes varying between subjects. This would indicate that, although a stable state of acoustics exists during neutral conversation, this state would be difficult to generalize.

Bachorowski [12] indicates humans' capability of inferring emotion from the spoken word alone, stating that is a basic part of human vocal communication. She too, however, notes variety in prediction strength between both different voices and different emotions. She claims this may be caused by certain emotions being more difficult to convey or detect using only the vocal dimension, but also attributes it to differences between speakers.

## 3 Experiment

A user experiment was conducted using an online survey for subjects to fill in. This survey consisted of videos of a human speaker and robotic listener, accompanied by two questions each. This section will proceed to discuss the implementation of selected backchannel models, the creation of survey stimuli, and the collection of experimental data.

### 3.1 Backchannel behaviour

Due to the the limited scope of this thesis research, there was only room for a testing three models of interest: the model by Ward and Tsukahara, a (semi-)random strategy, and a copy strategy, which performs backchannels at the same moments as a documented human listener. For the same reason, the research was limited to effects of happiness and anger, since these are easily distinguishable. The Ward & Tsukahara model and the semi-random strategy are defined as follows:

| Backchannel strategy as specified by Ward and Tsukahara [3] |
|---|
| *(P1) a region of pitch less than the 26th-percentile pitch level and* |
| *(P2) continuing for at least 110 milliseconds,* |
| *(P3) coming after at least 700 milliseconds of speech,* |
| *(P4) providing you have not output back-channel feedback within the preceding 800 milliseconds,* |
| *(P5) after 700 milliseconds wait* |

| (Semi-)Random Backchannel strategy |
|---|
| *(P1) Determine amount of backchannels n predicted by W&T for a sound fragment x, then*<br>*(P2) divide fragment x into n equal parts, and*<br>*(P3) provide one backchannel timing per fragment part using a uniform random number generator.* |

Lastly, the copy strategy makes use of backchannel timings from the actual human listener in the respective fragment. These were produced with a 4ms time resolution.
The Ward & Tsukahara strategy followed an implementation in Praat [13] implemented and used by Poppe et al. [5].[1] This implementation required an audio file as input, which are included in the Semaine Corpus [6]. The proper fragments were extracted from each sound file (see *Fragment selection*) and were subsequently analyzed using the models to determine the proper backchannel timings. The semi-random strategy did not rely on the input of sound files, and was thus easy to implement in Python. The copy strategy consisted of taking note of all listener backchannels from the relevant fragments.

The determined backchannel timings needed to be given shape. According to Gardner [14], the implications of different vocalisms during backchannels vary widely, and have significant effects on how the backchannel is perceived. In order to avoid a bias caused by 'erroneous' vocal utterances, they were chosen to be left out entirely. As has also been argued by Shinya et al. [9], the backchannel content is largely dependant on the content of the spoken dialogue. Making the robot perform the same backchannel for different conversational content could as such increase this bias in the results. Although leaving vocal backchannels out of the experiment would likely decrease perceived performance, it would not cause a variety of biases between different models, emotions and conversations. The previously determined BC timings were then implemented in a NAO [15], making it nod its head for every backchannel timing, taking 40ms for each nod.

## 3.2 Fragment selection and video creation

The goal of the experiment is to find out the effect of conversational emotion on the perceived performance of these models. In order to test this, a subset of the same videos used by Poppe et al. was taken from the *Semaine corpus of emotionally coloured conversation* [6]. This corpus features lengthy video material of human conversation, with a speaker fulfilling the role of a certain character that provides a certain emotion. The four different characters convey happiness, sadness, anger and calm. Each video is accompanied by extensive annotations of, among others, emotion values with a resolution of 2ms, ranging from a score of -1 to 1. These scores indicate how noticeable the respective emotion is at that point in the conversation, with higher scores indicating stronger prevalence. A simple Python script was used to filter these annotations and display all time windows where specific emotion scores were positive. From these time windows, five fragments were chosen for both emotions, selected on the length of the fragments, and the usability (i.e. as little interruption from the speaker as possible). The video fragments of the happy speaker averaged 12.2 seconds in length, while those of the angry speaker were on average 24.2 seconds. After collecting all suitable backchannel timings for each model and video, the backchannel timings were implemented and performed on camera by a NAO [15]. The videos were then placed side by side using a script in FFmpeg [16] such that the backchannels all happened according to their determined timing. Fig. 1 presents a still from one such video, featuring the human speaker on the right, and the robotic 'listener' to the left.

[1] Thanks to K.P. Truong for providing access to their implementation of the Ward&Tsukahara model

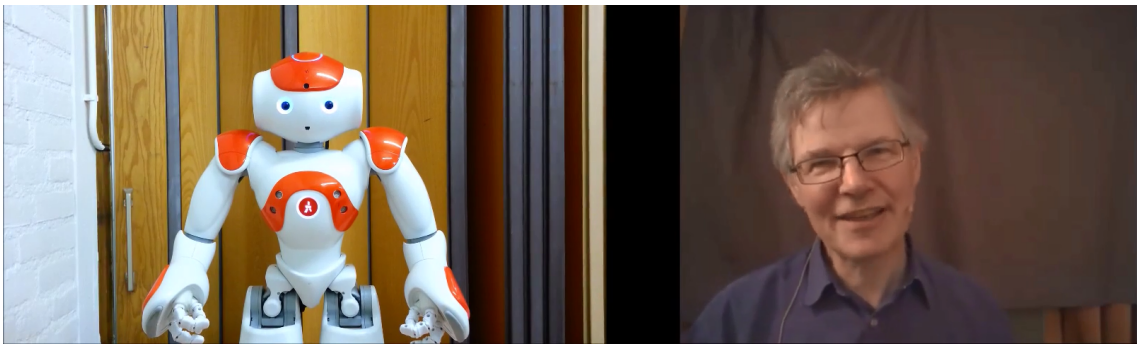

Figure 1: Still from one of the surveyed videos

### 3.3 Subjects

A collection of Students and their relatives were asked to participate in the experiment, of whom 28 subjects responded. The average age of the subjects was 23.125 (min. 19, max. 55), with 20 male and 8 female subjects. Among the subjects were 13 Artificial Intelligence students, 4 art students and 3 biologists; the other subjects all had different professions and occupations. The subjects filled in an online survey in Dutch, in which all subjects were fluent. Subjects were assigned to a subset of questions based on the moment at which they filled in the survey. Due to irregular frequency of subject influx, 16 subjects were assigned to the first subset, and 12 to the second. As a result, a slight difference in number of ratings between the model-emotion combinations has arisen, as can be seen in table 1.

### 3.4 Procedure

Subjects were shown the videos in an online survey. They were asked for their age, gender and profession, and were then shown two sample videos to prepare them for the survey. For each video they shown, subjects were asked to rate the robot's behaviour on a scale from 1 (not appropriate) to 10 (appropriate). They were also asked to rate the speaker's emotion on a scale from 1 (angry) to 10 (happy). Asking subjects to indicate the present conversational emotion allows for an integrity check of that independent factor, subsequently indicating or ruling out a major confounding variable in the research. At the end of the survey, subjects could provide textual feedback about the survey.

For each combination of model and emotion five separate videos were produced, providing a total of 30 videos. The speaker's side of the video was the same for each model over a specific emotion. That is to say, for the W&T strategy under the happy emotion the speaker's side of the videos is the same as that of the copy- and random strategy under the happy emotion. The same holds for the angry emotion. This meant that presenting each subject will all 30 videos would result in them hearing the speakers say everything three times over the course of the experiment, with the only difference being the robot's behaviour. This could cause subjects to believe they had already seen certain fragments, and give the same rating as they had previously done, creating a performance bias. In order to prevent this, the videos (and thus, questions) were split into two subsets of 15 questions with uniformly distributed model-emotion combination. Each subject would be shown only one subset, meaning they would only hear a speaker say something at most twice.

| **Perceived Performance** | W&T | Random | Copy |
|---|---|---|---|
| Angry | 70 | 65 | 70 |
| Happy | 65 | 70 | 65 |

Table 1: Division of number of subject responses per model-emotion combination

# 4 Results

The previous section will present the results provided by the experiment described here, and will also discuss the statistic analyses performed. The validity check of distinguishable conversational emotion will be discussed first, followed by the results from the perceived performances.

## 4.1 Distinguishing conversational emotion

In order to be able to account for a possible bias in results, subjects' judgement of the conversational emotion must be accurate. As described in *3.3* these judgements were collected, and a paired-samples T-Test was performed on the results. On a scale from 1 (Angry) to 10 (Happy), under the happy condition subject ratings of videos had a mean of 7.24 with a standard deviation of 1.32. On the same scale, videos under the angry condition received ratings with a mean of 3.16 and a standard deviation of 1.334. The two conversational emotions were found to differ significantly ($p < 0.001$), meaning that the emotion ratings under the happy condition differ significantly from those under the angry condition.

## 4.2 Perceived performance

The results from the survey were analyzed using a GLM-Repeated Measures to find out the differences between perceived performance of all six combinations of within-subject factors (3 models x 2 emotions). The means resulting from this can be seen in table 2, with the corresponding interaction plot in fig. 2. As can be seen, the perceived performance of the *random* model is approximately equal between both emotions, and the *Ward & Tsukahara* strategy as well as the *Copy* strategy score lower on perceived performance under the angry condition than the happy condition.

The within-subject contrast of perceived performance between *Ward & Tsukahara* and the *Random* model was not significant ($p > 0.4$), whereas the contrast between *Ward & Tsukahara* and the *Copy* strategy was significant ($p < 0.01, R = 0.118$). The effect of emotion (angry versus happy) was also found to be significant ($P < 0.05, R = 0.139$). There was, however, no significant interaction-effect for W&T versus Random ($p > 0.1$) nor *W&T* versus *Copy* ($p > 0.8$).

# 5 Conclusion & Discussion

The results presented in the previous section allow for some conclusions to be made with regards to the hypotheses and questions stated in this thesis. The following section will present such relevant conclusions with regards to these results. It will also discuss the general procedure and factors to be taken into account, as well as discussing some future work potential.

## 5.1 Discussion

One should take note that due to the limited scope of research possibilities, several factors have had to be left out. One example would be the fact that testing more than two emotions would quickly become unfeasible for a project of this scope. When taking into account more emotions

| | Mean | Std. Deviation | N |
|---|---|---|---|
| W&T x Happy | 6.42 | 1.793 | 65 |
| W&T x Angry | 5.71 | 1.860 | 65 |
| Random x Happy | 5.94 | 1.519 | 65 |
| Random x Angry | 5.89 | 1.650 | 65 |
| Copy x Happy | 5.80 | 1.962 | 65 |
| Copy x Angry | 5.00 | 2.312 | 65 |

Table 2: Division of number of subject responses per model-emotion combination

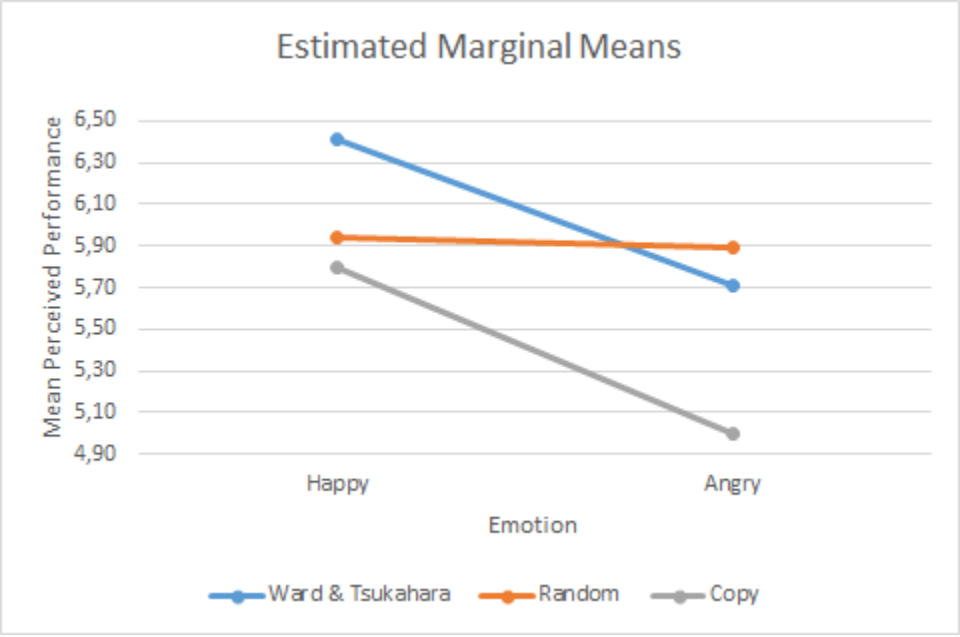


Figure 2: Perceived performances of the models per emotion

than simply anger and happiness (i.e. sadness, disgust or surprise) one might come to wholly different conclusions than I. Adding levels to this factor would certainly provide a more reliable set of results, but would also require more resources, time, and subjects. It is, however, a step that should be taken in order to get a better understanding of the overall effect of conversational emotion on the perceived performance of backchannel behaviour.

As mentioned in 3.1, the backchannel behaviour used consisted solely of nods. The presence of vocal utterances could have caused a severe bias due to the differences in implications between various vocalities. An option would have been to have vocal behaviour dictated by models as well. However, having such models would complicate the experiment, requiring more subjects in total, and requiring more time. Having no vocal feedback for backchannel behaviour, however, can be expected to decrease perceived performance over all models, since vocal backchannels are to be expected in symmetrical interactions. The bias caused by this may thus have affected my results. Many subjects communicated (at the end of the survey or through personal communication) that the absence of backchannels in some videos was very striking. Another factor which subjects also communicated repeatedly is the limitation of using a robot. A real NAO robot was used for this research, rather than a simulated one or a virtual agent. Every nod was actuated by mechanisms which can make noise and behave unnaturally. This too may have introduced a confounding variable that may be worth consideration for future research.

The low perceived performance of the copy strategy could be interpreted as an indication that human backchannel behaviour may not be perceived very well, and that the behaviour dictated by even the simplest of models may outperform human behaviour. However, it should again be noted that the model did not actually mimic human behaviour; all vocal utterances were left out and only the timing of human backchannels was copied. Changing the nature of the backchannels into a single modality while some might only be appropriate in an other one could introduce a certain bias to the results of this model. It should also be noted that Poppe et al. [5] found that the perceived performance of human backchannel timing was higher than that of the other models they tested on. This would indicate that human behaviour is in fact perceived to be appropriate. Their implementation, however, dictated every backchannel to be a nod accompanied by a vocal "mm-hm", and was performed by an animated virtual human. These differences form the set-up used in this research may have had a significant effect on perceived performance as well.

These differences would provide a basis for future research. The effects of using different models for vocalising backchannel behaviour (or leaving them out entirely) may provide significant insight in how to apply the various natures of backchannel behaviour, rather than focusing on the use of different models.

## 5.2 Conclusion

I have tested whether conversational emotion can have an effect on the perceived performance of certain backchannel models. Between the happiness and sadness emotion, a significant effect on the perceived performance of both the backchannel model by Ward & Tsukahara and a listener-copying strategy has been observed, providing supportive evidence that conversational emotion does in fact affect the perceived performance of backchannel behaviour. An effect on the perceived performance of a random strategy has not been observed, as can be expected from a behavioural model that is essentially random. Subjects were capable of distinguishing happiness and anger in conversation. This means that the effects of emotion were not a random outcome of the difference between videos or speakers, but can actually be attributed to conversational emotion.

The absence of a significant interaction-effect provides evidence against models outperforming others on average due to changes in conversational emotion. That is to say, on average, a change of conversational emotion affects the perceived performance of different models the same way. This would contradict my former hypothesis stating that the averaging of perceived performances

may have caused the *Random* model to have outperformed the *Ward & Tsukahara* model in the research by Poppe et al. [5]. However, it should be noted that, as can be seen in table 2 and figure 2, the *Random* model scores much lower than the *Ward & Tsukahara* model under the happy condition, but by staying stable under a change of conversational emotion, outperforms both other models under the angry condition. Although no significant difference has been found between the perceived performances of the *Ward & Tsukahara* model and the *Random* model, nor has an interaction-effect been observed, further research may well contradict the results and conclusions described here.

# 6 References


1 A. M. Turing (1950): Computing Machinery and Intelligence. Mind 49: 433-460.

2 Yngve, V. "On getting a word in edgewise," page 568. Papers from the Sixth Regional Meeting [of the] Chicago Linguistic Society, 1970.

3 Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207, 2000.

4 Terell, A., Mutlu, B.: A Regression-based Approach to Modeling Addressee Backchannels. Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 280–289, 2012.

5 Poppe, R., Truong, K.P., Reidsma, D., Heylen, D.: Backchannel Strategies for Artificial Listeners, IVA 2010, LNAI 6356, pp. 146–158, 2010.

6 Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic, 'The SEMAINE Corpus of Emotionally Coloured Character Interactions', Proc. IEEE Int'l Conf. Multimedia & Expo (ICME'10), pp. 1079-1084, Singapore, July 2010.

7 Morency, L. P., de Kok, I., Gratch, J. (2008, January). Predicting listener backchannels: A probabilistic multimodal approach. In Intelligent Virtual Agents (pp. 176-190). Springer Berlin Heidelberg.

8 Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282–289.

9 Shinya, F., Kenta, F., Tetsunori, K. (2004)."A Conversation Robot with Back-channel Feedback Function based on Linguistic and Nonlinguistic Information". 2nd International Conference on Autonomous Robots and Agents December 13-15, 2004 Palmerston North, New Zealand

10 Banse, R., Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), 614–36. Retrieved from

11 Cowie, R., Douglas-Cowie, E., Romano, A. (1999). Changing emotional tone in dialogue and its prosodic correlates. In ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody.

12 Jo-Anne Bachorowski. (1999). "Vocal Expression and Perception of Emotion". Current Directions in Psychological Science April 1999 vol. 8(2), 53-57.

13 Boersma, Paul (2001). Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.

14 Gardner, Rod. 2001. When Listeners Talk: Response Tokens and Listener Stance. Amsterdam: J. Benjamins Publishing.

15 NAO, Aldebaran Robotics. https://www.aldebaran.com/en/humanoid-robot/nao-robot

16 FFmpeg: http://www.ffmpeg.org