Practical Evaluation of Human and Synthesized Speech for Virtual Human Dialogue Systems

Kallirroi Georgila[†], Alan W. Black[‡], Kenji Sagae[†], David Traum[†]

[†]Institute for Creative Technologies, University of Southern California
[‡]Language Technologies Institute, Carnegie Mellon University
{kgeorgila,sagae,traum}@ict.usc.edu, awb@cs.cmu.edu

Abstract

The current practice in virtual human dialogue systems is to use professional human recordings or limited-domain speech synthesis. Both approaches lead to good performance but at a high cost. To determine the best trade-off between performance and cost, we perform a systematic evaluation of human and synthesized voices with regard to naturalness, conversational aspect, and likability. We also vary the type (in-domain vs. out-of-domain), length, and content of utterances, and take into account the age and native language of raters as well as their familiarity with speech synthesis. We present detailed results from two studies, a pilot one and one run on Amazon's Mechanical Turk. Our results suggest that a professional human voice can supersede both an amateur human voice and synthesized voices. Also, a high-quality general-purpose voice or a good limited-domain voice can perform better than amateur human recordings. We do not find any significant differences between the performance of a high-quality general-purpose voice and a limited-domain voice, both trained with speech recorded by actors. As expected, in most cases, the high-quality general-purpose voice is rated higher than the limited-domain voice for out-of-domain sentences and lower for in-domain sentences. There is also a not statistically significant trend for long or negative-content utterances to receive lower ratings.

Keywords: speech synthesis, professional and amateur human recordings, dialogue systems, virtual humans

1. Introduction

Virtual humans are artificial conversational agents designed to mimic the behavior of real humans (Gratch et al., 2002; Swartout et al., 2006; Traum, 2008; Traum et al., 2008). The perceived naturalness of a virtual human is a mixture of aspects, such as appearance, voice, gestures, etc. In fact studies have shown that behavior can be more important than appearance (Cassell and Tartaro, 2007). In this paper we focus on the aspect of voice. In order for virtual humans to successfully simulate real humans, they need to sound natural and also give the impression that they are engaged in the conversation when interacting with real humans.

When building a virtual human dialogue system, a major decision is whether to use human recordings or synthesized speech. The advantage of using human recordings is obviously performance. But this high quality comes at the high cost of hiring a professional actor to record the lines that the virtual human will utter. Furthermore, once we decide to expand the system by adding more lines, the actor has to be hired again to record the new lines. The alternative is to use speech synthesis. Current state-of-the-art speech synthesizers have reached high levels of naturalness and intelligibility for neutral read aloud speech. However, synthesized speech generated using neutral read aloud data lacks all the attitude, intention, and spontaneity associated with everyday conversations (Andersson et al., 2010). Note that building a synthesized voice usually requires hiring an actor to record a large number of sentences plus the additional effort of processing these recordings. To improve the quality of speech synthesis in a particular domain, an alternative approach to general-purpose speech synthesis is limited-domain speech synthesis (Black and Lenzo, 2000). A limited-domain synthesized voice is trained using material from the domain where it will be deployed in order to achieve high-quality speech synthesis within this domain. Note that recently there have been attempts to build synthesized voices especially for conversational purposes (Andersson et al., 2010; Marge et al., 2010; Andersson et al., 2012).

For the above reasons, the current practice in virtual human dialogue systems (especially systems that are targeted to real users) is to use professional human recordings or limited-domain speech synthesis. The problem is that, as previously discussed, both approaches can be very costly and time consuming. The question that arises is whether the performance of professional human recordings or limited-domain synthesized speech justify their cost. Are professional human recordings and limited-domain synthesized voices better than amateur human recordings or general-purpose synthesized voices? And if yes, how much better are they? Furthermore, how does the performance of human and synthesized voices vary depending on the length and content of the utterance? Also, what is the importance of the age of the listeners, their native language, and their familiarity with speech synthesis?

To our knowledge no one has systematically addressed these issues. There has been some work on comparing the effect of synthesized vs. human speech on the interaction with a dialogue system, e.g. a virtual patient dialogue system (Dickerson et al., 2006) and an intelligent tutoring dialogue system (Forbes-Riley et al., 2006), but none of these studies has compared professional vs. amateur human recordings or limited-domain vs. general-purpose synthesized voices. Moreover, none of these studies has considered the effect of the type (in-domain vs. out-of-domain), length, and content of utterances on performance.

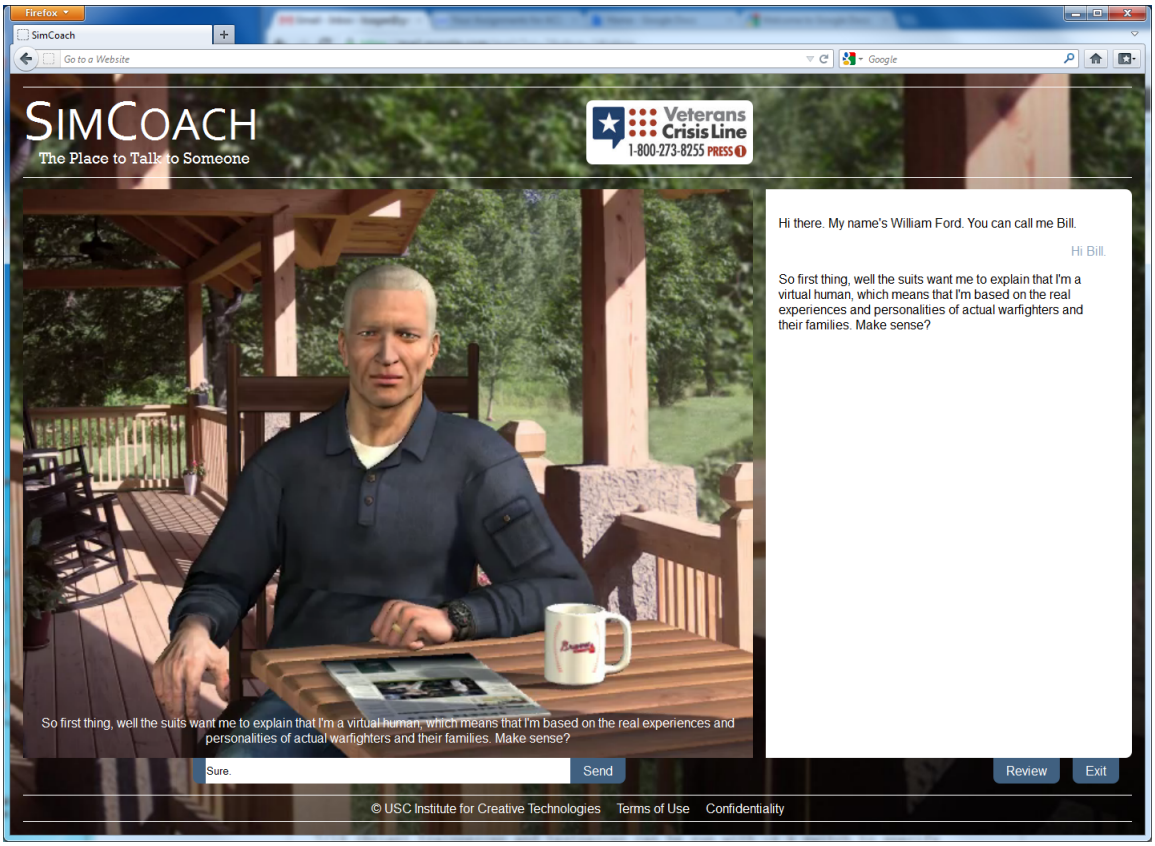

Figure 1: The Simcoach web interface.

In this paper we aim to answer these questions. We perform two studies, one small-scale pilot study and one large-scale study on Amazon's Mechanical Turk (MTurk, `http://www.mturk.com`), and compare a variety of human and synthesized voices based on several criteria, i.e. naturalness, conversational aspect, and likability. We also vary the type (in-domain vs. out-of-domain), length, and content of utterances (see section 2), and we take into account the age and native language of the participants as well as their familiarity with speech synthesis.

The paper is structured as follows. In section 2, we describe our corpus of sentences that were spoken by humans or synthesized by speech synthesis engines. In section 3 we present the voices that we compared. Section 4 is about our experimental setup and sections 5 and 6 present our results. Finally, section 7 discusses our findings and section 8 concludes.

2. Our Corpus of Sentences

Our virtual human dialogue system called Simcoach (Rizzo et al., 2011) aims to motivate military personnel and family members to take the first step and seek information and advice with regard to their health care, in particular depression and post-traumatic stress disorder. The system and the user communicate with natural language. More specifically, the user types text and the system responds using speech and non-verbal behavior. The current Simcoach system uses professional human recordings for the system utterances. In Figure 1 we can see the web interface of Simcoach.

We selected 200 sentences in the Simcoach domain, out of which 100 were included in the data set used for training our two limited-domain synthesized voices (see section 3). We also selected 30 out-of-domain sentences. An example of an in-domain sentence is "well I'm just trying to get some info so I can help you better" and an example of an out-of-domain sentence is "this TV show is hilarious, don't you think so?". The sentences (both in-domain and out-of-domain) varied in length (long, i.e. > 5 words and short, i.e. up to 5 words) and content (positive, neutral, and negative). An example of positive content is "that was the best meal ever", an example of neutral content is "I prefer reading the news online", and an example of negative content is "it makes me feel unhappy and hopeless". To measure inter-annotator agreement, two experienced annotators annotated 400 sentences for their content. For 332 sentences (83%) there was agreement between the annotators, i.e. the annotators agreed that these sentences had positive or neutral or negative content. From the sentences where there was disagreement (68 in total, 17%) only 4 were short in length. For our evaluation we selected only sentences where there was agreement between the annotators. In Table 1 we can see the distribution of sentences used in our evaluation with respect to the domain, length, and content. In our pilot study we used only sentences with positive content.

3. Human and Synthesized Voices

We compare professional and amateur human voices and 4 synthesized voices (2 general-purpose and 2 limited-

Abbreviation	Domain	Used for training the limited-domain voices?	Length	Content	Number of sentences
dom-train-lg-pos	in-domain	yes	long	positive	30
dom-train-lg-neu	in-domain	yes	long	neutral	30
dom-train-lg-neg	in-domain	yes	long	negative	20
dom-train-sh-pos	in-domain	yes	short	positive	10
dom-train-sh-neu	in-domain	yes	short	neutral	10
dom-notrain-lg-pos	in-domain	no	long	positive	35
dom-notrain-lg-neu	in-domain	no	long	neutral	35
dom-notrain-lg-neg	in-domain	no	long	negative	30
gen-lg-pos	out-of-domain	no	long	positive	6
gen-lg-neu	out-of-domain	no	long	neutral	5
gen-lg-neg	out-of-domain	no	long	negative	5
gen-sh-pos	out-of-domain	no	short	positive	4
gen-sh-neu	out-of-domain	no	short	neutral	5
gen-sh-neg	out-of-domain	no	short	negative	5

Table 1: Our corpus of sentences.

domain ones). One professional actor and one speech synthesis expert recorded approximately 1300 sentences in our Simcoach domain plus some general sentences to improve coverage. These recordings were used for building our two limited-domain unit-selection synthesized voices (Black and Taylor, 1997; Black and Lenzo, 2000). More specifically, for our studies we use the following human and synthesized voices:

- Professional actor's voice (PROF): The 100 in-domain sentences of our corpus (see section 2) were recorded by a professional actor. These were a subset of the recorded sentences used for training the first limited-domain voice (see below).
- Amateur person's voice (AMAT): The 230 sentences of our corpus were recorded by an amateur, neither actor nor speech synthesis expert. Due to some problems with these amateur recordings (low audio volume) we only kept 67 in-domain utterances for our pilot study.
- High-quality general-purpose voice (GEN-HIGH): The 230 sentences of our corpus were synthesized using a custom high-quality commercial unit-selection voice developed by CereProc Ltd (Aylett et al., 2006) for USC/ICT. This voice was trained on material unrelated to our domain. This material was recorded by a professional actor (different from the actor whose speech was used for our professional human recordings and one of the limited-domain voices).
- First limited-domain voice (LD1): The 230 sentences of our corpus were synthesized using a limited-domain unit-selection voice. This voice was trained with the recordings of the professional actor (PROF); approximately 1300 recorded sentences in our domain. Also, this voice works with the Flite speech synthesis engine (`http://www.speech.cs.cmu.edu/flite/`) developed at CMU.
- Second limited-domain voice (LD2): The 230 sentences of our corpus were synthesized using a second limited-domain unit-selection voice. This voice was trained with the recordings of a speech synthesis expert; approximately 1300 recorded sentences in our domain plus some additional general material to improve coverage. Again, this voice works with Flite.
- Lower-quality general-purpose voice (GEN-LOW): The 230 sentences of our corpus were synthesized using a lower-quality diphone-based voice (Microsoft Sam).

Note that for our pilot study we used only sentences with positive content, which means that we ended up with 40 utterances for PROF, 30 for AMAT, and 85 for each of the synthesized voices. For our MTurk experiment we used utterances from all content categories.

4. Experimental Setup

We performed two studies, one small-scale pilot study and one large-scale study run on Amazon's Mechanical Turk (MTurk). Below we provide details about our experimental setup in both studies.

4.1. Pilot Study

In our pilot study each participant was asked to listen to 12 utterances (2 from each human or synthesized voice category) and answer the following 3 questions on a 5-point Likert scale:

- Question 1: Does this utterance sound natural? (1=very unnatural, 2=somewhat unnatural, 3=neither natural nor unnatural, 4=somewhat natural, 5=very natural)
- Question 2: Is this an isolated prompt or taken from a conversation? (1=definitely isolated prompt, 2=maybe isolated prompt, 3=cannot decide, 4=maybe taken from a conversation, 5=definitely taken from a conversation)
- Question 3: Would you like to have a conversation with this speaker? (1=definitely not, 2=maybe not, 3=cannot decide, 4=maybe yes, 5=definitely yes)

The purpose of Question 1 is to measure the naturalness of the voice, the purpose of Question 2 is to measure whether the voice sounds conversational or not, and the intention of Question 3 is to measure whether the voice is likable or not.

4.2. Amazon Turk Study

For our MTurk study we used Questions 1 and 3, as for the pilot study. However, the participants of the pilot study reported that they found Question 2 confusing, thus for the MTurk study this question was rephrased as follows:

- Question 2: Does this utterance sound more like in an everyday conversation (as opposed to e.g. someone reading from a script)? (1=definitely not like in an everyday conversation, 2=perhaps not like in an everyday conversation, 3=cannot decide, 4=perhaps like in an everyday conversation, 5=definitely like in an everyday conversation)

In the MTurk study each participant could perform one or more HITs (Human Intelligence Tasks), and each HIT contained 5 utterances. To prevent spam HITs, if the 5 sentences of a HIT were all done in less than 30 seconds, then all the submissions in that HIT were discarded.

5. Results of the Pilot Study

Table 2 shows the results of our pilot study with 27 participants. We can see the results for all voices and for each sentence category ("dom-train-lg-pos", "dom-train-sh-pos", etc.) as well as for all sentence categories combined. We can also see the average result per voice for Questions 1-3. For explanations of the abbreviations used see Table 1. For example, "gen-sh-pos" means out-of-domain sentences, short in length, and positive in content.

As we can see in Table 2, the amateur recordings were rated higher than the professional recordings and all synthesized voices. This result was consistent across all questions. However, the difference between the professional and the amateur recordings was not statistically significant. Of the synthesized voices, overall GEN-HIGH performed best. Again this result was consistent across all questions. The limited-domain voices did better (at least in most cases) only for the in-domain sentences used for training these voices, which is not surprising. The two limited-domain voices on average performed similarly. The LOW-GEN voice performed the worst, as expected. The worst performance of the LOW-GEN voice was statistically significant for questions 1 and 3, and the average of all questions.

Once we take into account whether the participants were native speakers or not, the results change slightly. In our pilot study we had 16 native and 11 non-native speakers of English. Native speakers of English prefer the amateur recordings to the professional recordings for all questions apart from Question 3 (likability) but this result is not significant. On the other hand, non-native speakers always prefer the amateur voice (not significant). Of the synthesized voices, both native and non-native speakers prefer the GEN-HIGH voice (not significant on average). Only for Question 2 (conversational aspect) do native speakers prefer the limited-domain voices. Also, non-native speakers tend to prefer the LD2 voice to the LD1 voice (not significant). Detailed results are given in Table 3. As expected due to the small number of participants in the pilot study results are not statistically significant, which necessitates a second large-scale study (MTurk).

Note that the amateur voice, the high-quality general-purpose voice, and the second limited-domain voice (LD2) were all "young voices", whereas the professional voice and the first limited-domain voice (LD1) were "older voices". Our pilot study participants were mostly younger people. To see if this may have been an issue, in our MTurk experiment we ask participants to provide their age range.

6. Results of the Amazon Turk Study

In our MTurk study we had 826 participants who rated 24590 sentences. Table 4 shows the results of the MTurk study for all participants and all questions combined, Question 1, Question 2, and Question 3. For all questions combined we can see results for each sentence category as well as for all sentence categories combined. Here the result is rather different from the pilot study. The professional human voice is ranked as the best voice on average and this result is statistically significant. Both GEN-HIGH and LD1 are rated as significantly better on average than the amateur voice. GEN-HIGH is on average perceived as marginally better than LD1, but LD1 performs a little better than GEN-HIGH for in-domain sentences (not significant). LD1 is on average rated higher than LD2 (significant) but in individual sentence categories there is no significant difference between them. GEN-LOW is on average significantly worse than all voices. There is also a tendency for lower scores for utterances that are long in length or that have negative content but again this is not significant or is only marginally significant.

When we look into the individual questions (not just the average of all questions) the above results hold to a great extent. For Question 1 (naturalness), the order of voices in terms of scores remains the same as with all questions. In particular, PROF is rated as the best voice (significant), GEN-HIGH is perceived as better than AMAT (significant) and LD1 (significant), and LD1 is rated higher than LD2 (significant). For Question 2 (conversational aspect) PROF is perceived as the best voice (significant) and GEN-HIGH is rated as better than AMAT (not significant). LD1 is rated higher than LD2 (marginally significant) and GEN-HIGH (not significant). For Question 3 (likability) again PROF is ranked as the best voice (significant), GEN-HIGH is rated higher than AMAT (significant) and LD1 (not significant), and LD1 is rated higher than LD2 (significant). For all 3 questions, long sentences and negative-content utterances tend to receive lower scores (not significant). Also, for all 3 questions GEN-LOW performs the worst (significant).

In Table 5 we can see the results with regard to the native language of the participants. In total we had 173 US English speakers, 62 UK English speakers, 264 Indian English speakers, 303 non-English speakers, and 24 participants did not provide information about their native language. US English and Indian English scores were consistently higher than UK English scores. For US English speakers, PROF was rated as the best voice (significant)

	PROF mean CI	AMAT mean CI	GEN-HIGH mean CI	LD1 mean CI	LD2 mean CI	GEN-LOW mean CI
All questions combined						
dom-train-lg-pos	3.96 3.30-4.23	**4.06** 3.89-4.23	2.88 2.37-3.39	**3.17** 2.65-3.69	2.74 2.23-3.25	2.08 1.67-2.49
dom-train-sh-pos	**3.60** 3.33-3.87		**3.38** 2.93-3.82	3.07 2.56-3.57	2.84 2.45-3.23	2.42 1.95-2.90
dom-notrain-lg-pos			**3.07** 2.61-3.52	2.33 1.84-2.83	2.63 2.09-3.17	2.04 1.49-2.58
gen-lg-pos			**2.94** 2.53-3.36	2.36 1.99-2.73	2.77 2.28-3.25	1.78 1.38-2.18
gen-sh-pos			**2.82** 2.41-3.23	2.30 1.88-2.72	2.09 1.65-2.53	1.78 1.38-2.18
all utterances combined	3.79 3.60-3.98	**4.06** 3.89-4.23	**2.99** 2.80-3.18	2.60 2.40-2.80	2.62 2.42-2.83	2.03 1.83-2.22
Question 1 all utterances combined	3.94 3.60-4.29	**4.42** 4.17-4.68	**2.80** 2.45-3.14	2.28 1.91-2.66	2.31 1.96-2.67	1.40 1.21-1.58
Question 2 all utterances combined	3.70 3.34-4.06	**3.87** 3.54-4.19	**3.31** 2.96-3.67	3.15 2.82-3.48	3.20 2.85-3.55	2.96 2.58-3.35
Question 3 all utterances combined	3.74 3.44-4.03	**3.88** 3.60-4.17	**2.85** 2.54-3.17	2.36 2.04-2.68	2.35 2.04-2.67	1.72 1.46-1.97

Table 2: Results for all participants in our pilot study (CI: 95% confidence interval).

	PROF mean CI	AMAT mean CI	GEN-HIGH mean CI	LD1 mean CI	LD2 mean CI	GEN-LOW mean CI
native (16 raters)	4.08 3.85-4.30	**4.14** 3.94-4.35	**2.78** 2.53-3.03	2.65 2.34-2.95	2.41 2.15-2.66	1.91 1.66-2.17
non-native (11 raters)	3.39 3.08-3.71	**3.94** 3.65-4.23	**3.29** 2.98-3.59	2.53 2.29-2.77	2.94 2.61-3.27	2.18 1.88-2.49
all (27 raters)	3.79 3.60-3.98	**4.06** 3.89-4.23	**2.99** 2.80-3.18	2.60 2.40-2.80	2.62 2.42-2.83	2.03 1.83-2.22

Table 3: Results for native vs. non-native participants in our pilot study, all questions combined (CI: 95% confidence interval).

and AMAT was rated higher than GEN-HIGH (not significant). GEN-HIGH was perceived as better than LD1 (not significant), and LD1 as better than LD2 (marginally significant). GEN-LOW was rated as the worst voice (significant). For UK English speakers, PROF was also perceived as the best voice but not as significantly better than GEN-HIGH, which in turn was not perceived as significantly better than AMAT. GEN-HIGH was rated higher than LD1 (not significant). LD1 was perceived as significantly better than LD2, and GEN-LOW was rated higher than LD2 (significant). For Indian English speakers, PROF was judged as the best voice (significant), GEN-HIGH and LD1 as better than AMAT (significant), and GEN-HIGH as better than LD1 (not significant). Also, GEN-LOW performed the worst (significant). Non-native speakers (Other) preferred the PROF and GEN-HIGH voices, which were both judged as significantly better than the rest of the voices. LD1 was perceived as better than AMAT (not significant) and LD2 (significant), and GEN-LOW was rated as the worst voice (significant).

In Table 6 we can see the results with regard to the age of the participants. In total we had 14 participants under the age of 15, 344 between 15 and 25, 286 between 26 and 35, 111 between 36 and 45, 27 between 46 and 55, 19 between 56 and 65, and 1 above 65, whereas 24 participants did not provide information about their age. In all age groups, except for the age group 56-65, PROF is perceived as the best voice. GEN-HIGH is consistently rated as better than AMAT apart from the age group 56-65. Participants of the age groups 26-35, 36-45, and 46-55 prefer LD1 to AMAT, and generally the consensus is that LD1 performs better than LD2, which in turn performs better than GEN-LOW. The exception to this rule is that LD2 is perceived as slightly better than LD1 for the age group 56-65 (not significant).

In Table 7 we can see the results with regard to the familiarity of the participants with text-to-speech (TTS). In total we had 18 TTS developers, 72 participants who often use TTS, 298 who sometimes use TTS, and 414 who never use TTS, whereas 24 participants did not provide information about their familiarity with TTS. We avoided asking the question "are you a TTS expert" since people might think that an-

	PROF mean CI	AMAT mean CI	GEN-HIGH mean CI	LD1 mean CI	LD2 mean CI	GEN-LOW mean CI
All questions combined						
dom-train-lg-pos	**3.89** 3.80-3.98	3.51 3.41-3.61	**3.77** 3.67-3.87	3.75 3.66-3.85	3.63 3.53-3.73	3.22 3.11-3.32
dom-train-lg-neu	**3.83** 3.73-3.92	3.34 3.23-3.45	3.61 3.52-3.71	**3.71** 3.62-3.81	3.51 3.40-3.61	3.22 3.11-3.32
dom-train-lg-neg	3.66 3.55-3.78	3.22 2.99-3.44	3.40 3.28-3.53	**3.69** 3.58-3.81	3.29 3.16-3.42	3.26 3.14-3.38
dom-train-sh-pos	**4.13** 3.98-4.28		3.91 3.75-4.08	**4.06** 3.90-4.22	3.82 3.64-4.00	3.57 3.40-3.74
dom-train-sh-neu	3.96 3.80-4.13		**3.98** 3.82-4.14	3.91 3.75-4.07	3.69 3.51-3.87	3.10 2.92-3.28
dom-notrain-lg-pos			**3.68** 3.59-3.77	3.56 3.47-3.65	3.38 3.29-3.47	3.16 3.06-3.26
dom-notrain-lg-neu			**3.56** 3.47-3.65	3.32 3.22-3.41	3.26 3.17-3.36	3.12 3.03-3.22
dom-notrain-lg-neg			**3.37** 3.27-3.47	3.22 3.12-3.33	3.13 3.03-3.24	3.06 2.96-3.16
gen-lg-pos			**3.68** 3.45-3.90	3.38 3.15-3.61	3.55 3.32-3.78	2.99 2.75-3.23
gen-lg-neu			**3.56** 3.32-3.80	3.10 2.84-3.37	3.06 2.81-3.31	3.09 2.84-3.33
gen-lg-neg			3.14 2.88-3.40	**3.50** 3.26-3.74	3.25 3.01-3.49	3.01 2.76-3.25
gen-sh-pos			**3.82** 3.56-4.09	3.19 2.92-3.46	3.43 3.13-3.72	3.14 2.86-3.42
gen-sh-neu			3.69 3.46-3.93	3.53 3.30-3.77	**3.70** 3.45-3.95	3.36 3.10-3.61
gen-sh-neg			3.26 3.03-3.49	3.05 2.78-3.31	**3.40** 3.15-3.65	3.06 2.81-3.31
all utterances combined	**3.86** 3.80-3.91	3.40 3.33-3.47	**3.60** 3.57-3.64	3.54 3.50-3.58	3.41 3.37-3.44	3.17 3.14-3.21
Question 1 all utterances combined	**3.97** 3.89-4.06	3.43 3.30-3.56	**3.71** 3.65-3.77	3.57 3.50-3.63	3.41 3.34-3.47	3.10 3.03-3.16
Question 2 all utterances combined	**3.57** 3.47-3.67	3.25 3.13-3.37	3.34 3.27-3.41	**3.37** 3.30-3.43	3.27 3.20-3.34	3.13 3.07-3.20
Question 3 all utterances combined	**3.99** 3.90-4.07	3.50 3.39-3.62	**3.72** 3.66-3.77	3.66 3.60-3.72	3.52 3.46-3.58	3.28 3.22-3.34

Table 4: Results for all participants and all questions in the MTurk study (CI: 95% confidence interval).

swering "yes/no" would disqualify them. However, it is unclear whether participants understood the question. For example, the results for TTS developers (18 in total) were surprising; GEN-LOW was rated as better than all voices, except for LD2, which raises questions about how different people perceive the term "familiarity with TTS". It could be the case that these participants thought that synthesized voices were supposed to sound like TTS rather than human-like. The rest of the results were consistent with our overall results, i.e. PROF was perceived as the best voice, followed by GEN-HIGH and LD1. LD2 was rated higher than LD1 by the participants who often use TTS (not significant).

7. Discussion

The trends that we see in our results are more or less expected. What is surprising though is that in the MTurk study on average GEN-HIGH and LD1 were perceived as significantly better than AMAT. It is hard to generalize beyond the voices that we included in our study but this result shows that speech synthesis has reached a high level of quality. As our results show, even the best synthesized voices cannot reach the level of professional voices, however, the differences in scores were not very large. This suggests that investing on a high-quality voice rather than using professional or amateur human recordings can be a good trade-off between performance and cost.

One limitation of our study though is that we do not deal with the issue of how different voices are perceived during a full interaction with a virtual human dialogue system. This is something to address in our future work. One would expect people to be less tolerant of poor quality voices if they had to listen to them during a complete dialogue ses-

	PROF mean CI	AMAT mean CI	GEN-HIGH mean CI	LD1 mean CI	LD2 mean CI	GEN-LOW mean CI
US English	**3.88**	3.51	**3.46**	3.42	3.30	3.15
(173 raters)	3.78-3.97	3.39-3.64	3.39-3.53	3.36-3.49	3.23-3.37	3.08-3.23
UK English	**3.46**	3.13	**3.27**	3.24	2.75	3.07
(62 raters)	3.24-3.69	2.77-3.49	3.14-3.40	3.10-3.37	2.61-2.88	2.92-3.21
Indian English	**3.83**	3.46	**3.67**	3.65	3.61	3.36
(264 raters)	3.76-3.90	3.36-3.56	3.62-3.72	3.60-3.70	3.56-3.66	3.31-3.41
Other	**3.71**	3.35	**3.60**	3.49	3.28	3.04
(303 raters)	3.62-3.80	3.22-3.47	3.54-3.67	3.43-3.55	3.22-3.35	2.98-3.11
not specified	**4.32**		**3.71**	3.35	2.92	2.63
(24 raters)	3.97-4.67		3.42-3.99	3.06-3.64	2.64-3.20	2.26-3.01
all	**3.86**	3.40	**3.60**	3.54	3.41	3.17
(826 raters)	3.80-3.91	3.33-3.47	3.57-3.64	3.50-3.58	3.37-3.44	3.14-3.21

Table 5: Results for native vs. non-native participants in the MTurk study, all questions combined (CI: 95% confidence interval).

	PROF mean CI	AMAT mean CI	GEN-HIGH mean CI	LD1 mean CI	LD2 mean CI	GEN-LOW mean CI
under 15	**4.00**		**3.78**	3.49	2.97	2.96
(14 raters)	3.12-4.88		3.51-4.04	3.19-3.80	2.64-3.30	2.57-3.36
15-25	**3.79**	3.46	**3.48**	3.43	3.39	3.20
(344 raters)	3.71-3.86	3.36-3.56	3.43-3.54	3.38-3.48	3.34-3.45	3.14-3.25
26-35	**3.78**	3.14	**3.60**	3.49	3.29	3.20
(286 raters)	3.69-3.87	3.00-3.27	3.53-3.66	3.43-3.55	3.23-3.36	3.13-3.27
36-45	**4.25**	3.52	**3.97**	3.90	3.68	3.34
(111 raters)	4.16-4.33	3.38-3.66	3.90-4.04	3.83-3.97	3.61-3.76	3.26-3.42
46-55	**3.86**	2.94	**3.51**	3.49	3.31	3.04
(27 raters)	3.69-4.03	2.70-3.17	3.38-3.65	3.35-3.62	3.17-3.45	2.88-3.20
56-65	3.39	**3.63**	2.93	2.85	**2.96**	2.17
(19 raters)	3.02-3.75	3.05-4.22	2.61-3.25	2.54-3.16	2.69-3.23	1.93-2.41
above 65				**3.60**		
(1 rater)				1.71-5.49		
not specified	**4.32**		**3.71**	3.35	2.92	2.63
(24 raters)	3.97-4.67		3.42-3.99	3.06-3.64	2.64-3.20	2.26-3.01
all	**3.86**	3.40	**3.60**	3.54	3.41	3.17
(826 raters)	3.80-3.91	3.33-3.47	3.57-3.64	3.50-3.58	3.37-3.44	3.14-3.21

Table 6: Results for all age ranges of participants in the MTurk study, all questions combined (CI: 95% confidence interval).

sion (rather than just listen to isolated utterances). Another issue is that a virtual human that uses human recordings will be able to generate a limited number of phrases, which could make people perceive it as unrealistic and boring.

8. Conclusions

We performed a systematic evaluation of human and synthesized voices with regard to naturalness, conversational aspect, and likability. We also varied the type (in-domain vs. out-of-domain), length, and content of utterances, and took into account the age and native language of the raters as well as their familiarity with speech synthesis. Our results suggest that professional human voices are perceived as better than both amateur human voices and synthesized voices. Also, a high-quality general-purpose voice or a good limited-domain voice can perform better than amateur human recordings. There were not any significant differences between the performance of a high-quality general-purpose voice and a limited-domain voice, both trained with speech recorded by actors. As expected, in most cases, the high-quality general-purpose voice was rated higher than the limited-domain voice for out-of-domain sentences and lower for in-domain sentences. We also found a not statistically significant trend for long or negative-content utterances to receive lower ratings. There was some variation in the ratings with regard to whether the raters were native or non-native speakers of English, and results were generally consistent across age groups.

9. Acknowledgements

The work presented here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily re-

	PROF mean CI	AMAT mean CI	GEN-HIGH mean CI	LD1 mean CI	LD2 mean CI	GEN-LOW mean CI
TTS developer	3.20	**3.27**	2.68	3.02	**3.70**	3.45
(18 raters)	2.50-3.90	2.37-4.16	2.20-3.15	2.61-3.42	3.28-4.12	3.11-3.79
often use TTS	**4.02**	3.68	**3.95**	3.79	3.82	3.60
(72 raters)	3.91-4.12	3.52-3.84	3.87-4.02	3.71-3.87	3.75-3.89	3.53-3.68
sometimes use TTS	**3.88**	3.31	**3.68**	3.64	3.47	3.27
(298 raters)	3.81-3.95	3.20-3.41	3.63-3.72	3.59-3.69	3.42-3.52	3.22-3.31
never use TTS	**3.74**	3.38	**3.48**	3.43	3.26	3.07
(414 raters)	3.68-3.81	3.29-3.47	3.43-3.53	3.38-3.48	3.21-3.31	3.03-3.12
not specified	**4.32**		**3.71**	3.35	2.92	2.63
(24 raters)	3.97-4.67		3.42-3.99	3.06-3.64	2.64-3.20	2.26-3.01
all	**3.86**	3.40	**3.60**	3.54	3.41	3.17
(826 raters)	3.80-3.91	3.33-3.47	3.57-3.64	3.50-3.58	3.37-3.44	3.14-3.21

Table 7: Results for the TTS expertise of participants in the MTurk study, all questions combined (CI: 95% confidence interval).

flect the position or the policy of the United States Government, and no official endorsement should be inferred.

10. References

Sebastian Andersson, Kallirroi Georgila, David Traum, Matthew Aylett, and Robert A. J. Clark. 2010. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Proc. of Speech Prosody*, Chicago, IL, USA.

Sebastian Andersson, Junichi Yamagishi, and Robert A. J. Clark. 2012. Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication*, 54(2):175–188.

Matthew P. Aylett, Christopher J. Pidcock, and Mark E. Fraser. 2006. The Cerevoice Blizzard Entry 2006: A prototype small database unit selection engine. In *Proc. of the Blizzard Challenge*, Pittsburgh, PA, USA.

Alan W. Black and Kevin A. Lenzo. 2000. Limited domain synthesis. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

Alan W. Black and Paul Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis. In *Proc. of the European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece.

Justine Cassell and Andrea Tartaro. 2007. Intersubjectivity in human-agent interaction. *Interaction Studies*, 8(3):391–410.

Robert Dickerson, Kyle Johnsen, Andrew Raij, Benjamin Lok, Amy Stevens, Thomas Bernard, and D. Scott Lind. 2006. Virtual patients: Assessment of synthesized versus recorded speech. In *Studies in Health Technology and Informatics*.

Kate Forbes-Riley, Diane Litman, Scott Silliman, and Joel Tetreault. 2006. Comparing synthesized versus prerecorded tutor speech in an intelligent tutoring spoken dialogue system. In *Proc. of the International Florida Artificial Intelligence Research Society Conference*, Melbourne Beach, FL, USA.

Jonathan Gratch, Jeff Rickel, Elisabeth Andre, Norman Badler, Justine Cassell, and Eric Petajan. 2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 17(4):54–63.

Matthew Marge, Joao Miranda, Alan W. Black, and Alexander I. Rudnicky. 2010. Towards improving the naturalness of social conversations with dialogue systems. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue (SIGdial)*, Tokyo, Japan.

Albert A. Rizzo, Belinda Lange, John G. Buckwalter, E. Forbell, Julia Kim, Kenji Sagae, Josh Williams, Barbara O. Rothbaum, JoAnn Difede, Greg Reger, Thomas Parsons, and Patrick Kenny. 2011. An intelligent virtual human system for providing healthcare information and support. In *Studies in Health Technology and Informatics*.

William R. Swartout, Jonathan Gratch, Randall W. Hill Jr., Eduard H. Hovy, Stacy Marsella, Jeff Rickel, and David R. Traum. 2006. Toward virtual humans. *AI Magazine*, 27(2):96–108.

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proc. of the International Conference on Intelligent Virtual Agents (IVA)*, Tokyo, Japan.

David Traum. 2008. Talking to virtual humans: Dialogue models and methodologies for embodied conversational agents. In *I. Wachsmuth and G. Knoblich (Ed.) Modeling Communication with Robots and Virtual Humans*, pages 296–309.