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

Beyond the Listening Test: An interactive approach to TTS Evaluation

Joseph Mendelson[1], *Matthew Aylett*[2]

[1]KTH Royal Institute of Technology, Sweden
[2]CereProc Ltd. and University of Edinburgh, UK
josephme@kth.se, matthewa@cereproc.com

Abstract

Traditionally, subjective text-to-speech (TTS) evaluation is performed through audio-only listening tests, where participants evaluate unrelated, context-free utterances. The ecological validity of these tests is questionable, as they do not represent real-world end-use scenarios. In this paper, we examine a novel approach to TTS evaluation in an imagined end-use, via a complex interaction with an avatar. 6 different voice conditions were tested: Natural speech, Unit Selection and Parametric Synthesis, in neutral and expressive realizations. Results were compared to a traditional audio-only evaluation baseline. Participants in both studies rated the voices for naturalness and expressivity. The baseline study showed canonical results for naturalness: Natural speech scored highest, followed by Unit Selection, then Parametric synthesis. Expressivity was clearly distinguishable in all conditions. In the avatar interaction study, participants rated naturalness in the same order as the baseline, though with smaller effect size; expressivity was not distinguishable. Further, no significant correlations were found between cognitive or affective responses and any voice conditions. This highlights 2 primary challenges in designing more valid TTS evaluations: in real-world use-cases involving interaction, listeners generally interact with a single voice, making comparative analysis unfeasible, and in complex interactions, the context and content may confound perception of voice quality.
Index Terms: TTS evaluation, subjective evaluation, listening tests, interactive virtual agents, user experience, unit selection, statistical parametric speech synthesis, expressive speech synthesis, voice interface design, human-computer interaction

1. Introduction

As we incorporate speaking devices into our lives with increasingly sophisticated and human-like capabilities, systems begin to enter the social domain. This means, for many applications, expressive, emotional and conversational speech synthesis are required [1]. Subsequently, text-to-speech (TTS) researchers are attempting to push synthetic speech ever-closer to these human-like realms. The primary tool researchers use to subjectively evaluate their progress is the 'listening test', in which participants, generally using headphones, evaluate synthetic speech generated by various systems, and rate them on various scales [2–4]. This method has become the standard, but it is not clear that it is the best tool; it utilizes an artificial construct that bears little resemblance to the actual user experience (UX) [2], especially for interactive products.

In this paper we investigate an alternative methodology for TTS evaluation, by taking the listener out of the rarefied construct of a traditional listening test, and inserting them into an imagined end use, where they interact with an intelligent, speaking, embodied agent. For comparison, we also conduct a standard, audio-only listening test as a baseline. We ask participants in both experiments to rate the speech they heard on the same scales, and we compare the results. In the interactive study, we also investigate cognitive and affective responses to different voice systems, as a means to explore human reactions to changes in TTS voice quality that cannot be revealed under traditional listening test conditions.

Three primary voice categories of speech are evaluated: Natural Speech, Unit Selection Synthesis (US), and Statistical Parametric Speech Synthesis (SPSS) using a Deep Neural Network model (DNN). For each of these 3 categories, we generate both a 'Neutral' and an 'Expressive' version of the voice, for a total of 6 Voice Conditions. More details are provided in section 3.

2. Background

Wester et al [5] made a close inspection of listening test practices based around the 2014 Blizzard Challenge, a well known annual TTS research event [3, 6], and found significant deficiencies in many research teams' methodology. In [2], King explicitly notes the ecological invalidity of traditional listening tests, naming their 'idealised environment' as the most serious issue. He references one example of an attempt to 'de-idealise' the environment in an evaluation, by deliberately adding noise, which was incorporated in the 2009 version of the Blizzard Challenge [7]. Another example is [4], who devised a novel protocol for evaluating TTS in the context of audio books, which was later adopted by the Blizzard Challenge. Other researchers have looked at variations in voice synthesis in the context of interactions, but without the explicit purpose of evaluating a specific TTS-building technique or system [8–11]. In general, the subject remains under-researched, due in large part to the inherent challenges in developing evaluation methods that allow subtle differences to be discerned, while still emulating real-world use cases.

3. Experimental Elements

3.1. Experimental Overview

Two experiments were developed: A traditional audio-only listening test (the 'baseline'), and an interactive dialogue with a high-quality virtual agent (the 'avatar interaction'). In both experiments, participants listened to the 6 different Voice Conditions (see Table 1), and were asked to rate them on 5 point Mean Opinion Score (MOS) scales, for Naturalness, Affect, and Speaking Style. Participants in the interactive dialogue completed multiple additional questionnaires. There were 75 participants for the listening test, and 72 participants in the avatar interaction. There was no participant overlap between tests. In both cases, all were native English speakers, equally balanced by gender, and by UK/Non-UK place of English acquisition. In the audio-only test, all 75 participants heard all 6 voice condi-

http://dx.doi.org/10.21437/Interspeech.2017-1438

tions. In the avatar interaction, the 72 participants were divided into 6 groups of 12. Each group heard only one of the 6 voice conditions. No group of 12 had more than 7 of a single gender, or more than 7 native UK English speakers. All participants in both experiments were paid. None of the participants in either experiment had backgrounds in AI, Machine Learning, or Speech Technology.

3.2. The Voice Conditions

Six Voice Conditions were used in this report (Table 1). All voices were based on recordings of the same voice actor, a 45 year-old British Male. For conditions 3-6, an existing database was provided by Cereproc, Ltd. [12]. This consists of 3600 sentences based on a phonetically balanced script, for a total of 310 minutes of recorded speech. These were recorded in a 'neutral' speaking style, similar to documentary film narration. An additional 660 sentences were recorded in a 'tense' speaking style, for a total of 46 minutes of recorded speech. Via informal listening tests it was determined that these 'tense' recordings sounded 'irritated'; hereafter we will refer to them as such. All data used for synthesis was recorded with the same microphone in a professional recording studio.

3.2.1. Natural Speech

For conditions 1 and 2, the Natural Speech, the voice actor who performed the original database was brought to a professional recording studio and directed to perform the 100 prompts for the project (see section 3.4) in both Neutral and Irritated styles. For reference, he was played samples of his database recordings, and every attempt was made to match that voice tonality, style, and effort. This produced Natural voice conditions that are directly comparable to the synthesis systems.

3.2.2. Unit Selection

For conditions 3 and 4, the CereVoice system was used to generate the required prompts. This is considered to be a good example of a high-quality Unit Selection voice, and as such it is currently in use in multiple commercial products. To generate the expressive speech, this system uses mark-up tags to bias the unit selection algorithm to first look for appropriate units in the expressive portion of the database, backing off to the neutral database when it cannot find a sufficiently low-cost unit, as in [13]. Additionally, Digital Signal Processing is applied to further refine the expressivity [14]. This technique works better on certain expressive qualities for some voice data than for others, based on how the original voice actor performed during recording, and the inherent qualities of their 'normal' voice [14]. Informal in-lab testing determined that the default CereVoice settings for a 'cross' voice had the best combination of low-artifacting and high expressivity.

3.2.3. DNN Parametric

For conditions 5 and 6, the DNN parametric system was built with the open source Merlin toolkit [15]. Both conditions used the same DNN architecture: the acoustic models had 4 feedforward *tanh* layers of 1024 neurons, plus 2 Simplified-LSTM (SLSTM, [16]) layers of 512 neurons. Duration models had 2 feedforward *tanh* layers of 512 neurons and 2 SLSTM layers of 512 neurons. The Festival toolkit was used as the 'front end' [17], to generate linguistic specifications for each sentence in the database, including a phonetic transcription. The Hidden Markov Toolkit (HTK) [18] was then used to perform a 'forced alignment', generating duration data for each of these phones. The WORLD vocoder [19] was used to parameterise the speech waveforms (MGCs, BAPs and log F0s), and the output of the DNN predicted these parameters, which are then input to the vocoder to generate the final speech output. To generate expressive speech, the data from the expressive portion of the database was tagged with a feature at the input of the network during training, and aligned with its associated expressive acoustic output. This allowed the model to differentiate, and hence 'learn' to predict the acoustic properties of the expressive speech. The same expressive tag was used during synthesis to generate the expressive prompts.

3.3. The Avatar

The avatar was provided by Speech Graphics, Ltd. [20], who specialize in procedural animation of facial features driven by speech audio [21, 22]. Using proprietary techniques, acoustic features are extracted in real time and used to trigger specific facial movements. This real-time capability allowed us to easily switch between any of our 6 Voice Conditions, with the avatar's speech articulation animating in perfect synchronicity, creating a high degree of realism. Speech Graphics' software enables fine control of an array of non-verbal facial features. By hand-tuning these features, we restricted the avatar's perceived emotional range, while simultaneously keeping his appearance life-like. This was a key constraint in our multi-modal interaction, in order to minimize the confounding effects of the visual facial expression versus the expressivity we were comparing in the Voice Conditions [8, 23].

3.4. The Dialogue Prompts

The avatar's side of the experimental dialogue was composed by the researchers, consisting of 100 sentences and/or phrases (hereafter 'prompts'). Sixty-one of these constitute a structured progression of questions, answers and statements, based on the premise that the avatar would lead and guide the conversation. By moving through these prompts in the same order, all participants could have very similar interactions. The remaining 39 prompts consisted of short phrases that the researcher could use at his discretion, to maintain the smoothness of the conversation. These include discourse markers, such as 'OK', 'That's interesting', 'So…', 'Well…'; short general answers, such as 'Yes', 'No', 'Maybe', 'I don't know'; backchannels, such as 'Wow', 'Oh', 'Ahh'; interruption-handling phrases, such as 'Excuse me, please go on', 'I'm sorry, you were saying?'; and deflections, such as 'I don't have an answer for that', 'We will have to talk about that another time', 'Can I ask you another question now?'. The dialogue moves through several topics, comprised of a few prompts each. It starts with introductions, as if meeting an inquisitive stranger, such as 'What is your name?', and 'where are you from?'. It then moves through discussions of food, music, philosophy, movies, loneliness, and memory. All 100 prompts were synthesized for conditions 3-6 via their respective systems, and recorded as natural speech for conditions 1-2, as described in Section 3.2.1, to create the 6 Voice Conditions.

4. Methodology

4.1. The Baseline Listening Test

Twelve of the 100 dialogue prompts were chosen, for each of the 6 Voice Conditions, for a total of 72 prompts. They

were intentionally selected to be spread from throughout the 61 prompts of the structured dialogue section, but with the constraint that they could not be consecutive prompts, or have any clear semantic continuity between them. This was to insure that listeners would hear a representative sampling of the same prompts as the interactive participants, but without any confounding factors associated with the underlying meaning of the dialogue. To minimize ordering effects, the order of all prompts were randomized for each listener, in 2 groups of 36. Within each group of 36, each listener heard 6 prompts in each of the 6 conditions. They were then asked to rate each prompt on 5-point scales for 'Naturalness', 'Emotional Character - Negative-to-Positive', and 'Speaking Style - Irritated-to-Calm'. The test was performed online, using the Qualtrics survey software [24], with paid participants via the Prolific Academic crowd-sourcing platform [25]. In order to ensure that participants listened to the samples and to minimize 'cheating' as in [26], participants were asked to state which brand of headphones they were using. Additionally, since natural speech was included in 2 of the 6 voice conditions, these served as 'gold standard' anchors - any listener who rated these as un-natural would be discarded. Listeners were equally distributed between male/female, and UK/Non-UK Native English speakers. Each participant was paid £2, and spent 20 minutes completing the test.

4.2. The Avatar Interaction

Figure 1: *The researcher's WoZ setup on the left, the participant setup on the right*

The interaction was performed 'Wizard of Oz' (WoZ) style, with the researcher controlling the avatar's speech from a hidden location, via midi keyboard (Figure 1). Participants were seated facing the avatar on a screen, with a speaker hidden behind it (Figure 1). The researcher began the interaction by pressing the first key on the midi keyboard, which triggers the prompt 'Hello'. From this point onward, each participant progressed through the dialogue with the avatar. Each interaction was unique; participants responded to questions and statements posed by the avatar in their own way, and some asked questions in return. The structured nature of the experimental setup insured that all participants heard c.95% of the primary 61 prompts. Some prompts were missed by some participants in cases where their own response necessitated skipping ahead to maintain the illusion that the avatar could understand them, and to maintain conversational flow. After the interaction, participants were asked to complete several questionnaires:

- The 3 rating scales to evaluate the voice, as described in Section 4.1
- The Positive and Negative Affect Schedule (PANAS) [27].
- The Godspeed Questionnaire, reduced to Anthropomorphism, Animacy, and Likeability, and Intelligence [28].
- A recall test of specific preferences stated by the avatar at the beginning, middle, and end of the conversation.

Each participant was paid £10, and spent c.30 minutes completing the study.

5. Results

Mann-Whitney statistical significance tests were used to compare all MOS scores. Bonferroni corrections were applied as needed for repeated tests of the same data sets. Results from both experiments are combined where possible for efficiency.

Table 1: *The 6 voice conditions and their abbreviations, used in all subsequent figures*

Condition	Description/System	Abbreviation
Voice 1	Natural, Neutral Speech	NN
Voice 2	Natural, Irritated Speech	NI
Voice 3	Unit Selection - Neutral	USN
Voice 4	Unit Selection - Irritated	USI
Voice 5	Parametric/DNN - Neutral	DNN-N
Voice 6	Parametric/DNN - Irritated	DNN-I

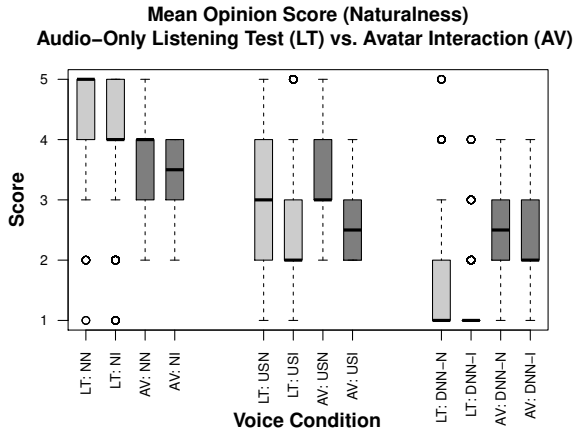

Figure 2: *MOS scores for Naturalness across all conditions, Listening Test (LT) in light grey, Avatar Interaction (AV) in dark grey. Results were significant in both experiments, but with much larger effect size in LT.*

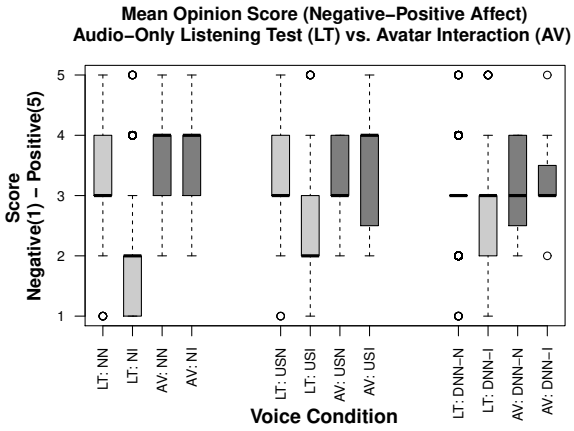

Figure 3: *MOS scores for Neg-Pos Affect across all conditions, Listening Test (LT) in light grey, Avatar Interaction (AV) in dark grey. LT shows significant difference between expressions, AV does not.*

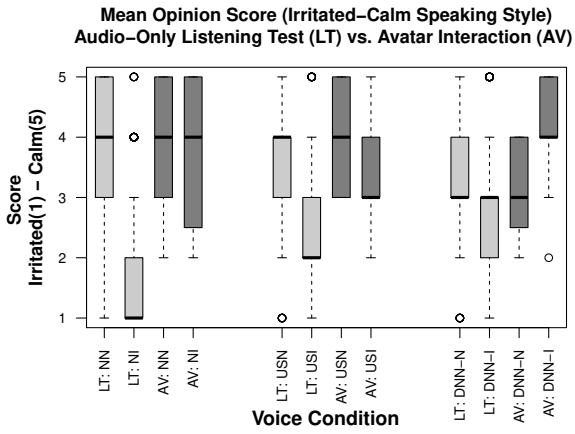

Figure 4: *MOS scores for Speaking Style Irritated-Calm across all conditions, Listening Test (LT) in light grey, Avatar Interaction (AV) in dark grey. LT shows significant difference between expressions, AV does not, except for the DNN, in the opposite direction.*

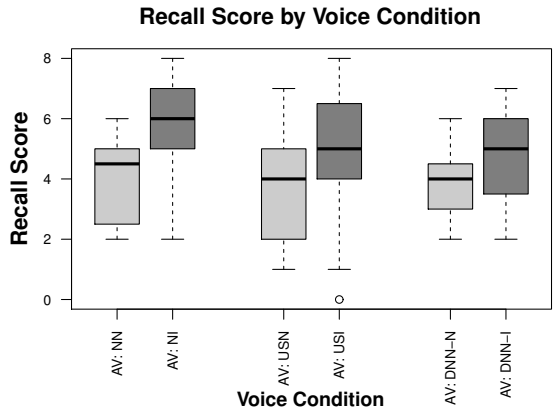

Figure 5: *Participant recall scores from post-interaction memory test, by voice condition, Neutral condition in light grey, Irritated condition in dark grey. There is a non-significant trend that Irritated condition lead to higher recall scores (See Discussion).*

5.1. Audio-Only Listening Test: Baseline

MOS 'Naturalness' scores show the canonical, expected ranking of Natural speech highest, then Unit Selection, then Parametric. In every pairwise comparison, we see highly significant results: $p < .01$ with Bonferroni corrections. (Figure 2).

Listeners also rated Neutral speech as significantly more positive, and more calm, than Irritated speech, across all 3 systems, firmly establishing our baseline. The effect size of this difference is highest for Natural, then US, then DNN. All pairwise comparisons between systems were highly significant: $p < .01$ with Bonferroni corrections. (Figures 3, 4).

5.2. Comparison of Listening Test to Avatar Interaction

Interaction participants rated Natural speech as most natural, followed by Unit Selection, then Parametric (Figure 2). Significant difference is found between Natural and US, and US and DNN: $p < .05$ with Bonferroni corrections. Effect size is smaller overall compared to the baseline. For Affect, Interaction participants on average rated all voice conditions as neutral-positive, and for Style as neutral-calm (Figures 3, 4). In some cases, the Irritated version of the system was rated as more positive, and/or more calm - the opposite of our baseline. For the DNN system, this effect was significant for speaking style: $p < .05$ with Bonferroni correction (Figure 4). Neither system nor expressive quality had significant effect on ratings in any other pairwise comparisons: $p > .05$. Moreover, no correlations were found between *any* of the voice conditions and any of the cognitive (recall), affective (PANAS), or perceptual (Godspeed) responses reported by participants. In the case of recall, there was a tendency for more expressive voices to induce higher recall scores, but the effect was not significant (Figure 5).

6. Discussion

The MOS naturalness results for the interactive study validate our central theme: TTS evaluation *can* be achieved through innovative, end-use style testing. Participants could make judgements regarding the naturalness of the voice, despite only hearing one condition, that corresponded to the baseline listening test, albeit with smaller effect size.

It is important to note that in the baseline listening test, listeners heard more than one condition, which enhances their ability to discriminate between different systems, as well as different kinds of expressivity. They also used headphones (rather than a speaker), and evaluated each utterance directly after hearing it (rather than evaluating the entire interaction). We believe these methodological differences contributed to the smaller effect size in the interactive participants' MOS naturalness scores. (However, a typical commercial interactive device is likely to only use one voice, played through a speaker, so in that sense our UX *was* realistic). These differences may also have contributed to the interaction participants' inability to discern differences in expressivity. (Figures 3, 4). It is particularly likely that the *unchanging* nature of the expressivity played a role. While easily discernible by our listening test participants, in the context of a complex dialogue, our interaction participants simply accepted the expressive nature of the voice as 'normal'. This implies that voices that can dynamically change their expressivity should be considered in interface deisgn.

We also postulate that the very nature of the interaction experiment itself contributed to the less discriminatory results compared to the baseline. It seems likely that participants were engaged by the interaction to the point where their ability to isolate and evaluate the voice quality alone was confounded. This has implications for interaction design in that voice quality concerns may be relative to the content and context of the interaction. Higher quality voices may be required for simple interactions, but voice quality may wane in importance as user experiences become more complex.

Lastly, we note the findings shown in Figure 5. While not significant, they show a promising direction: expressivity in voices in general, *regardless of valence of affect*, seems to have an impact on aspects of engagement, as measured by recall. This seems counterintuitive, and therefore makes it a prime area for future study.

7. Conclusions

We have shown that alternative methodologies for TTS evaluation are viable, although by attempting to simulate real-world scenarios, we face new challenges: multi-system comparisons become more difficult, and the content and context of a scenario may confound participants' evaluative discernment.

8. References

[1] M. P. Aylett, P. O. Kristensson, S. Whittaker, and Y. Vazquez-Alvarez, "None of a CHInd: relationship counselling for HCI and speech technology," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 26 Apr. 2014, pp. 749–760.

[2] S. King, "Measuring a decade of progress in Text-to-Speech," *Loquens*, vol. 1, no. 1, p. e006, 30 Jun. 2014.

[3] A. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," *Proceedings of interspeech*, 2005.

[4] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Proc. Blizzard Challenge Workshop*. International Speech Communication Association (ISCA), 2011.

[5] M. Wester, C. Valentini-Botinhao, and G. Henter, "Are we using enough listeners? no! an empirically-supported critique of interspeech 2014 tts evaluations," in *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 9 2015, pp. 3476–3480.

[6] S. King, "The blizzard challenge 2016," *Proc. Blizzard Challenge Workshop, Cupertino, CA*, 2016.

[7] S. King and V. Karaiskos, "The blizzard challenge 2009," in *Proceedings of Blizzard Challenge Workshop*, 2009.

[8] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press, 2007.

[9] S. Hennig and R. Chellali, "Expressive synthetic voices: Considerations for human robot interaction," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012, pp. 589–595.

[10] R. K. Atkinson, R. E. Mayer, and M. M. Merrill, "Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice," *Contemp. Educ. Psychol.*, vol. 30, no. 1, pp. 117–139, 2005.

[11] R. E. Mayer, K. Sobko, and P. D. Mautone, "Social cues in multimedia learning: Role of speaker's voice," *J. Educ. Psychol.*, vol. 95, no. 2, p. 419, Jun. 2003.

[12] "Welcome to CereProc — CereProc Text-to-Speech," https://www.cereproc.com/, accessed: 2016-8-10.

[13] G. O. Hofer, K. Richmond, and R. A. J. Clark, "Informed blending of databases for emotional speech synthesis," in *Proceedings, Interspeech'2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal*. International Speech Communication Association, 2005.

[14] M. P. Aylett and C. J. Pidcock, "The CereVoice characterful speech synthesiser SDK," in *Pelachaud C., Martin JC., André E., Chollet G., Karpouzis K., Pelé D. (eds) Intelligent Virtual Agents. IVA 2007*. Lecture Notes in Computer Science, vol 4722. Springer, Berlin, Heidelberg, 2007.

[15] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." *Proc. SSW, Sunnyvale, USA*, 2016.

[16] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis." *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5140–5144, 2016.

[17] R. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system." *Speech Commun.*, vol. 49, pp. 317–330, 2007.

[18] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering, 1993.

[19] "WORLD," http://ml.cs.yamanashi.ac.jp/world/english/, accessed: 2016-8-14.

[20] "Speech graphics facial animation," https://www.speech-graphics.com/, 21 Apr. 2014, accessed: 2016-8-10.

[21] M. A. Berger, "Static and dynamic 3-D human face reconstruction," Patent 7 239 321, 3 Jul., 2007.

[22] M. A. Berger, G. Hofer, and others, "Carnival—combining speech technology and computer animation," *IEEE Comput. Graph. Appl.*, 2011.

[23] M. P. Aylett and B. Potard, "Synthesising and evaluating Cross-Modal emotional ambiguity in virtual agents," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds. Springer Berlin Heidelberg, 12 Sep. 2012, pp. 471–473.

[24] "Qualtrics," Provo, Utah, USA.

[25] Prolific, "Prolific – find participants fast," https://prolific.ac/, accessed: 2016-8-10.

[26] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating." in *INTERSPEECH*. ISCA, 2011, pp. 3053–3056.

[27] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales," *J. Pers. Soc. Psychol.*, vol. 54, no. 6, pp. 1063–1070, Jun. 1988.

[28] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Adv. Robot.*, vol. 1, no. 1, pp. 71–81, 20 Jan. 2009.