INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Perception of Standard Arabic Synthetic Speech Rate

Yahya Aldholmi, Rawan Aldhafyan, Asma Alqahtani

Department of Linguistics and Translation Studies, King Saud University, Riyadh, Saudi Arabia
yaldholmi@ksu.edu.sa, rawanaldhfayan@gmail.com, y.asma.colt@gmail.com

Abstract

This experiment investigated how Arabic speakers perceive synthetic Standard Arabic speech rate produced by Google TTS, at normal vs. accelerated rates. Twenty syntactically identical Standard Arabic sentences with a similar length (*M*= 22 syllables per sentence, *SD*= 1) were auditorily presented in a female voice to thirty female participants who were instructed to rate the tempo of the normal (*M*≈ 4.5 syllable per second) and accelerated (by 10%, 20%, and 30%) stimuli on a 1-7 Likert scale (1= extremely slow, 4= normal, 7= extremely fast). The results show that differences in the four-condition synthetic speech rates were reflected in the ratings provided by the participants: the more the speech was accelerated, the higher rating it received. More importantly, the findings support the observation that the current normal speech rate of Google TTS synthetic speech is not perceived as normal by Arabic speakers, but rather is perceived as slow. This may negatively affect the likelihood that users are comfortable using this technology. Hence, the outcome of this study does not only call for further investigation into Standard Arabic synthetic speech rates, but also reveals the need to define a baseline for a natural speech rate in Arabic.

Index Terms: synthesized speech, perception, Standard Arabic, speech rate

1. Introduction

Synthetic speech research has maintained a focus on intelligibility, comprehensibility, and perception, chiefly in relation to one (or a combination of more than one) factors including environmental conditions like noise [e.g., 1], context [e.g., 2, 3], age [e.g., 4], (non)nativeness [e.g., 5, 6, 7], speech disorders and/or hearing impairment [8, 9], blindness [10], experience/ exposure and/or training [e.g., 11, 12], and speech rate [e.g., 13]. This last factor (namely, speech rate of synthetic speech) in particular has been further explored in different dimensions, mainly with other relevant factors such as age and nativeness. For instance, Higginbotham et al. [13] experimented with synthetic speech rate and showed that a slower speech rate resulted in better comprehensibility. Sutton et al. [14] investigated young and adult listeners' speech rate preferences for synthetic speech and found that both age groups are comfortable with a 150 to 200 word-per-minute (wpm) speech rate. Von Berg et al. [15] examined children and adults' perception of the speech rate of synthetic speech and discovered that adults prefer the speech rate to be faster at an average of 157 wpm, while children prefer it to be slower at a mean of 127 wpm. [16] examined the effect of the speech rate of text-to-speech recordings of a description of banking products on native and non-native speakers' comprehension of synthetic speech and found that accelerated speech resulted in lowered comprehension for both native and non-native listeners. More recently, Christenson [17] found that slowing down the speech rate of digital assistants did not have a substantial impact on intelligibility, but rather on the likelihood that users would want to utilize such technology.

One can posit four observations pertinent to the existing research on synthetic speech rate. First, a large body of the research on perception of synthetic speech rate is devoted to languages other than Standard Arabic, which severely lacks similar studies. Second, the normal rate of natural speech in the languages that have received such attention, unlike in Arabic, has been well explored and established, allowing for comparisons with that of synthetic speech; Arabic lacks this baseline. Third, stimuli acceleration or deceleration in most of these studies were either not documented in great detail or experimentally and linguistically uncontrolled. Specifically, since speech rate is a complex phenomenon that involves both linguistic and extralinguistic factors, the stimuli used for the perception experiments must factor out (and report on) potential variables such as sentence length, syntactic structure, and pauses that could influence perception of speech rate. Fourth, the unit of measure in such studies, namely words-per-minute, is not a precise measure. Many recent studies on speech rate have recruited syllable or segment/mora per second to measure speech rate [e.g., 18, 19, 20, 21].

Hence, in the current study, we seek to investigate the perceived synthetic speech rate of Standard Arabic, the variety of Arabic used in most (if not all known) digital assistants, virtual agents, and text-to-speech technologies and recently (as of July 2020) supported by the Google Text-to-Speech app. We utilize stimuli that have undergone a highly laboratorial control in terms of length, structure, prosody, and recording, recruiting single-gender (female) participants similar in age, educational background, and dialectal Arabic variety (Najdi). The study will address two overarching questions: *Is the Standard Arabic synthetic speech rate (using Google Text-to-Speech "TTS" as an exemplar) perceived as normal by native speakers of Arabic? If not, what rate is perceived as normal?* We hypothesize that current technology is perceived as slower than normal and that a minimum acceleration of 10% will be necessary for users to perceive the rate of current synthetic speech technologies as normal.

2. Methodology

2.1 Stimuli construction and recording

Twenty thematically-diverse Standard Arabic sentences were carefully constructed to be identical in syntactic structure in an attempt to avoid the impact of structural and prosodic disparities on perceived speech rate. All sentences were identically eight words in length, with a syllable rate of *M*= 22, and *SD*= 1 (note that length is measured in syllables per sentence). Each sentence was initially produced through the medium of Google TTS with a female voice as the normal rate (*M*≈ 4.5). The sentences were then accelerated (Acc) by 10%

http://dx.doi.org/10.21437/Interspeech.2021-39

three times to form four levels of speech rate: normal, 10% Acc, 20% Acc, and 30% Acc. The acceleration process was conducted using the PSOLA (Pitch-Synchronous Overlap-and-Add) feature in Praat [22], which allows for manipulation of the duration of speech while maintaining its original pitch. Table 1 below shows the speech rate for the four conditions.

Table 1: *Mean speech rate in all conditions.*

Condition	Mean speech rate (rounded)
Normal	4.5 syllables per second
10% Acc	5.0 syllables per second
20% Acc	5.5 syllables per second
30% Acc	6.0 syllables per second

2.2 Participants, task, and procedure

Thirty Arabic-speaking raters of a similar age (*range*= 21-25, *M*= 22.7, *SD*= 2.68) volunteered to participate in this study. In order to rule out any bias or influence due to gender, it was decided to match the gender of the recordings to that of the participants; hence, no male raters participated in this study. All participants speak, in addition to Standard Arabic, the same vernacular, Najdi Arabic, and have the same tertiary level of education. None of them reported any current or previous hearing impairment or loss.

A rating task was implemented in this study. The participants were instructed to rate the speed of the 20 sentences (5 sentences per condition) on a 1-7 Likert scale (with 1 being extremely slow and 7 being extremely fast). Every participant completed the task individually and was allowed to proceed at her own pace. The stimuli from all levels were presented in a randomized order, and no sentence was heard by the participant in more than one condition, nor were any two sentences from the same condition presented in a sequential order.

3. Results and discussion

The thirty participants provided 600 responses (5 per condition x 4 conditions x 20 per participant x 30 participants). A Repeated-Measures Logistic Regression was performed to determine the effect of different synthetic speech rates (as a predictor variable) on the listeners' perception (the predicted variable).

The rate of the synthetic speech had a statistically significant effect on the participants' rate perception, Wald $\chi^2(1) = 279.721, p < .001$. The odds of individuals considering the normal speech rate of synthetic speech to be slow was 0.286835 (95% CI, -5.193 to -4.069) times higher than that for the 30% accelerated speech rate, a statistically significant effect, $\chi^2(1) = 260.660, p < .001$.

The odds that the 10% accelerated speech rate was considered slow was 0.2615, 95% CI [-3.654, -2.629] times higher than that of the 30% accelerated speech rate, a statistically significant effect, $\chi^2(1) = 144.357, p < .001$. Likewise, the odds of the 20% accelerated speech rate being considered slow was 0.2297, 95% CI [-1.894, -0.993] times higher than that of the 30% accelerated speech rate, a statistically significant effect, $\chi^2(1) = 39.472, p = > .001$.

Figure 1 below shows that the more the speech was accelerated, the higher its speed ratings.

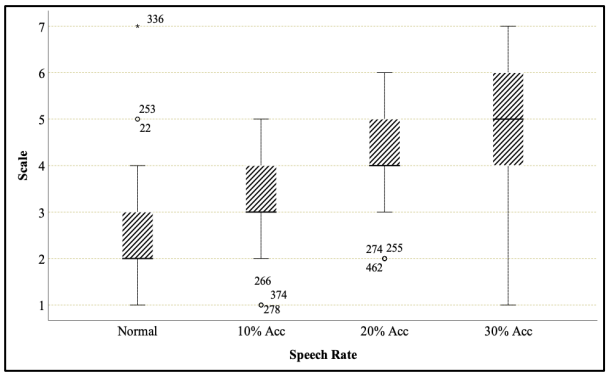

Figure 1: *Ratings for the four synthetic speech rates (normal, 10% Acc, 20% Acc, & 30% Acc).*

As shown in Table 2, the average response for the normal condition was *Slightly Slow* (*Median/Mode*= 2, *Truncated M* [*TM*]= 2.36, *M*= 2.41 *SD* = 0.97); the normal condition did not receive a *Normal* rating in more than approximately 7.33% of the responses.

Table 2: *Detailed and overall ratings for the Normal rate.*

Perceived rate	Scale points	N	Percentage
Extremely Slow	1	25	16.66%
Slightly Slow	2	58	38.66%
Slow	3	53	35.33%
Normal	4	11	7.33%
Slightly Fast	5	2	1.33%
Fast	6	0	00.00%
Extremely Fast	7	1	0.006%
Overall measures			**Value**
Median/Mode			2
Mean (SD)			2.41 (0.97)
Truncated mean (5%)			2.36

As shown in Table 3, the average given responses for the 10% accelerated speech rate was *Slow* (*Median/Mode*= 3, *TM*= 3.24, *M*= 3.24 *SD*= 0.85). It is not what the participants consider to be the normal speed. (see Table 4).

Table 3: *Detailed and overall ratings for the 10% Acc.*

Perceived rate	Scale points	N	Percentage
Extremely Slow	1	3	2.00%
Slightly Slow	2	26	17.33%
Slow	3	59	39.33%
Normal	4	56	37.33%
Slightly Fast	5	6	4.00%
Fast	6	0	00.00%
Extremely Fast	7	0	00.00%
Overall measures			**Value**
Median/Mode			3
Mean (SD)			3.24 (0.85)
Truncated mean (5%)			3.24

As reported in Table 4, the average response in ratings for the 20% accelerated speech rate was *Normal* (*Median*= 4, *TM*= 4.27, *M*= 4.25 *SD*= 0.86).

Table 4: *Detailed and overall ratings for the 20% Acc.*

Perceived rate	Scale points	N	Percentage
Extremely Slow	1	0	00.00%
Slightly Slow	2	6	4.00%
Slow	3	12	8.00%
Normal	4	82	54.66%
Slightly Fast	5	38	25.33%
Fast	6	12	8.00%
Extremely Fast	7	0	00.00%
Overall measures			**Value**
Median/Mode			4
Mean (SD)			4.25 (0.86)
Truncated mean (5%)			4.27

Table 5: *Detailed and overall ratings for the 30% Acc.*

Perceived rate	Scale points	N	Percentage
Extremely Slow	1	7	4.66%
Slightly Slow	2	13	8.66%
Slow	3	10	6.66%
Normal	4	14	9.33%
Slightly Fast	5	33	22.00%
Fast	6	48	32.00%
Extremely Fast	7	25	16.66%
Overall measures			**Value**
Median & Mode			5 & 6
Mean (SD)			4.98 (1.70)
Truncated mean (5%)			5.09

Finally, as provided in Table 5, the average response in the ratings for the 30% accelerated speech rate was *Slightly Fast* (*Median*= 5, *TM*= 5.09, *M*= 4.98 *SD*= 1.70). Note that the most frequent response was *Fast (Mode= 6)*.

Figure 2 shows that the more accelerated the speech was, the more the ratings shift to the right side of the continuum (extremely fast), but only in the 20% Acc condition do the ratings cluster and constitute a peak around *Normal*. The findings from this experiment are threefold. First, Arabic speakers are sensitive to minor increments in synthetic speech rate. Second, the speech rate of Google TTS technology is perceived as being relatively slow and not normal. Third, a speech rate that will be perceived as normal can be obtained by accelerating the current rate by a proportion between 10% and 20%. These two last points, which address the research questions and support the hypothesis, are discussed in further detail below.

Although previous research has not specifically investigated the normal produced and perceived speech rates of Standard Arabic, Gósy [23] reports that the average speech rate for Arabic ranges from 4.6 to 7.0 syllables per second. The normal speech rate produced by Google TTS in the current study is 4.5 syllables per second. This value is only slightly below the lower cut-off of the range reported in [23], but the participants consistently perceived it as being too slow. One obvious difference between Gósy and the current work is that the latter deals with synthetic speech. One would think that synthetic speech lacking some natural features would make it less intelligible than natural speech; hence, listeners would have a preference towards a slower synthetic speech. Notwithstanding, previous studies have shown the opposite. The findings from [15] albeit differences in the language and technology used, confirm that adults, but not children, prefer a faster speech rate of synthesized speech. Current findings are not easily comparable to those of [14] for two reasons.

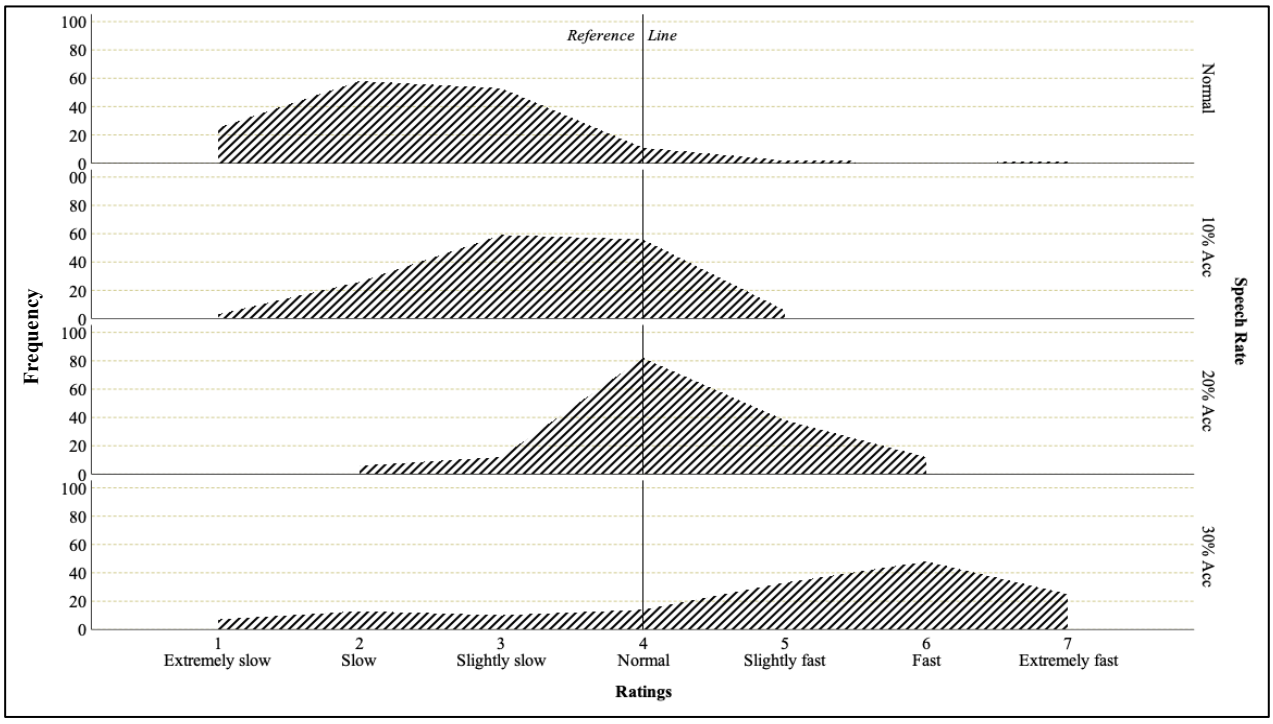

Figure 2: *Distribution of ratings across condition (note the reference line and the shift to the right side of the scale).*

First, the units of measurement are completely different (words vs. syllables and minutes vs. seconds), and any attempt to convert between measures becomes problematic due to the fact that the two languages have different morphophonological systems. Second, the sentences used in [14] were generated by MultiVoice and DECtalk, which are two relatively old technologies and were not meant to be assistive technologies for visually impaired or blind users. The quality and naturalness of speech would be presumably lower than they are now.

The highest proportion of Normal responses was obtained in the 20% Acc condition (approximately 54.66%), followed by the 10% Acc condition (approximately 37.33%), while the lowest proportion of Normal responses was obtained in the normal condition (approximately 7.33%), followed by the 30% Acc (approximately 9.33%). This shows that the preferred *normal* rate lies between 5.0 and 5.5 syllables per second, with a higher tendency for the participants to prefer 5.5 syllables per second (that is 20% Acc). This speech rate is within the range reported by Gósy [23]. The range reported by Gósy [23] seems to have been adopted from Vaane [24] who conducted an experiment on multiple languages including Arabic and reported an identical range. Vaane [24] reported the mean in syllables per second, *M*= 5, which is close to the rate reported in this study (from 5.0 to 5.5). Nevertheless, we must bear in mind that while [23] did not specify the Arabic variety reported in her study, [24] explicitly stated that the participants were natives of Moroccan Arabic, a variety that is substantially different from Standard Arabic.

4. Conclusion

The current study contributes to a body of research on speech rate in Standard Arabic, be it natural or synthetic, that is severely impoverished. The chief finding is that the current speech rate of Google TTS, a popular TT technology used by visually impaired Arabic-learning users and linked to other third-party applications, is slow. An increase in its rate is necessary, otherwise its slow speech rate may turn away users from the technology. This conclusion should assist software engineers working on speech synthesis development and establish a baseline normal speech rate for Standard Arabic. As this study was limited to normal-hearing listeners, further research may replicate the study with a focus on hearing-impaired users, a group of users who develop high-level communication skills that make them better able to process faster speech.

5. References

[1] D. Fucci, M. Reynolds, R. Bettagere, and M. D. Gonzales, "Synthetic speech intelligibility under several experimental conditions," *Augmentative and Alternative Communication,* vol. 11, no. 2, pp. 113-117, 1995.

[2] M. A. Merva and B. H. Williges, "Context, Repetition and Synthesized Speech Intelligibility," in *Proc. Human Factors Society Annual Meeting*, *vol*. 31, no. 9, Sep. 1987, pp. 961-965.

[3] K. D. Drager and J. E. Reichle, "Effects of discourse context on the intelligibility of synthesized speech for young adult and older adult listeners," *Journal of Speech, Language, and Hearing Research,* vol. 44, pp. 1052-1057, 2001.

[4] P. Mirenda and D. Beukelman, "A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups," *Augmentative and Alternative Communication*, vol. *6*, no. 1, pp. 61-68, 1990.

[5] D. M. Alamsaputra, K. J. Kohnert, B. Munson and J. Reichle,"Synthesized speech intelligibility among native speakers and non-native speakers of English," *Augmentative and Alternative Communication*, vol. *22*, no.4, pp. 258-268, 2006.

[6] M. Mack, J. Tierney, and M. E. Boyle, "The intelligibility of natural and LPC-vocoded words and sentences presented to native and non-native speakers of English," Massachusetts Inst of Tech Lexington Lincoln Lab, 1990.

[7] M. Reynolds, Z. S. Bond, and D. Fucci, "Synthetic speech intelligibility: Comparison of native and non-native speakers of English," *Augmentative and Alternative Communication*, vol.*12, no. 1, pp*. 32-36, 1996.

[8] L. M. Huntress, L. Lee, N. A Creaghead, D. D. Wheeler, and K. M. Braverman, "Aphasic subjects' comprehension of synthetic and natural speech," *Journal of Speech and Hearing Disorders*, vol. *55*, no. 1, pp. 21-27, 1990.

[9] K. A. Kangas and G. D. Allen, "Intelligibility of synthetic speech for normal-hearing and hearing-impaired listeners," *Journal of Speech and Hearing Disorders*, vol. *55, no*. 4, pp. 751-755, 1990.

[10] J. Gunderson, "Limits of intelligibility of accelerated synthesized speech by inexperienced sighted and experienced blind listeners," *in Proc. Human Factors Society Annual Meeting*, vol. 35, no.6, Sep. 1991, pp. 496-500.

[11] D. McNaughton, K. Fallon, J. Tod, F. Weiner, and J. Neisworth, "Effect of repeated listening experiences on the intelligibility of synthesized speech," *Augmentative and alternative communication*, vol.*10*, no.3, pp.161-168, 1994.

[12] M. Reynolds, C. Isaacs-Duvall, B. Sheward, and M. Rotter, "Examination of the effects of listening practice on synthesized speech comprehension," *Augmentative and Alternative Communication*, vol.*16*, no.4, pp. 250-259, 2000.

[13] D. J. Higginbotham, A. Drazek, K. Kowarsky, C. Scally, and E. Segal,"Discourse comprehension of synthetic speech delivered at normal and slow presentation rates," *Augmentative and Alternative Communication,* vol. 10, no. 3, pp. 191-202, 1994.

[14] B. Sutton, J. King, K. Hux, and D. R. Beukelman, "Younger and Older Adults Rate Performance When Listening to Synthetic Speech," *AAC: Augmentative and Alternative Communication*, vol. 11, no. 3, pp.147–153. 1994.

[15] S. Von Berg, A. Panorska, D. Uken, and F. Qeadan, "DECTALK and VERIFOX: intelligibility, likeability, and rate preference differences for four listener groups," *AAC: Augmentative & Alternative Communication*, vol. 25, no. 1, pp. 7–18, 2009

[16] C. Jones, L. Berry and C. Stevens, "Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners," *Computer Speech & Language*, vol. *21*, no.4, pp. 641–651, 2007.

[17] B. Christenson, "Speaking fast and slow: how speech rate of digital assistants affects likelihood to use," unpublished PhD Dissertation.

[18] Y. Aldholmi, "Syllable rate vs. segment rate in perceived speech rate," *in Proc. 11th International Conference of Experimental Linguistics*, Athens, Greece, 21-24, 2020.

[19] M. O'Dell and T. Nieminen, "Syllable rate, syllable complexity and speech tempo perception in Finnish," *in Proc. 19th International Congress of Phonetic Sciences,* Melbourne, Australia, 2019, pp. 622-626.

[20] L. Plug and R. Smith, "Segments, syllables and speech tempo perception," *in Proc. 9th International Conference on Speech Prosody,* Poznan, Poland, 2018, pp. 279-283.

[21] L. Plug, R. Lennon, and R. Smith, "Listeners' sensitivity to syllable complexity in speech tempo perception," in *Proc. 10th International Conference on Speech Prosody,* Tokyo, Japan, 2020.

[22] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program],* version 6.0.37, retrieved from http://www.praat.org/

[23] M. Gósy, "The perception of tempo," in M. Gósy et al. (Eds.), *Temporal factors in speech,* Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences, 1991, pp. 63-106.

[24] E. Vaane, "Subjective estimation of speech rate," *Phonetica*, vol. *39*, no. 2-3, pp.136-149, 1982.