Which Synthetic Voice Should I Choose for an Evocative Task?

Eli Pincus, Kallirroi Georgila & David Traum
USC Institute for Creative Technologies
12015 Waterfront Dr
Playa Vista, CA 90094, USA
pincus,kgeorgila,traum@ict.usc.edu

Abstract

We explore different evaluation methods for 4 different synthetic voices and 1 human voice. We investigate whether intelligibility, naturalness, or likability of a voice is correlated to the voice's *evocative function potential*, a measure of the voice's ability to evoke an intended reaction from the listener. We also investigate the extent to which naturalness and likability ratings vary depending on whether or not exposure to a voice is extended and continuous vs. short-term and sporadic (interleaved with other voices). Finally, we show that an automatic test can replace the standard intelligibility tests for text-to-speech (TTS) systems, which eliminates the need to hire humans to perform transcription tasks saving both time and money.

1 Introduction

Currently there are a wealth of choices for which output voice to use for a spoken dialogue system. If the set of prompts is fixed and small, one can use a human voice actor. If a wider variety and/or dynamic utterances are needed, then text-to-speech synthesis (TTS) is a better solution. There are high quality commercial solutions as well as toolkits for building voices. While many of these are getting better, none are completely natural, especially when it comes to emotional and conversational speech. It can be difficult to decide which voice to choose for a specific system, given multiple criteria, and also since TTS evaluation is a labor-intensive process, without good automated understudies.

In this paper, we perform a comparative evaluation of several natural and synthetic voices using several different criteria, including subjective ratings and objective task measures. In particular, we compare the relationship of a voice's *evocative function potential*, a measure of the voice's ability to evoke an intended reaction from the listener, to the voice's intelligibility and to the listener's perception of the voice's naturalness and likability.

Our first hypothesis is that voice quality is a multi-dimensional construct, and that the best voice for some purposes may not be the best for all purposes. There may be different aspects that govern subjective perceptions of a voice and objective task performance, and different aspects may facilitate different tasks. For example, a neutral highly intelligible voice may be perfect for a system that provides information but very unpleasant for a story-telling system that is trying to express strong emotion.

Our second hypothesis is that naturalness and likability perceptions of a voice may depend on whether or not the user's exposure to a voice is extended and continuous vs. short-term and sporadic (interleaved with other voices). The current practice in speech synthesis evaluation is to ask human raters to rate isolated audio clips, usually in terms of naturalness and intelligibility (Fraser and King, 2007; Karaiskos et al., 2008), without extended exposure to a voice. This approach can certainly inform us about the general quality of a synthetic voice; but it cannot necessarily provide any insight about the appropriateness of this voice for a task that requires that the listener be exposed to that voice for a considerable amount of time. Furthermore, as the environments where these dialogue systems are deployed become increasingly immersive involving multiple agents, e.g., virtual and augmented reality environments, it becomes critical to determine how subjective perceptions of a voice change if voice exposure is sporadic and interleaved with other voices[1].

[1] From now on, we will assume that sporadic voice exposure implies that the user is exposed to multiple voices interleaved.

Proceedings of the SIGDIAL 2015 Conference, pages 105–113,
Prague, Czech Republic, 2-4 September 2015.

Noting that it is not always feasible to evaluate a voice in the context of a full dialogue task we seek to determine whether results from standard voice evaluation experiments can act as a valid proxy for results from experiments that feature voice evaluation in a manner that more closely approximates the full dialogue task. Taking this idea one step further, we explore whether or not standard TTS evaluation tests such as transcription tasks (designed to assess the intelligibility of a voice) can be fully automated by using automatic speech recognition (ASR) output rather than manual transcriptions.

To test our hypotheses we perform 5 experiments using 4 synthetic voices (covering a range of speech synthesis techniques) and 1 human voice. Each experiment is defined by a unique set of stimuli, subjects, and measures. In the first two experiments, we perform standard speech synthesis evaluation, i.e., human raters rate isolated audio clips with regard to naturalness in one experiment and likability in the other experiment (each rater has short-term sporadic exposure to the voices). Experiments 3 and 4 are intelligibility experiments; in one, participants transcribe the utterances that they hear; in the other, we send audio files through an ASR engine. The fifth experiment is conducted in the context of a guessing game with extended continuous naturalness and likability ratings collected from participants. The evocative intention of an utterance is the behavior of the addressee that a speaker intends to evoke (Allwood, 1976; Allwood, 1995). In the case of the guessing game, a clue is given to evoke the expression of a target word. We ascertain a voice's *evocative function potential* (EVP) by calculating the ratio of targets that a clue evokes from listeners. Each participant listens to many consecutive clues uttered with the same voice (extended continuous exposure). Our participants are recruited using the Amazon Mechanical Turk (AMT) service[2] in the same fashion as in (Wolters et al., 2010; Georgila et al., 2012). To the best of our knowledge, our work is the first to systematically attempt to validate or disprove the hypotheses mentioned above, and compare the results of human transcriptions to ASR results in order to determine whether or not the latter can be used as an automatic intelligibility test for TTS system evaluations. This is also a first important step towards speech synthesis evaluation in a full dialogue context. Finally, this is the first time that a systematic evaluation is conducted on a voice's EVP.

The rest of the paper is organized as follows. First, we discuss previous work in Section 2 on TTS system evaluations. In Section 3 we present the voices that we use as well as meta-data about the clues that the voices spoke. In Section 4 we delineate the experiment methodology, and in Section 5 we report the results of our experiments and some inferences we can draw from them. Finally, Section 6 concludes.

2 Previous Work

Our ultimate goal is to evaluate synthetic voices in the context of a full interaction with a dialogue system, and analysis of the effects of extended/continuous vs. short-term/sporadic exposure of a listener to a voice is a first important step towards this goal. There has been some work on comparing the effect of synthetic 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 a large variety of voices or conditions, e.g., length and content of utterances, etc.

Recently, Georgila et al. (2012) performed a systematic evaluation of human and synthetic voices with regard to naturalness, conversational aspect, and likability. They also varied the type (in- vs. out-of-domain), length, and content of utterances, and took into account the age and native language of raters as well as their familiarity with speech synthesis. However, this study was based on the standard speech synthesis evaluation.

3 Data

3.1 Materials

Our experiments use 4 different synthetic voices and 1 human voice, all male, with standard American accents.

- ***Human voice (HUM)***: The audio clips were recorded by the first author using a high-quality microphone with noise cancellation features. The resulting audio clips were very clear, almost studio-quality.

- ***Commercial voice 1 (US1)***: This is a high-quality commercial stylized voice based on

[2] https://www.mturk.com

Table 1: Example Clues

Clue	Type	Source	Target Word
"an explosive device fused to explode under specific conditions"	Definition	WordNet	Bomb
"a blank to talk too much"	Example Usage	Dictionary.com	Tendency
"taxi"	Word Relation	Human	Cab
"a mixture containing two or more blank elements or blank and nonblank elements usually fused together or dissolving into each other when molten"	Definition	WordNet	Metal
"elephants may look alike to you and me, but the shapes of their blank flaps and their tusks set them apart"	Example Usage	Dictionary.com	Ear
"um not video but"	Word Relation	Human	Audio

Unit-**S**election (Hunt and Black, 1996; Black and Taylor, 1997).

- ***Commercial voice 2 (US2)***: This is a high-quality commercial customized **U**nit-**S**election voice developed specifically for our institute.

- ***Hidden Markov model -based voice (HMM)***: This voice is based on **HMM** synthesis (Zen et al., 2009), in particular, speaker-adaptive HMM-based speech synthesis (Yamagishi et al., 2009). First an average voice was built using the CMU ARCTIC speech databases[3]. Then this average voice was adapted to the voice characteristics of a speaker using approx. 15 minutes of speech from that speaker (studio-quality recordings). We built this voice using the HTS toolkit with its standard vocoder (Zen et al., 2007).

- ***Lower quality voice (SAM)***: We used Microsoft **Sam**.

We measure a voice's EVP for the guessing task by providing clues for listeners to guess a specific target word. We used 54 clues from a corpus of automatically and human generated clues. The material for the automatically generated clues came from two sources: WordNet (Miller, 1995) and the Dictionary.com pages associated with the target word. We replaced any occurrence of the target word or inflected forms of the target word in the clues used with the word *"blank"*. The human clues were culled from the rapid dialogue game corpus which contains audio and video recordings of human pairs playing a word guessing game (Paetzel et al., 2014). We only used clues that were able to elicit at least one correct guess in a previous study designed to measure clue effectiveness (Pincus et al., 2014). Some example clues used in this experiment, their source, their type, and the target word they intend to evoke can be found in Table 1. Each of the 54 clues was synthesized in each of the voices.

We categorized the 54 clues into 3 main clue types: a *definition* type which provided a definition of the target word, an *example usage* type which is generally a commonly used sentence that contains the word, and a *word relation* type which refers to clue types such as synonyms, hyponyms, hypernyms, antonyms, etc. of the target word. Human clues were annotated according to this taxonomy (Pincus and Traum, 2014). For our analysis we looked at cumulative statistics for the full set of clues as well as statistics for two different partitions of the clue corpus; by type and by length ($> 5\ words$ and $\leq 5\ words$). The relative frequency for each type of clue can be found in Table 2; 24% or 13/54 of the clues are composed of 5 or fewer words while 76% (41/54) of the clues are composed of more than 5 words. The average clue length is 10.75 words and the standard deviation of clue lengths is 7.86 words.

3.2 Participants

We crowdsourced data collection for this experiment via Amazon Mechanical Turk. All Turkers who completed the task were required to have a 90% approval rating or higher and have at least 50

[3] `http://www.festvox.org/cmu_arctic/`

approved HITs. Note that no Turker participated in more than one of any of the experiments described in Section 4.

Table 2: Clue Type Frequency

Clue Type	Relative Frequency (absolute # / 54)
Definition	63% (34)
Example Usage	24% (13)
Word Relation	13% (7)

4 Method

A summary of the 5 experiments conducted in this study, introduced in section 1, and the measures obtained from each experiment can be found in Table 3. The standard naturalness, likability and intelligibility experiments featured short-term sporadic exposure to the 5 voices and were designed using the online survey software Qualtrics[4]. In these experiments all participating Turkers listened to 20 audio recordings (human or synthetic speech) of clues randomly selected from the 54 clues described previously. Each set of 20 audio recordings was balanced so that the participant would listen to 4 clips per voice. The order of clues and voices was randomized, i.e., there was constant switching from one voice to another (short-term sporadic exposure to a voice). Generally, each participant never heard a clue more than once. Turkers were instructed to listen to an audio file only once in these experiments in order to more accurately model a normal spoken language situation such as transcribing a lecture or simultaneous interpretation.

54 different Turkers participated in the standard naturalness experiment. After listening to an audio file a Turker answered the following question: "For the utterance you just heard, how did the voice sound?" (1=very unnatural, 2=somewhat unnatural, 3=neither natural nor unnatural, 4=somewhat natural, 5=very natural). We will call this a Turker's **short-term/sporadic (S/S) naturalness measure**.

54 different Turkers participated in the likability experiment. After listening to an audio file a Turker answered the following question: "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). We will call this a Turker's **short-term/sporadic (S/S) likability measure**.

Table 3: Experiments & Obtained Measures

Experiment	Obtained Measures
1. Standard Naturalness	1. Short-Term/Sporadic (S/S) Naturalness
2. Standard Likability	1. Short-Term/Sporadic (S/S) Likability
3. Standard Intelligibility	1. Human Wrd. Err. Rate 2. Human Miss. Word %
4. ASR Intelligibility	1. ASR Wrd. Err. Rate 2. ASR Miss. Word %
5. Guessability	1. Extended/Continuous (E/C) Naturalness 2. Extended/Continuous (E/C) Likability 3. Guessability

The standard intelligibility experiment was designed as a transcription task. 55 Turkers listened to audio recordings of the clues described previously and then wrote into a text box what they heard. 6 of the 55 Turkers' transcription results were discarded; 2 Turkers did not appear to make a best effort and 4 misread the instructions and provided guesses for the clues they heard rather than transcribing the audio. We compared the transcriptions with the actual text of the clue that was synthesized or recorded (reference). In order to compare the results of this intelligibility experiment with the results from an automatic test of intelligibility (ASR intelligibility experiment) we send the 54 audio recordings of each clue for each voice through the Google Chrome ASR[5]. For both standard and ASR intelligibility, we calculated **word error rate (WER)** (Equation 1), and the percentage of words contained in the reference but not in the target transcription (**missing word %**).

$$WER = \frac{Subs.\ +\ Delets.\ +\ Inserts.}{\#\ Of\ Words\ In\ Reference} \tag{1}$$

A web application was developed for the guessability experiment, and Turkers were redirected to this application from the AMT site to participate in the experiment. Each Turker in the guessing experiment had extended continuous exposure to 3 of the 5 voices, listening to 18 clues in each voice, for a total of 54 clues. We collected a full set of 54

[4] http://www.qualtrics.com/

[5] https://www.google.com/intl/en/chrome/demos/speech.html

recordings from 59 different Turkers and almost a full set (53/54) recordings from a 60th Turker (who failed to make a guess for the last clue). Note that many more Turkers attempted the experiment but failed to finish for unknown reasons. We do not consider this partially collected data except for the 60th Turker's data just mentioned. Turkers heard only one instance of each clue. The order of voices was balanced (there are 60 permutations of the voices possible with our experimental set up; so each Turker heard 3 voices in a unique order), but clues were presented in a fixed order. Each Turker, when listening to a clue, was instructed to make as many guesses as he could before a pop-up alert appeared (six seconds later), indicating that recording had ended and revealing the target word. After each clue the Turker was asked to rate the naturalness of the voice he had just heard on a Likert scale as in the previously described experiments except the word "clue" replaced the word "utterance" in the question. The average of these 18 naturalness scores for each Turker will be called a Turker's **extended/continuous (E/C) naturalness score**. After each set of 18 clues with the same voice, the Turker was asked whether or not he would like to have a conversation with the speaker the Turker had just been exposed to for the last 18 clues (same question as in the previously described likability experiment). We will call this a Turker's **extended/continuous (E/C) likability score**.

We annotated the 60 sets of audio recordings (3,239 audio files) of Turkers' guesses for whether or not the recording contained a correct guess. An audio recording was annotated as correct if it contained a guess composed of the target word or an inflected form of the target word for the previously spoken clue. We define a **guessability score** for a voice as the percentage of correctly guessed clues out of the total number of clues played to participants with that voice.

All the likability and naturalness measures we categorize as subjective measures while the intelligibility and guessability measures we categorize as objective measures.

5 Results

This section contains the results of our experiments including the S/S and E/C naturalness ratings in Table 4, and the S/S and E/C likability ratings in Table 5, and all the objective measures in Table 6. The general ranking of the voices across the various subjective and objective dimensions measured were (starting with the highest ranking voice and proceeding in decreasing order): human (HUM), commercial (US1), commercial (US2), hidden Markov model (HMM), lower quality voice (SAM). We will refer to this as the standard order. The existence of a standard order indicates that we did not find good evidence to support hypothesis 1. At first glance any measure is a good proxy for another measure; however there are some exceptions. If there is a statistically significant exception we will explicitly mention it. A marking of "***" by a measure in one of the three tables indicates that the difference between that measure with the measure for the next ranked voice is highly significant ($p<.001$)[6]. A marking of "**" by a measure in one of the three tables indicates that the difference between that measure with the measure for the next ranked voice is significant ($p<.01$). Finally, a marking of "#" by a measure in one of the three tables indicates that the difference between that measure and the voice ranked 2 below is significant ($p<.01$).

5.1 Subjective & Objective Measures

Table 4: S/S & E/C Naturalness Means

Voice	S/S Naturalness Avg.	E/C Naturalness Avg.
HUM	4.15***	4.59***
US1	3.93***	3.48***
US2	2.92***	2.04***
HMM	2.04***	1.83***
SAM	1.81	1.57

Table 5: S/S & E/C Likability Means

Voice	S/S Likability Avg.	E/C Likability Avg.
HUM	$3.78^{\#}$	4.17**
US1	3.63***	3.36***
US2	2.66***	1.69
HMM	1.81	1.53
SAM	1.72	1.35

The voices follow the standard order for both S/S and E/C mean naturalness, and all pair-wise

[6] Statistical tests conducted were paired or unpaired t-tests (based on the relationship of the data sets tested) with the use (if needed) of the Holm - Bonferroni method to counteract the issue of multiple comparisons.

Table 6: Objective Measure Means

Voice	Guessability	Human Word Err. Rate	Human Missing Word %	ASR Word Err. Rate	ASR Missing Word %
HUM	57.10%$^{\#}$	18.35% $^{\#}$	15.64%$^{\#}$	5.41%**	5.24%**
US1	59.72%**	23.31%***	20.53%***	6.11%$^{\#}$	4.54%$^{\#}$
US2	50.39%$^{\#}$	29.65%$^{\#}$	25.18%$^{\#}$	21.82%**	18.5%**
HMM	46.45%	29.32%***	25.44%***	13.26%$^{\#}$	10.3%$^{\#}$
SAM	42.44%	35.43%	32.36%	28.27%	24.78%

comparisons for both S/S and E/C show differences in means that were highly statistically significant. This indicates that synthetic voices, at least the ones tested, have still not reached human-level naturalness. There were no significant violations to this pattern in various subsets of clues tested. The S/S and E/C likability scores can be found in Table 5 for all clues. Again, both measures follow the standard order. It is interesting that the US1 and HUM voices do not have a significant difference in their S/S likability but do for their E/C likability ($p = 0.008$). In terms of naturalness and likability we believe the HMM scored low due to the fact that it was not trained on a large amount of data (only 15 minutes of speech was used for adaptation) and also the fact that it did not use a more advanced vocoder such as STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) (Kawahara, 1997). Overall, this data suggests that synthetic voices are catching up faster in the likability dimension to HUM voices than in the naturalness dimension, although an experiment with more human voices is needed for more evidence of this trend.

For standard intelligibility results the standard order is followed for both WER and missing word %. The HUM voice performs best although its performance over US1 is not significant, demonstrating that synthetic voices are able to match human voices in intelligibility measures. We see from Table 6 that the overall intelligibility of US2 and HMM is comparable. However, the HMM voice outperformed US2 significantly ($WER : p = 0.002$, $missing\ word\ \% : p = 0.017$) on example usage clues. Noting that the HMM voice extended the pronunciation of the word "blank" (which appeared in almost all of the example usage clues) this could provide some support for a hypothesis that unnatural sounding words remained in the listeners' short-term memory more readily. However, further experiments are needed to verify whether or not this is just an aberration. For the ASR intelligibility results although the standard order was violated, HMM outperformed US2 for both WER and missing word % and US1 outperformed HUM for missing word %, these deviations were not significant. Overall, the intelligibility results indicate that Google Chrome ASR is much better than real-time Turkers at the transcription task (where Turkers have only a single opportunity to hear the audio).

In the guessability dimension the standard order is violated because US1 outperformed HUM there but we draw no conclusions from this as it is not a statistically significant difference. The performance of US1 for guessability is significantly ($p = 0.001$) better than US2 but has comparable performance to the HUM voice indicating that synthetic voices have reached an EVP approaching human level for the clue guessing task. One hypothesis on why US2 has significantly worse guessability than US1 and HUM is that although US2 is a high-quality voice, more effort has been put in making this voice expressive rather than making sure that all phonetic units are fully covered in all possible contexts. In terms of the guessability for the various sub-groups of clues it appears all voices are performing much better for long clues except for HUM which has similar performance for both long and short clues. SAM is particularly bad for short clues, with guessability 33.3% (compared to 45.3% for long clues).

These results indicate that if one is concerned with the subjective perception of the system carrying out the task or its intelligibility rather than only the task performance measure then HUM is the undeniable best voice. However, if one is only concerned with maximizing the EVP of a dialogue system then US1 might be the preferred choice; as it eliminates the need for human recordings.

5.2 Time/Continuity-Exposure

In order to determine if time/continuity of voice exposure is an important variable in determining people's subjective evaluations of a voice (note that hypothesis 2 was that this is an important variable) we consider the difference between 3 different pairs of statistics for each voice for all clues. The first pair of statistics we compare are the average S/S likability scores and the average E/C likability scores. These statistics are found in Table 5. We see that the likability scores decreased for all the synthetic voices (decrease in US2's likability scores highly statistically significant: $p = 3.6e^{-05}$) but increased for the human voice ($p = 0.04$) . The second pair of statistics we compare are the S/S naturalness scores and the E/C naturalness scores. These statistics are given in table 4. We see the same pattern with S/S and E/C naturalness scores that we saw with S/S and E/C likability scores for the 5 voices; increasing naturalness scores for the HUM voice and decreasing naturalness scores for the synthetic voices. Moreover, every difference is highly significant here ($HUM : p = 3.08e^{-16}$, $US1 : p = 1.01e^{-12}$, $US2 : p = 6.72e^{-33}$, $HMM : p = 0.06e^{-2}$, $SAM : p = 6.53e^{-05}$).

Table 7: First vs. Last Naturalness Scores

Voice	First Three Naturalness Avg.	Last Three Naturalness Avg.
HUM	4.25	4.81*
US1	3.42	3.52
US2	2.58	1.833*
HMM	1.69	1.78
SAM	1.67	1.31

An attempt to examine whether or not time exposure alone has an effect on subjective evaluation of a voice leads us to examine a third pair of statistics: comparing the average of the first three naturalness scores from a Turker in the guessability experiment to the average of the last three naturalness scores (of 18 total) of the same voice (first voice heard only). This comparison provides evidence that the pattern we are discussing is not simply due to the difference in the types of tasks participants were asked to perform. These scores can be found in Table 7. A "*" in the second column indicates that the corresponding increase or decrease is statistically significant ($HUM : p = 0.017, US2 : p = 0.013$). Although US1's and HMM's naturalness averages increase, these increases are not significant. One issue to point out here is that the order of clues was fixed so the synthetic voices might have had worse performance on the last clues vs. the first clues.

We now note that this study has results from two experiments where synthetic voices have a statistically significant decrease and where a human voice has a statistically significant increase in subjective evaluation ratings when comparing the ratings from people who had S/S vs. E/C exposure to the voices. These findings provide support for hypothesis 2 indicating that extended/continuous exposure to a synthetic voice negatively affects subjective perception of that voice. Furthermore, this study has shown results from one experiment which suggests that people's subjective perceptions of synthetic voices degrade over time while their subjective perceptions of human voices improve over time. Additional experiments with more human voices and a balanced order of clues could be conducted to provide further support for this phenomenon.

5.3 Correlation Analysis

Table 8 presents the results of a correlation analysis between guessability and the other dimensions previously discussed. The correlation results for guessability and the two naturalness scores do not lead us to any clear conclusions. The only statistically significant correlation is between E/C naturalness, which had ratings collected after a participant had received feedback on the correctness of their guess (which could affect the rating), and guessability.

Table 8: Guessability Correlations

Categories	r_s[7]	P-Value
Guessability & S/S Natural.	0.122	0.051
Guessability & E/C Natural.	0.31	$0.002e^{-4}$
Guessablity & S/S Likability	0.108	0.085
Guessability & Stand. Word Error Rate	-0.108	0.081
Guessability & % Stand. Missing Word %	-0.129	0.035

[7] Spearman's Rank-Order Correlation Coefficient

Table 9: Intelligibility Correlations

Voice	Word Error Rate Standard ASR Corr. (ρ)[8](p-val)	Missing Word % Standard ASR Corr. (ρ)[8] (p-val)
HUM	0.06 (0.37)	0.07 (0.29)
US1	0.27 ($1.66e^{-36}$)	0.26 ($3.97e^{-05}$)
US2	0.55 ($1.37e^{-05}$)	0.58 ($5.21e^{-23}$)
HMM	0.78 ($7.17e^{-52}$)	0.74 ($2.52e^{-43}$)
SAM	0.07 (0.29)	0.17 (0.007)

We find weak negative correlations between guessability and both of the measures from the standard intelligibility experiments. Note that only the correlation between missing word % and guessability is statistically significant. This indicates that while intelligibility measures of a voice could be useful information when evaluating a voice's EVP the correlation is not strong enough to suggest that they are valid proxy measures for a voice's EVP. Furthermore, performing voice evaluation in an experiment that features the full context of the system being evaluated might still be required for precise voice evaluation results of a dialogue system.

Table 9 shows the correlations for each voice between the ASR intelligibility experiment results and the standard intelligibility experiment results. For almost all of the synthetic voices there is a strong or somewhat strong positive correlation between the ASR intelligibility experiment results and the standard intelligibility results that has high statistical significance. The one exception to this is SAM's ASR WER which shows no significant relationship with the human transcriptions WER. It is also interesting that for the HUM voice the ASR intelligibility results show basically no correlation to the standard intelligibility results. Overall though, it appears that for synthetic voices intelligibility results can be obtained automatically by sending recordings of the voice to a well-trained ASR engine such as Google Chrome ASR; and these should be able to predict the results from a standard (human participant) intelligibility test.

6 Conclusion

We presented the results of an evaluation for 4 synthetic voices and 1 human voice that featured collection of data for subjective perception measures as well as for objective task measures of the voices. We demonstrated that synthetic voices do not always have significantly lower EVP than a human voice (US1 is similar); although they do significantly differ in subjective ratings assigned to them by listeners. For this reason, we would choose a human voice for a dialogue system designed to evoke an intended reaction from a listener only if subjective perceptions were important enough to the system designer to warrant the extra cost and time of making human audio recordings.

We showed via comparison of measures of the voice's EVP with measures of subjective perceptions and intelligibility that while you cannot always use standard measures of synthetic voice evaluation as a proxy for a new task, in determining the voice's effectiveness at that new task, the results from standard tests can provide useful information. Some of our data led us to suggest that synthetic voices' likability and naturalness perceptions degrade based on time/continuity of exposure while human voices' likability and naturalness perceptions improve with increasing time/continuity. Finally, we provided evidence that the automatic method of sending synthetic voice audio recordings through an ASR engine can serve as an adequate substitute for standard (human participant) intelligibility experimental results, and that the automatic method even outperforms Turkers' transcription ability (when Turkers hear the audio only once).

Future work includes additional experiments that will control for the order of the clues as well as cover a wider variety of tasks. Finally, we would like to evaluate EVP in the context of a full dialogue, where users can clarify and perform moves other than guesses, and multiple clues might contribute to a guess.

7 Acknowledgements

The effort described here is supported by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Jens Allwood. 1976. *Linguistic Communication as Action and Cooperation*. Ph.D. thesis, Göteborg University, Department of Linguistics.

[8] Pearson Product-Moment Correlation Coefficient

Jens Allwood. 1995. An activity based approach to pragmatics. Technical Report (GPTL) 75, Gothenburg Papers in Theoretical Linguistics, University of Göteborg.

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.

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 pre-recorded tutor speech in an intelligent tutoring spoken dialogue system. In *Proc. of the International Florida Artificial Intelligence Research Society Conference*, Melbourne Beach, FL, USA.

Mark Fraser and Simon King. 2007. The Blizzard Challenge 2007. In *Proc. of the ISCA Workshop on Speech Synthesis*, Bonn, Germany.

Kallirroi Georgila, Alan W. Black, Kenji Sagae, and David Traum. 2012. Practical evaluation of human and synthesized speech for virtual human dialogue systems. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Andrew J. Hunt and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, GA, USA.

Vasilis Karaiskos, Simon King, Robert A. J. Clark, and Catherine Mayo. 2008. The Blizzard Challenge 2008. In *Proc. of the Blizzard Challenge*, Brisbane, Australia.

Hideki Kawahara. 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *IEEE International Conference On Acoustics, Speech, And Signal Processing*, Munich, Germany. Acoustics, Speech, and Signal Processing.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Maike Paetzel, David Nicolas Racca, and David DeVault. 2014. A multimodal corpus of rapid dialogue games. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Eli Pincus and David Traum. 2014. Towards a multimodal taxonomy of dialogue moves for word-guessing games. In *Proc. of the 10th Workshop on Multimodal Corpora (MMC)*, Reykjavik, Iceland.

Eli Pincus, David DeVault, and David Traum. 2014. Mr. Clue - A virtual agent that can play word-guessing games. In *Proc. of the 3rd Workshop on Games and NLP (GAMNLP)*, Raleigh, North Carolina, USA.

Maria K. Wolters, Karl B. Issac, and Steve Renals. 2010. Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proc. of the ISCA Workshop on Speech Synthesis*, Kyoto, Japan.

Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, and Steve Renals. 2009. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1208–1230.

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda. 2007. The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. of the ISCA Workshop on Speech Synthesis*, Bonn, Germany.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.