Manipulating artificial speech: the effects of prosodic manipulation of the response latency

Tomas Schulkes
University of Twente
(+31) 6 – 44 311 366
t.a.schulkes@student.utwente.nl

ABSTRACT
Prosody is defined to be the rhythm, stress and intonation of speech. Studies have shown that prosody is affected by the rapport of two interlocutors. This raises the question whether or not rapport could be affected when prosodic elements in speech are manipulated. One of these prosodic elements is the response latency – the pause time between switching speakers.

This paper describes a research focusing on testing if manipulation of the response latency of artificial speech agents improves their rapport as perceived by human test subjects.

To achieve this, an experiment is conducted using synthetically generated voices in a Text-To-Speech engine. Three audio-fragments are generated, all having a different manipulation on response latency. Test-subjects listened to one of the three fragments and judge that conversation on a number of points related to the rapport of the artificial agents.

Keywords
Response latency manipulation, prosody, artificial speech, rapport.

1. INTRODUCTION
Efforts to create electronic Text-To-Speech (TTS) engines have been around at least since the 1930s when Bell Labs developed the ‘vocoder’, an analysis/synthesis system, used mainly for speech. In the early days, and even now, the generated voices sounded monotonous, tinny and choppy. This made for little rapport between a human and the speaking machine.

These issues stem from two intertwined problems. The first problem is how to generate the correct sounds of the speech and the second problem is how to model prosody. Prosody is defined to be the rhythm, stress and intonation of speech.

A lot of research has been performed to deal with these issues. The generations of the sounds have been addressed from different angles, from all-mechanical generation in [7] to all-digital solutions in TTS engines like Open MARY, Loquendo TTS and others. The problem of prosody has been addressed in other research, for example [5] and [7].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

According to [8], rapport can be measured in three essentials. Participants in an interaction that experience themselves to have a high degree of rapport experience a feeling of intense mutual interest in the other participant(s). ‘Mutual attentiveness’ is the first essential component of rapport. It creates a focused and cohesive interaction. ‘Positivity’ is the second essential component for humans feeling in rapport to one another. A higher degree of mutual friendliness and caring makes for a higher degree of feeling rapport. The last essential component of rapport is ‘coordination’ between participants.

In simple human terms, rapport is the feeling of being “in-sync” with another person.

In [6] it is shown that some aspects of prosody are affected by rapport between conversing dyads. In particular, the individual response latencies of the dyads tended to converge to a common value. One would expect this effect to go both ways, so this raises the question whether or not manipulating prosody of an artificial agent increases its rapport as perceived by its human interlocutor. In this research, focus will lie on manipulation of the response latency. In this study, an experiment will be performed to evaluate the effect on rapport of two artificial speech agents when the response latency is manipulated.

The experiment will have 2 artificial agents speaking to each other. Manipulation will occur on the response latency: one version the response latencies of the agents diverging, one version will have no manipulation and the last version will have the response latencies converging.

Based on [6], it is expected that the version of the experiment with the converging response latencies will be evaluated to have a higher rapport than the other two versions. Also, the version with no manipulation is expected to have a higher rapport than the version with the diverging response latencies.

Should a manipulation yield positive results, then future agents that have speech-ability should be programmed with a proper model of prosody that incorporates response latency manipulation.

In particular, the following questions will be investigated:

1. What is the role of response latency in human interaction?
2. What TTS engines are available and which one of these is most suited for this study?
3. What question should the test subjects be asked to measure the effect on rapport?
4. Does the manipulation of the response latency yield an effect on rapport as perceived by human test subjects?
To find the answers to these questions, literature to the subject of prosody in human-to-human interaction has been evaluated to setup a suitable experiment.

To find the best TTS engine, some mainstream TTS engines have been evaluated on a number of issues. Of these issues, the ability to speak Dutch and the quality of the voices were two heavy-weight factors.

To test whether manipulation yields an effect on rapport, the subjects answered a questionnaire. This questionnaire has been made based on questions in other studies that have shown the ability to effectively test rapport and have valid evaluation methods.

This paper first shows the role of response latency in human interaction. Then, the experiment that has been conducted is described. Finally, the experiments findings will be evaluated and a conclusion is drawn.

2. PROSODY AND RESPONSE LATENCY

Prosody is a generic term to describe all suprasegmental features accompanied by human speech [4]. A suprasegmental feature is a feature of speech extending over more than one single speech sound, for example tone, stress or response latency. Response latency is one of the major items of these features, even though it partly results from time required for processing what the other person has said and generating a proper response. All of the features together provide clues to the meaning of the sentence when the speaker reaches a clause boundary.

Most features co-occur and overlap in providing cues, so the absence of one feature can be (partly) compensated by the presence of the other features.

In [6], conversing dyads have been studied on prosody, in particular the synchronization of response latency during 15 minute conversations.

The response latency is defined to be the duration of the pause that results from 1 speaker finishing his sentence, and another speaker starting theirs.

Results of the study in [6] have shown that during the conversations, the response latency tended to gradually converge to a common value when the dyads felt high rapport to each other. Despite what one may expect, this value wasn’t necessarily the mean of the two individual response latencies. Indeed, it could even be a value above the highest or below the lowest initial response latency.

3. EXPERIMENT

To test whether or not manipulating the response latency of an artificial speaker affects its rapport, an experiment was conducted. The experiment had the test subjects listening to a conversation between 2 artificial speakers. One of the speakers is a teacher. The other speaker is a student.

The conversation that the experiment subjects listened to had a student and a teacher speaking for 9 minutes about homework and everyday life. The initial response latencies were 300 milliseconds for the student, and 600 milliseconds for the teacher.

Three versions using the same transcript were created. The first version (DI) had the response latencies gradually diverging to 750 milliseconds for the teacher and 200 milliseconds for the student. The second version (NA) had no adjustments made to the response latencies. The third version (CO) had the response latencies gradually converging to 450 milliseconds for both speakers. Except for the response latency, all three generated versions of the conversation were equal.

The content of the conversation was designed to be neutral. The speakers did not express their feelings nor did they state any opinions that would affect the feelings of the other person.

To evaluate the influence of the manipulation, the test subjects were to answer 36 questions of a questionnaire that was created based upon other studies that presented ways of testing for rapport and presence.

3.1 Experiment Setup

The setup of the experiment was a computer system which had a headphone or speaker system on which the audio was played. The environment in which the experiment was conducted was made as little distractive as possible.

The experiment itself had thirty subjects listening to one of the three conversations, ten subjects for each of the three differently manipulated versions of the conversation. The group of subjects was a homogenous group of people, existing of all male subjects between the ages of 18 and 27 and studying a technical study.

The experiment was purely based on speech, so the agents had no visual representation.

3.2 TTS engines

To generate the experiment, a TTS engine was required. Five mainstream TTS engines have been tested for suitability, namely:

- Open MARY
- Loquendo TTS
- NaturalReader
- MS SAPI
- Fluency TTS

The TTS engines have been evaluated on the following objective points, in descending order of importance:

1. Availability of Dutch voices
2. Prosodic manipulation available in interface (especially pause control)
3. Client user interface availability
4. License availability
5. Windows OS compatibility

Since the test subjects were all native Dutch speakers, the availability of Dutch voices was of highest priority requirement when choosing a TTS engine. The hearing quality of the voices was a subjective point on which the TTS engines have been evaluated. Should the Dutch voices have sounded far inferior to English voices, the experiment would have been conducted in English. Also, engines with better voice quality would give preference to others.

Loquendo TTS and Fluency TTS were the only engines with Dutch voices available. Since the voice quality of the Dutch Loquendo TTS voices were not significantly inferior to the
English voices of other engines, the choice was made to conduct the experiment using Dutch voices.

A close second requirement was the availability of prosodic manipulation using a client user interface. The advantage of such an interface is that no programming is required to generate voice samples. Since Fluency TTS was only available using a demonstration license, it had no interface that allowed for pause control. The Loquendo TTS engine was fully licensed, and contained a client user interface that allowed for pause control.

Even though manual editing of the response latency is possible, it is not preferred. Since Fluency TTS’s voice quality was inferior to that of Loquendo TTS, combined with the lack of a decent client user interface, the choice was made to generate the voice samples for the experiment using Loquendo TTS.

### 3.3 Conversation contents

The transcript of the conversation had to be as neutral as possible in order to minimize manipulation on rapport caused by semantics. As stated before, the speakers did not express their feelings nor did they state any opinions that would affect the feelings of the other person.

One could argue that stating feelings or opinions would not make a difference, since the transcript of the conversation is equal in all three different versions and it does not influence prosody. Therefore, rapport would be affected equally in all three versions. However, the number of subjects in the experiment is small, and affecting rapport by stating feelings or opinions would probably have a higher possibility of rendering the experiment useless.

### 3.4 Questionnaire

To test rapport as perceived by the experiment subjects, a questionnaire was generated. One part of the questionnaire was taken from [1]. Bartneck et al. describe a number of constructs that are able to measure the perception of artificial agents, called Godspeeds. These Godspeeds are ‘Anthropomorphism’, ‘Animacy’, ‘Likeability’, ‘Perceived Intelligence’ and ‘Perceived Safety’. Since this experiment is based purely on speech, the Animacy Godspeed has been left out of the questionnaire.

Anthropomorphism refers to the extent to which non-human things are considered to be human. Perceived Safety measures the subject’s emotional state in reference to the artificial agent [1]. This construct includes 5 questions. For example, one of the questions is “Artificial – Lifelike”.

The Likeability construct also had 5 questions. One example of a question on which the subject had to judge the teacher is “Unfriendly – Friendly”.

The Perceived Intelligence measure the intelligence of the teacher as perceived by the subjects on 5 questions. One of these questions was “Unintelligent – Intelligent”.

The Perceived Safety construct was measured by 3 questions. One of these questions was “Agitated – Calm”.

Not all of these constructs may directly measure rapport of the agent as per definition of rapport. However, a more humanlike, intelligent, likeable agent to whom subjects feel safe can be expected to have higher rapport than agents that are considered machinelike, dumb and unlikeable and to whom subjects feel unsafe.

Since the flow of information in human to machine conversations is most likely to be from the artificial agent to the human, the most obvious roles of speaking artificial agents are roles that are similar to that of a teacher. For example, your navigation system telling you where to go, a robot teaching your children mathematics, your personal digital assistant warning you that you forgot to take your medicine, etc. For this reason, the questions taken from Bartneck et al. are reflected on the teacher. For example, the question “Anxious – Relaxed” was about the amount of anxiousness (1 for anxious, 5 for relaxed) of the teacher as perceived by the subject.

The other part of the questionnaire was taken from [2]. [2] describes a study in which a measure of social presence was developed. The study resulted in a set of questions arranged in 6 sub-dimensions. These are ‘co-presence’, ‘attentional allocation’, ‘perceived message understanding’, ‘perceived affective understanding’, ‘perceived emotional interdependence’ and ‘perceived behavioral interdependence’.

‘Attentional allocation’ is the attention allocated from partner to partner. The questions of this sub-dimension are left of the questionnaire by design of the experiment. For the same reason, the questions of the ‘perceived behavioral interdependence’ sub-dimension are left out. The questions of the sub-dimension ‘perceived affective understanding’ are already covered in the first part of the questionnaire.

The questions of the other three sub-dimensions are included in the questionnaire. The questions of the ‘Co-presence’ sub-dimension measure the level of peripheral or focal awareness between the interlocutors as perceived by the subject and the degree to which the subject feels isolated to the interlocutors. This construct existed of 6 questions, one of these being “The presence of the student was obvious to the teacher”.

The questions of the ‘Perceived Message Understanding’ (PMU) sub-dimension measure the extent to which the subject feels that messages between the interlocutors are understood by each other. This construct also existed of 6 questions. One example is “The teacher had difficulty understanding the student”.

Finally, the questions of the ‘Perceived Emotional Interdependence’ (PEI) sub-dimension measure the degree to which the subject feels that the emotional states of the interlocutors are affected by each other’s emotional state. This construct, too, had 6 questions. An example of such a question is “The feelings of the teacher influenced the mood of the interaction”.

<table>
<thead>
<tr>
<th>Construct</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropomorphism</td>
<td>0.92</td>
</tr>
<tr>
<td>Likeability</td>
<td>0.88</td>
</tr>
<tr>
<td>Perceived Intelligence</td>
<td>0.75</td>
</tr>
<tr>
<td>Perceived Safety</td>
<td>Not Indicated</td>
</tr>
<tr>
<td>Co-Presence</td>
<td>0.84</td>
</tr>
<tr>
<td>PMU</td>
<td>0.87</td>
</tr>
<tr>
<td>PEI</td>
<td>0.85</td>
</tr>
</tbody>
</table>
The questionnaire existed of a total of 36 questions, 18 questions from [1] and 18 questions from [2]. All questions from [1] were rated on a semantic differential scale. For example, the subject was to judge the teacher from 1 to 5 on the scale “Anxious – Relaxed”.

All questions from [2] had a five-point Likert scale, where 1 and 5 where the extreme ends. For example, 1 could mean ‘the speakers totally did not understand each other’ and then 5 would mean ‘the speakers totally understood each other’.

Table 1 shows the Cronbach’s Alpha values for each construct as indicated in their research. [1][2] The values for Anthropomorphism and Likeability have been taken for the android condition. The value taken for Perceived Intelligence is the lowest of the values given in [1]. [1] does not indicate a Cronbach’s Alpha value for the Perceived Safety construct.

The questionnaire can be found in the appendix.

4. RESULTS

In order to come to the results of the experiment, the questions for each version (diverging, no adjustment and converging) are grouped together in their (7) categories. The means and standard deviations can be found in table 2. A visual representation of these results can be found in figure 1.

Judging from figure 1, significant results are not expected to be found. To test whether or not differences exist between the three different versions within each construct, an ANOVA test was performed on each construct. A significance value of 0.05 was maintained. The ANOVA tests showed no significant difference for any of the constructs.

Table 3 shows the reliability of the scale for each construct. The constructs Perceived Safety and Perceived Message Understanding also show a sub-construct with an optimized reliability scale. The numbers within the curly brackets show which of the questions of the construct have been included in that particular sub-construct.

The Perceived Safety sub-construct had a Cronbach’s Alpha of less than 0.6, even when question 3 was removed from the construct. Therefore, ANOVA tests have been performed on the individual questions of this sub-construct. These tests did not show any significant difference between the three versions.

Since the original PMU construct has a Cronbach’s Alpha of less than 0.6, question 2 is left out of the construct and treated independently. An ANOVA test on the sub-construct did not show a significant difference between the three versions.

Table 2. Questionnaire results: mean and standard deviation.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Diverging</th>
<th></th>
<th></th>
<th></th>
<th>No Adjustment</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Converging</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Anthropomorphism</td>
<td>2.56</td>
<td>.479</td>
<td>2.28</td>
<td>.755</td>
<td>2.38</td>
<td>.643</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likeability</td>
<td>3.40</td>
<td>.533</td>
<td>3.34</td>
<td>.943</td>
<td>3.40</td>
<td>.533</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Intelligence</td>
<td>3.70</td>
<td>.662</td>
<td>3.58</td>
<td>.358</td>
<td>3.72</td>
<td>.590</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Safety</td>
<td>3.70</td>
<td>.429</td>
<td>3.20</td>
<td>.571</td>
<td>3.43</td>
<td>.353</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Co-Presence</td>
<td>3.93</td>
<td>.504</td>
<td>4.13</td>
<td>.292</td>
<td>4.18</td>
<td>.493</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Message Understanding</td>
<td>3.85</td>
<td>.338</td>
<td>4.07</td>
<td>.446</td>
<td>4.07</td>
<td>.486</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Emotional Interdependence</td>
<td>2.88</td>
<td>.875</td>
<td>3.00</td>
<td>.521</td>
<td>3.00</td>
<td>.854</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Visual representation of questionnaire results.

Table 3. Reliability of scales.

<table>
<thead>
<tr>
<th>Construct</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropomorphism</td>
<td>0.737</td>
</tr>
<tr>
<td>Likeability</td>
<td>0.818</td>
</tr>
<tr>
<td>Perceived Intelligence</td>
<td>0.781</td>
</tr>
<tr>
<td>Perceived Safety</td>
<td>0.094</td>
</tr>
<tr>
<td>» Perceived Safety {1, 2}</td>
<td>0.550</td>
</tr>
<tr>
<td>Co-Presence</td>
<td>0.701</td>
</tr>
<tr>
<td>PMU</td>
<td>0.597</td>
</tr>
<tr>
<td>» PMU {1, 3, 4, 5, 6}</td>
<td>0.623</td>
</tr>
<tr>
<td>PEI</td>
<td>0.800</td>
</tr>
</tbody>
</table>

Performing individual ANOVA tests on the question that was left out of the PMU construct showed that there was a result. Using t-test, it was showed that the mean of version DI (3.3) was significantly less than the mean of version NA (4.3):
t(18) = -2.518, p = 0.022. The particular question can be found in the appendix.

5. DISCUSSION
It is shown that there is little to no significance in results in the three different versions of the experiment. The expectation was that the CO version had a (significantly) higher rapport than the version DI and a higher rapport than NA version.

The lack of results may be due to the amount of experimental subjects in relation to the impact of the manipulation of just one aspect of prosody. As described in section 2, prosodic features overlap. Therefore, the effect of only modifying response latency may not be measurable using only 30 subjects.

The reliability of the scales show that 6 out of 7 constructs are properly designed to measure their aspects, so the lack of results it is not assumed to stem from the questionnaire being invalid.

By increasing the number of subjects, the variance is expected to decrease. Judging from the means and standard deviations in figure 1, increasing the amount of subjects could have a positive impact on the constructs from [2] (Co-Presence, PMU and PEI) as the standard deviations would decrease. However, this should be tested in another study.

Another possible reason for the lack of significant differences between the different versions was put forwards by Hauber et al. [3]. In that study there were no significant results on social presence measures found. Their conclusion is that social presence measures might not be sensible enough to show significant differences.

In this research, the only construct not having a sufficient reliability is Perceived Safety. In [1], no Cronbach's Alpha value is provided for this construct. A study might be performed to test whether or not this construct is reliable enough to be used in questionnaires.

Even though the results of this study show no data to prove that there is a significant difference between the three versions, there were multiple comments from subjects listening to the DI version that the teacher was slow and annoying.

6. CONCLUSION
The experiment performed in this study shows no significant difference in rapport as perceived by human experimental subjects due to manipulation of response latency. Neither a positive nor a negative influence can be shown due to the manipulation.

Response latency is one of the features that together are known as prosody. Partly, response latency results from processing a spoken sentence from another person and generating a proper response. However, response latency is also a suprasegmental feature accompanied in speech, so it serves to help transmit the message from speaker to listener.

Research on speech synthesis has made for the availability of TTS Engines. In this research, 5 major TTS Engines have been investigated: Open MARY, Loquendo TTS, NaturalReader, MS SAPI and Fluency TTS. Not all of these engines supported Dutch speech synthesis. Also, for this research there was not a full license available for every engine. Speech quality did not significantly differ between the best sounding Dutch voice and the best sounding English voice. Only Loquendo TTS and Fluency TTS had Dutch voices available. The voice quality was worst in Fluency TTS. Combined with the fact that a full license was not available, Loquendo TTS was selected as best suited for this study.

The questionnaire that the subjects were asked to answer consisted of a combination of constructs from two studies that designed questionnaire to test rapport. Seven constructs have been incorporated in this study, namely Anthropomorphism (extend to which non-human things are considered to be human), Likeability, Perceived Safety, Co-Presence, Perceived Intelligence, Perceived Message Understanding and Perceived Emotional Understanding.

It would be interesting to test the results of manipulation of the response latency on a larger scale. Especially the results to the questions of the constructs Co-Presence, PMU and PEI show a positive upward trend. However, due to the small amount of test-subjects, the standard deviations were too large to proof influence of response latency manipulation.

7. REFERENCES

APPENDIX - Questionnaire

The questions of the questionnaire are here ordered by construct. The original questionnaire had the questions mixed up.

Rate the teacher (the male voice) on a scale of 1 to 5 on the following items:

**Anthropomorphism**
1) Fake – Natural
2) Moving rigidly – Moving elegantly
3) Unconscious – Conscious
4) Machinelike – Humanlike
5) Artificial – Lifelike

**Likeability**
1) Dislike – Like
2) Unpleasant – Pleasant
3) Unfriendly – Friendly
4) Awful – Nice
5) Unkind – Kind

**Perceived Intelligence**
1) Incompetent – Competent
2) Irresponsible – Responsible
3) Foolish – Sensible
4) Ignorant – Knowledgeable
5) Unintelligent – Intelligent

**Perceived Safety**
1) Anxious – Relaxed
2) Agitated – Calm
3) Quiescent – Surprised

Answer the following questions on a scale of 1 to 5, where 1 is “totally agree” and 5 is “totally disagree”.

**Co-Presence**
1) The student noticed the teacher.
2) The presence of the student was obvious to the teacher.
3) The student caught the teacher’s attention.
4) The presence of the teacher was obvious to the student.
5) The teacher noticed the student.
6) The teacher caught the student’s attention.

**Perceived Message Understanding**
1) The teacher found it easy to understand the student.
2) The student had difficulty understanding the teacher.
3) The teacher had difficulty understanding the student.
4) The thoughts of the teacher were clear to the student.
5) The thoughts of the student were clear to the teacher.
6) The student found it easy to understand the teacher.

**Perceived Emotional Interdependence**
1) The attitude of the student influenced the feelings of the teacher.
2) The student was influenced by the teacher’s mood.
3) The teacher’s mood influenced the feelings of the student.
4) The feelings of the teacher influenced the mood of the interaction.
5) The teacher was influenced by the student’s mood.
6) The feelings of the student influenced the mood of the interaction.