A systematic study of open source and commercial text-to-speech (TTS) engines

Jordan Hosier[1], Jordan Kalfen[2], Nikhita Sharma[1], and Vijay K. Gurbani[1,2]

[1] Vail Systems, Inc., Chicago (USA)
{jhosier, nsharma, vkg}@vailsys.com
[2] Illinois Institute of Technology, Chicago (USA)
jkalfen@hawk.iit.edu, vkg@iit.edu

Abstract. The widespread availability of open source and commercial text-to-speech (TTS) engines allows for the rapid creation of telephony services that require a TTS component. However, there exists neither a standard corpus nor common metrics to objectively evaluate TTS engines. Listening tests are a prominent method of evaluation in the domain where the primary goal is to produce speech targeted at human listeners. Nonetheless, subjective evaluation can be problematic and expensive. Objective evaluation metrics, such as word accuracy and contextual disambiguation (is "Dr." rendered as Doctor or Drive?), have the benefit of being both inexpensive and unbiased. In this paper, we study seven TTS engines, four open source engines and three commercial ones. We systematically evaluate each TTS engine on two axes: (1) contextual word accuracy (includes support for numbers, homographs, foreign words, acronyms, and directional abbreviations); and (2) naturalness (how natural the TTS sounds to human listeners). Our results indicate that commercial engines may have an edge over open source TTS engines.

1 Introduction

As voice enabled devices gain prominence in our daily lives, it is increasingly important that such technologies possess human-like speech capabilities. The perceptual quality of TTS speech synthesis technology impacts the acceptability of such systems. For this reason, there is a push among TTS researchers to make synthetic speech more naturalistic. The applications for such technologies are vast, including solutions for the visually impaired, hands-free technology, customer-service centers, etc. In the market today, there are many TTS engines with varied capabilities. The goal of this study is to propose a set of evaluation metrics which can be used to evaluate TTS engines. The proposed evaluation is based on a measure we call *contextual word accuracy* (formally defined in Section 3), and the necessary, though subjective measure of *naturalness*, i.e., how do human listeners rank the TTS engines?

Seven TTS engines were considered: four open source engines and three commercial engines. The open source engines were:

– Mimic[3]: Mimic is the light-weight TTS component based on Carnegie Mellon's FLITE software (see below).
– CMU FLITE[4]: FLITE is a small TTS synthesis engine developed at Carnegie Mellon University (CMU) and is designed for small, embedded machines as well as large servers.
– MaryTTS [5]: MaryTTS is a Java-based multilingual TTS synthesis platform using a Hidden Markov Model (HMM-) model.
– DeepVoice3[6] [2]: DeepVoice3 is a fully convolutional attention-based neural TTS system.

The following commercial engines were evaluated using their respective cloud-based interfaces:

– Voicery[7]: Voicery is a commercial start-up offering a deep neural network.
– Acapela[8]: Acapela is a European company specializing in personalized digitized voices.
– Selvy[9]: Selvy a TTS synthesis engine from a South Korean company.

There are additional commercial engines such as Amazon Polly[10], Google Tacotron [3], and IBM Watson Text to Speech[11]. While we are not aware of any scientific study comparing these engines in a formal manner, it is widely assumed by practitioners that these engines are the state-of-art in TTS. Given this assumption, we use Amazon Polly as a control variable and benchmark on which to evaluate the seven TTS engines under consideration.

The remainder of the paper is organized as follows: Section 2 motivates the work, Section 3 presents our evaluation corpus of 21 test utterances, Section 4 details the evaluation methodology, and Section 5 presents results and discusses findings. Beyond Section 6 are several appendices that provide the raw data to elaborate on results.

2 Related work and contribution

Despite recent advancement in speech synthesis, the evaluation of such technology has seen little advancement and lacks an established gold standard of evaluation metrics. The classic approach for TTS evaluation is to synthesize a set of samples, present the samples to listeners, and to draw conclusions about the systems based on listener evaluation.

[3] https://mycroft.ai/documentation/mimic (last visit: April 23, 2020)
[4] http://www.festvox.org/flite/ (last visit: April 23, 2020)
[5] http://mary.dfki.de (last visit: April 23, 2020)
[6] https://github.com/r9y9/deepvoice3_pytorch (last visit: April 23, 2020)
[7] https://www.voicery.com (last visit: March 2020)
[8] https://www.acapela-group.com/ (last visit: April 23, 2020)
[9] http://speech.diotek.com/en/text-to-speech-demonstration.php (last visit: April 23, 2020)
[10] https://aws.amazon.com/polly/ (last visit: February 2020)
[11] https://www.ibm.com/Watson/services/text-to-speech/ (last visit: May 2019)

Objective measures have been developed for speech quality evaluation in telecommunication systems, such as measuring mel cepstral distortion [4, 5]. While these serve as a proxy for how well the TTS model represents natural speech, automating this process is challenging and often requires a benchmark natural speech signal [6]. While some measures do not require a reference speech signal [9], subjective listening tests remain the gold standard in the literature. The most common listening tests are Mean Opinion Score tests (MOS, ITU-T Rec. P.10, 2006), MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA, ITU-T Rec. BS.1543, 2015), preference tests, and transcription tasks. The attributes measured by such tests include measures of naturalness, intelligibility, similarity, etc.

The Blizzard Challenge was developed to better understand and compare research techniques in building corpus-based speech synthesizers on the same data [7]. Competitors present the results from a standard listening test and describe their systems. These tests included listening to a fixed number of utterances and subsequently assigning a domain-specific MOS score based on the test set. While this challenge has a well developed listening test, it is also subjective.

Primary contributions: TTS engines are used in a variety of applications and it is important that such technologies are flexible enough to adapt to the properties of novel environments. However TTS systems can be fragile, and often break down with minor changes in the lexicon. This work proposes a corpus (Section 3) of diverse set of English phonological and morphological artifacts (homographs, foreign loan words, acronyms, directional abbreviations, etc.) that present potential challenges to TTS engines. We seek to establish this corpus as a canonical corpus for evaluating TTS engines. Furthermore, we propose two evaluation methodologies (Section 4): an objective metric that allows for impartial evaluation of TTS response to complex input, and while the second metric is a subjective listening test, we attempt to control for subjectivity in evaluating it through using multiple advanced voting techniques.

3 Evaluation corpus

The set of 21 vectors used to evaluate the TTS engines is shown in Table 1. These sentences represent a diverse set of English phonological and morphological grammatical constructs that present potential challenges to TTS engines. While these sentences would be easily produced and understood by humans, they include ambiguities and homographs that could present challenges to a TTS system - challenges which potentially indicate inadequate training of the system. Thus, we test if the system can render these vectors with the accuracy that a human reader could easily achieve.

These stimuli included sentences with homographs (Test cases 10-12) and foreign words (Test cases 18 and 21). We also evaluate forms of abbreviations, including context dependent abbreviations (i.e. “Dr.” as a prefix to a name will be expanded as “Doctor”, while “Dr.” as a suffix to an address is expected to be expanded as “Drive”), abbreviations in addresses (i.e. “Apt.”, “Pl.”, “Pkwy.”), and abbreviations of names (i.e. “Chas.” for “Charles”). Finally, we evaluate

Test case	Sentence
1	American Communications & Engineering, Inc. is located at 123 NW. Main St., Apt. 1A, St. Paul, MN 60655
2	Natoma Professional Ctr. 555 Oakdale Pkwy., is located at 123 S. 2nd Pkwy., Ste. 700, Ft. Lauderdale, FL.
3	Valor Telecom Ltd. 1910 E. Kimberly Pl. P.O.B 93425, Old Village Sq., CA.
4	Sec. of State Hillary Clinton and Sen. Lisa Murkowski spoke with Pres. Mahmaud Abbas to discuss FASB.
5	Ex-Gov. Sarah Palin and Ex-HP CEO Carly Fiorina met with Israeli Ex-Prime Minister Ariel Sharon to discuss RBOC.
6	Treasury Sec. Timothy F. Geithner used to be the COO at JPMorgan and earned $4.5-million-a-year and earned an MBA from Harvard.
7	Rep. Chas. Rangel Ph.D was censured by PETA.
8	Mr. John Smith Sr. and Mrs. Jane Smith worked at Levi Strauss & Co. with their son, John Smith Jr., and daughter Ms. Judy Smith.
9	Gen. Douglas MacArther was tired of receiving SPAM from the NYSE.
10	They were too close to the door to close it.
11	The dove dove into the water
12	The team lead had lead us to victory.
13	After I read a book I add it to my list of books that I've read.
14	The farm was used to produce produce.
15	People who use are of no use.
16	Prof. Robt. B. Reich is a bona fide rocket scientist.
17	Dr. Albert Einstein, Phd had a lot of chutzpah turning down the presidency.
18	Jas. A. Barone III said bon voyage to Capt. Wm. O. Barnett before the coup d'etat.
19	I was born Mon., Sept. 25, 1989 at 12:30 AM
20	Lt. Cmd. Jas. W. Marks was born Wed. the 3rd. Of Mar. at 2:30 p.m.
21	I live in La Crosse county, Wisconsin. This is close to Eau Claire and Prarie du Chien.

Table 1: Corpus for evaluating the TTS engines

numbers (i.e. roman numerals, numbers in street addresses, and numbers occurring in a string denoting times or dates) and symbols ("&").

In summary, these 21 cases present non-trivial challenges to TTS engines to unambiguously pronounce the sentence in a manner consistent with expectations.

4 Evaluation methodology

We present two metrics of evaluation. The first of these metrics is ϕ, or *contextual word accuracy*. To evaluate ϕ, a sentence is considered as a bag of words. With that assumption, ϕ is defined as:

$$\phi = \frac{1}{n} \sum_{i=1}^{n} I(x_i), \tag{1}$$

where n is the total number of words in the bag, x_i enumerates over all the words in the bag, and $I(x_i)$ is the word pronunciation identity function defined as:

$$I(x) = \begin{cases} 1 : x \text{ is pronounced as expected} \\ 0 : \qquad\qquad \text{otherwise} \end{cases} \tag{2}$$

The range of ϕ is $[0, 1]$, and we seek to maximize ϕ. If all of the words in the bag are rendered in the expected manner, ϕ will be 1.0. Thus, contextual word accuracy measures both the phonological and morphological effects of a TTS engine producing all words in the sentence. Scoring word-level accuracy was done manually, and was a rather straighforward process. When determing

accuracy, we were tracking word stress and phonetic realization to determine whether a word was rendered correctly or not.

The second metric is *naturalness*. We evaluated the TTS engines on naturalness by synthesizing the 21 sentences and presenting them to listeners. The listening test asked participants to rate engines by placing them in ranked order from most to least human-like.

We recruited 14 participants, each of whom ranked, in descending order of preference, the seven TTS engines according to how natural they deemed the rendering to be. The participants ranged in age from 16 years to 64 years, with a median age of 28. They were asked to listen to a portion of a passage called "The Rainbow Passage" rather than the 21 test vectors used in the previous evaluations in an effort to make the grammatical artifacts that were the target of the accuracy evaluation less salient to participants. "The Rainbow Passage" is a standard reading passage, commonly used in speech evaluations, reading comprehension tests, and for testing language recognition software[12]. The result was an audio file rendered by each TTS engine. (Appendix C contains a link to these files.) In addition, the participants were asked an open-ended question: "What cues in the speech made you find it more (or less) robotic?" (Results in Appendix B.) To minimize selection bias, we explicitly chose individuals who are not in the field of linguistics, and excluded colleagues at our respective academic or industrial institutions. Instead, we chose participants who were not involved in any area related to speech technologies. To eliminate confirmation bias, each subject was presented the recordings in isolation from other participants.

We score the resulting TTS engine rankings in two ways; Condorcet voting and the Borda Count method [1]. These methods are preferred over others (e.g., averaging the votes across all participants) as they are robust and less influenced by presence of outliers. The Condorcet method selects the best candidate (i.e. TTS engine) by considering pairwise head-to-head elections among the candidates, and selects the candidate that would win the majority of the votes in all such pairwise contests. Under certain circumstances (presence of cycles in voting, e.g., A is preferred over B, B is preferred over C, C is preferred over A), the Condorcet method may not elect an authoritative winner, however, this turned out not to be the case with our voting. The Borda Count method asks participants to rank candidates in order of preference. Then, each engine, for each ballot, is given a certain number of points corresponding to the number of engines ranked lower. After counting all the votes, the candidate with the most points is the winner. The advantage of this method is that it selects a broadly acceptable candidate instead of those preferred by a majority.

[12] "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon."

5 Results and discussion

5.1 Contextual word accuracy (ϕ)

The results for contextual word accuracy are presented in Figure 1. In tabulating these results, we included Amazon Polly as our control variable as we discussed in Section 1. Appendix A shows in detail how each TTS engine fared against each test case, resulting the specific value of ϕ.

Results demonstrate that ϕ is high among commercial TTS engines, with Acapela reaching a word accuracy rate of 0.975 with minimal variance across the accuracy rate for each of the 21 sentences. Amazon Polly is a close second with an average word accuracy rate of 0.967, with some dispersion around the 1^{st} and 3^{rd} quartiles with respect to the median.

The open source engines are less accurate; the best accuracy is seen by FLITE (0.844) and the lowest accuracy by DeepVoice3 (0.761). This is surprising given that DeepVoice3 uses convolutional sequence learning and is considered a state-of-art neural speech synthesis system.

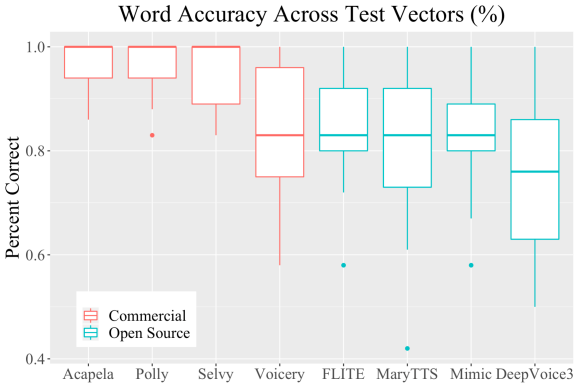

Fig 1: Contextual word accuracy (%) across the evaluation corpus.

5.2 Naturalness

A method of ranked comparison was used to evaluate naturalness. As mentioned in Section 4, we produced an audio file containing the rendering of "The Rainbow Passage" from each engine. (Appendix C contains a link to a ZIP archive of these files; DeepVoice3 only rendered 9s with what appears to be an abrupt, premature termination, and Acapela also terminates prematurely after 12s.) The participants were asked to rank the audio files and answer an open ended question, i.e., "What cues in the speech made you find it more (or less) robotic?" (Answers to the question provided by the participants are in the link in Appendix B).

The identity of each TTS engine was hidden from the participants. Instead, an opaque name ("Engine 1", ..., "Engine 7") was provided for ranking. Participants were told to rank each engine from 1 (most natural sounding) to 7 (least natural sounding), and were permitted to rank more than one TTS engine at the same level. Results of the ranking are in the table in Table 2. The result of Condorcet voting and the Borda Count indicated a unanimous winner, Voicery. Condorcet declares as winner the candidate that wins every comparison against all other candidates. Thus, for 7 candidates, Condorcet performs 21 pairwise comparisons and chose the winner to be the candidate who wins every comparison with all other candidates. That candidate is Voicery.

	Participants													
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	E7	E6, E7	E7	E7	E7	E7	E6, E7	E6	E2	E7	E7	E7	E6	E7
2	E6	E1	E4	E1, E6	E2	E2	E2	E2, E7	E7	E1	E6	E6	E2	E6
3	E1	E4	E3	E2, E4	E4	E3	E3, E5	E5	E6	E6	E1	E1	E1	E2
4	E2	E5	E2	E3, E5	E6, E3, E5	E6	E4	E1	E1	E4	E4	E5	E7	E1
5	E4	E3	E5	-	E1	E5	E1	E4	E4, E5	E2	E5	E2	E4	E4
6	E3	E2	E6	-	-	E4	-	E3	E3	E5	E3	E4	E5	E5
7	E5	-	E1	-	-	E1	-	-	-	E3	E2	E3	E3	E3

Table 2: Raw rankings of 14 participants (**En** implies TTS Engine N; a - implies that the participant did not vote for any TTS engine at that rank.)

TTS engine	Points
Engine 7 (Voicery)	53
Engine 6 (Selvy)	42
Engine 1 (Acapela)	32
Engine 2 (DeepVoice3)	32
Engine 4 (MaryTTS)	27
Engine 5 (Mimic)	20
Engine 3 (FLITE)	18

Table 3: Borda Count of TTS engines based on votes received by each engine

The Borda Count method assigns points to each candidate in the ranked lists corresponding to the number of candidates that were ranked lower. After counting all the votes, the candidate with the most points is the winner. An advantage of the Borda Count is that, in addition to declaring an absolute winner, it provides a ranking of the remaining candidates. As Table 3 shows, Voicery received the highest score with Selvy receiving the second highest. The contrast between Figure 1 and Table 3 is instructive. The ϕ value for Voicery is not the strongest as is evident from Figure 1, however, it is deemed the most naturalistic. DeepVoice3 did not receive a high score in the ϕ metric, but tied for third place in the naturalness metric. This discrepancy demonstrates a need for further research in new metrics for evaluating TTS engines.

6 Conclusion

In this paper, we study evaluating open source and commercial TTS engines on both subjective and objective measures. The two metrics used — aggregate accuracy (objective), and naturalness (subjective) — demonstrate their viability for use in business and academic contexts. We have attempted to control for the subjectivity in naturalness by using robust voting techniques such as Condorcet and Borda Count that have advantages over simple techniques like majority vote.

Our results indicate that the commercial TTS engines are superior to their open source counterparts. While some open source engines receive high marks for aggregate accuracy, they fall short on measures of naturalness. From the seven TTS engines evaluated, none emerge as a clear winner across both metrics. Acapela is the winner among commercial TTS engines with respect to ϕ (c.f., Figure 1), while FLITE gets the nod in the open source category. The naturalness metric clearly points to Voicery as the winner, but Voicery is not the preferred

engine with respect to ϕ. Assuming each metric is weighed evenly, Acapela would be declared the winner, but clearly, naturalness is an important metric where Acapela does not perform as expected.

In summary, although open source TTS engines do not reach the level of naturalness of the commercial engines, they demonstrate aggregate accuracy that show promise for deployment in business and academic settings. Future work should explore ranking TTS engines with weighted contributions of subjective and objective metrics. It could also prove interesting to evaluate the TTS engines using an automated speech recognition (ASR) system, however such a method would only evaluate accuracy as it would not be able to evaluate naturalness. Finally, future work should consider expanding the corpus we propose in Section 3, perhaps including non-English langauges, and explore evaluating the naturalness of TTS renderings through more objective measures, including feature extraction for similarity comparisions to human speech.

References

1. Pacuit, Eric. "Voting methods," The Stanford Encyclopedia of Philosophy (Available online at http://plato.stanford.edu/entries/voting-methods/) (2012).
2. Ping, Wei, et al. "Deep voice 3: 2000-speaker neural text-to-speech." arXiv preprint arXiv:1710.07654 (2017).
3. Wang, Yuxuan, et al. "Tacotron: A fully end-to-end text-to-speech synthesis model." arXiv preprint arXiv:1703.10135 (2017).
4. Yamagishi, Junichi, et al. "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm." IEEE Transactions on Audio, Speech, and Language Processing 17.1 (2009): 66-83.
5. Tribolet, J. M., et al. "A study of complexity and quality of speech waveform coders." ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 3. IEEE, 1978.
6. Möller, Sebastian, and Tiago H. Falk. "Quality prediction for synthesized speech: Comparison of approaches." Intl. Conf. on Acoustics. 2009.
7. Black, Alan W., and Keiichi Tokuda. "The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets." Ninth European Conference on Speech Communication and Technology. 2005.
8. Ma, Jeff Z., and Li Deng. "Target-directed mixture dynamic models for spontaneous speech recognition." IEEE Transactions on Speech and Audio Processing 12.1 (2004): 47-58.
9. Stoll, G., and F. Kozamernik. "A method for subjective listening tests of intermediate audio quality." ITU Working Party (2001).

Appendix A: Evaluation of TTS engines on our corpus
URL: http://www.cs.iit.edu/~vgurbani/tsd2020/appendix-a.pdf
SHA-1 Hash: b14f7632306c2c9aa4154882d97c1c829ee48224
Appendix B: Survey answers by participants
URL: http://www.cs.iit.edu/~vgurbani/tsd2020/appendix-b.pdf
SHA-1 Hash: f92c24fd84c35ee0be210801122deccf17ab0818
Appendix C: Rendering of "The Rainbow Passage"
URL: http://www.cs.iit.edu/~vgurbani/tsd2020/tsd-paper1023.zip
SHA-1 Hash: 8ef25f33b2f95300abb1e3200d0d7cc9ead856e8