Prosody Modifications for Question-Answering in Voice-Only Settings

Aleksandr Chuklin[1], *Aliaksei Severyn*[1], *Johanne R. Trippas*[2], *Enrique Alfonseca*[1], *Hanna Silen*[3], *Damiano Spina*[2]

[1] Google – Zürich, Switzerland
[2] RMIT University – Melbourne, Australia
[3] Google – London, UK

{chuklin, severyn, ealfonseca, silen}@google.com
{johanne.trippas, damiano.spina}@rmit.edu.au

Abstract

Many popular form factors of digital assistants—such as Amazon Echo, Apple Homepod or Google Home—enable the user to hold a conversation with the assistant based only on the speech modality. The lack of a screen from which the user can read text or watch supporting images or video presents unique challenges. In order to satisfy the information need of a user, we believe that the presentation of the answer needs to be optimized for such voice-only interactions.

In this paper we propose a task of evaluating usefulness of prosody modifications for the purpose of voice-only question answering. We propose a crowd-sourcing setup where we evaluate the quality of these modifications along multiple dimensions corresponding to the informativeness, naturalness, and ability of the user to identify the key part of the answer.

In addition, we propose a set of simple prosodic modifications that highlight important parts of the answer using various acoustic cues. Our initial results suggest that some of the modifications lead to better comprehension at the expense of slightly degraded naturalness.

Index Terms: speech generation, human-computer interaction, prosody

1. Introduction

Recent advances in technology have transformed the ways we access information. For example, with the rise of voice-only digital assistant devices, such as Amazon Echo[1], Apple Homepod[2], or Google Home[3] users can express their information need verbally and receive an answer back via voice. However, providing answers via voice in the absence of a screen is a challenging task which leads to different interaction strategies employed by both users and the system. Previous research shows that users tend to express more complex queries and engage in a dialog with the system [1], while for the system it appears to be beneficial to summarize and shorten the answers [2].

There is no commonly agreed task and evaluation guidelines for assessing prosody modifications for the task of voice-only question answering. Similar setup was used by Filippova et al. [2] for evaluating sentence compression techniques which were evaluated using human raters in terms of readability and informativeness. In contrast to this work, we propose an evaluation setup based on listening and assessing the voice-only answers across multiple dimensions.

Speech prosody is one of the major quality dimensions of synthetic speech alongside voice naturalness, fluency, and intelligibility [3]. It refers to the suprasegmental characteristics of speech such as tune and rhythm. Acoustically, prosody manifests itself in pitch, duration, intensity, and spectral tilt of speech.

Prosody has an important cognitive role in speech perception. Sentence stress seems to ease comprehension of stressed words and has been shown to lower reaction time independent of word's syntactic function [4]. Human listeners attend to those word onsets they are least able to predict [5] and high activation levels allow extra resources to be allocated for processing these words [6]. At signal-level, low-probability regions of pitch and energy trajectories show strong correlation with the perception of stress, providing further evidence of the connection between attention and unpredictability [7].

Pauses in speech convey information about intonational boundaries [8] and changes in pausing can alter syntactic parsing of a sentence [9]. However, interruptions also have a role in comprehension. Filler words, pauses, or even artificial tones have been reported to improve the human word recognition [10]. This kind of natural delays and filler words are frequent in spontaneous speech but typically omitted in synthetic speech.

While these features of prosody in the natural speech have been associated with positive effects, it remains unclear which effects it would have when incorporated in a speech synthesis system and how these effects can be evaluated at scale. We propose to address this problem by asking the following research questions:

RQ1 Can we use crowdsourcing to quantify the utility of the prosody modifications for voice-only question-answering?

RQ2 Which effects do prosody modification techniques have on informativeness and perceived naturalness of the audio response?

2. Methodology

Let us assume that for a user's *question* we have a *short answer* as well as a *support sentence*, which provides more context to the answer.[4] Table 1 provides an example of such a tuple. This pattern is used by commercial search engines, e.g., Google's featured snippets [11] or Bing Distill answers [12], where most important parts of the answer are highlighted or called-out separately. Additionally, there are datasets available for researchers to study text-based question answering, such as the Stanford

[1] https://www.amazon.com/echo
[2] https://www.apple.com/homepod/
[3] https://madeby.google.com/home

[4] We use the terms *answer* and *short answer* interchangeably.

Question-Answering Dataset (SQuAD) [13] or Microsoft Machine Reading Comprehension Dataset (MS MARCO) [14].

When surfacing an answer to a user's query on display it is possible to use visual cues, such as highlighting or bolding of key answer phrases or important terms, which may ease and speed up comprehension of the answer. In contrast, when serving voice-only answers one could employ prosody modifications to the speech to cue the user about the key answer phrase—the *short answer*—in the support sentence.

Table 1: Example question and short answer within the support sentence which is provided via audio to the user.

Question	Which guitarist inspired Queen?
Support Sentence	Queen drew artistic influence from British rock acts of the 60s [...] in addition to American guitarist **Jimi Hendrix**, with Mercury also inspired by the gospel singer Aretha Franklin.
Short Answer	Jimi Hendrix

In the current work we propose to use *crowd workers* to evaluate the prosody modifications. Given a *question* and a verbalization of a corresponding *support sentence* (see Figure 1), crowd workers need to identify the phrase or a token span in the audio that serves as a *short answer* to the user's question as well as give feedback on the quality of the audio response (see the Evaluation section below).

We hypothesize that highlighting the short answer by modifying the prosody, during the audio generation step, makes it easier for the worker to identify the answer and makes the whole response more relevant, potentially at the expense of naturalness of the audio.

Prosody modification. We perform four different prosody modifications in the Text-to-Speech (TTS) generation:

- **pause**: inserted before and after the short answer;
- **rate**: the speaking rate of the short answer is decreased;
- **pitch**: the short answer is spoken in a higher pitch than the rest of the support sentence;
- **emphasis**: the short answer is spoken with prominence. Emphasis is typically implemented as a combination of prosody modifications such as speaking **rate** and **pitch**.

The audio using TTS with no intervention is used as a baseline.

Collection of judgments. The judgments were collected using the CrowdFlower[5] crowdsourcing platform for the different audio responses.[6] Figure 1 shows a question-response pair as presented in the crowdsourcing interface. Question-audio pairs (and the different prosody modifications in the audio response) are randomly assigned by the platform to workers resident in English-speaking countries.

[5] http://www.crowdflower.com

[6] All experiments were performed under Ethics Application BSEH 10-14 at RMIT University.

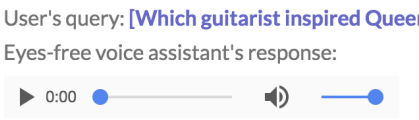

Figure 1: Question and audio response as presented to crowdsourced workers.

Evaluation. We study the following dimensions to evaluate the utility of short answer highlighting via prosody modifications: *informativeness* (how satisfactorily the audio response answers the user's question), *elocution* (were the words in the full support sentence pronounced correctly), appropriateness of the audio *length*, and presence of unwarranted *interruptions*. These dimensions are based on the guidelines for evaluating speech in Google Assistant [15], and the exact version used for crowdsourcing will be released together with the paper.

In addition to collecting the aforementioned explicit judgments we also calculate the *correctness* of the workers' answers. In order to compute correctness we compare the answer typed by the worker against the given short answer from the dataset. We do this by converting both into a Metaphone representation [16] to account for typos and misheard words, we then compute the difference using Ratcliff-Obershelp algorithm [17]. The value of this metric ranges from 0 to 1.

Quality control. In order to detect whether a worker is reliable, we use two different types of *test questions*: (i) we ask the worker to type in the short answer after listening to the full audio and compare the provided short answer against the ground-truth, and (ii) we include questions that are off-topic and do not contain an answer. In the first case we filter out workers who achieve answer *correctness* below 0.5 while in the second case we expect the lowest rating on the *informativeness* scale instead.

In order to commence annotating actual tasks, workers have to successfully complete the initial task comprised of three test questions. In addition, we limit the number of tasks each worker is allowed to contribute to avoid attracting spammers.

Note that the question of how much context is actually needed to support the answer is beyond the scope of the current paper. One can imagine different strategies for different scenarios[7], however, in general, we presume that providing context is useful to corroborate the short answer, or provides a way for the user to identify incorrect answers when this is the case.

3. Experimental Setup

In our study we use question/audio pairs obtained from the SQuAD question-answering dataset [13]. The SQuAD dataset consists of crowdsourced questions related to a set of Wikipedia articles. Each Wikipedia article has a set of questions. Answers for those questions correspond to text segments in sentences of the corresponding article. In our experiments, the sentence containing the short answer is fed to a TTS to generate the audio response.

In particular, for our set of experiments we used the first 300 Wikipedia articles and their corresponding question/audio pairs. We further split these 300 articles into four groups of 75 question/audio pairs (one group per modification: pause, rate, pitch, and emphasis). Choosing different questions for each

[7] E.g., for different question/answer complexity, answer quality or user context (at home, on the go, etc.).

Table 2: Prosody modification settings: `strength` parameter of the `<break>` SSML tag, `rate` / `pitch` parameters of `<prosody>`, and `level` parameter of `<emphasis>`.

TTS engine	Voice	**pause**	**rate**	**pitch**	**emphasis**
IBM	Lisa	`strong`	`x-slow`	`x-high`	n/a
Google	en-US-Wavenet-F	`strong`	`slow`	`+2st`	`strong`

Table 3: Worker agreement scores as measured by Krippendorff's α [23] and ratio of items where majority (two out of three) raters agree on the rating.

score	α (IBM)	α (Google)	maj. (IBM)	maj. (Google)
inform.	0.27 to 0.31	0.06 to 0.22	0.84 to 0.87	0.79 to 0.89
elocution	0.15 to 0.27	–0.04 to 0.08	0.87 to 0.97	0.99 to 1.00
interrupt.	0.00 to 0.08	–0.01 to 0.12	0.99 to 1.00	1.00 to 1.00
length	0.27 to 0.37	0.17 to 0.43	0.97 to 0.99	0.99 to 1.00

prosody modification reduces the chance that each crowdworker is exposed to the same question many times. We then generated the audio of the baseline and modified versions of the support sentence with the included short answer. Each of the resulting question/audio pairs were rated by three crowd workers.

We use two different TTS platforms: IBM Watson [18, 19] and Google Wavenet [20, 21]. The settings are summarized in Table 2.[8] Note that these settings are chosen ad-hoc based on the subjective perception of the authors. The perceived size of the effect depends on the TTS engine and voice used, as well as on the support sentence being modified. We leave the optimization of the level of prosodic modifications for the future work.

The following judgments were collected using the CrowdFlower crowdsourcing platform. After removing judgments used for quality control, we have 1,454 rows of judgments for the IBM engine from 99 workers for 450 question-audio pairs (75 for each of the three modifications (emphasis was not collected for the IBM engine), plus equal number of baseline pairs) for the IBM engine; 1,820 rows of judgments from 85 workers for 600 question-audio pairs (75 for each of the four modifications, plus equal number of baseline pairs) for the Google TTS engine. Minimum three judgments per question-audio pair were collected for each variation.

The agreement between crowd workers is rather low when measured by Krippendorff's alpha [23], especially for *elocution* and *interruption* scores (e.g., only for **pauses** modification did we have meaningful agreement for *interruption* score). For *informativeness* and *length* the scores are low, but are comparable with similar crowdsourcing judgment collections [24]. When it comes to majority agreement, however, it was substantially high across all dimensions/modifications/voices, meaning that two out of three crowdworkers selected the exact same rating label for almost all items.

Judgments are converted to Likert scale and, in case of *length*, the absolute deviation from the optimal length score is taken, effectively making it binary. We use median to aggregate judgments per item. Given that the ratings are not on the interval scale, we use the Wilcoxon signed-rank test [25] on per-item level to report statistical significance. We use * (**) to indicate statistical significance with $p < 0.05$ ($p < 0.01$ respectively).

[8] The `emphasis` feature is currently only available in the Google TTS engine. The exact combination of prosody parameters is not specified by the standard [22] and left unspecified in the documentation [21].

Almost identical results were obtained when t-test was used.

4. Results and Discussion

Table 4 shows the difference in terms of the judgments obtained for the proposed prosody modifications, using the two TTS systems. Note that the results are not comparable across TTS engines as the prosody modifications have noticeably different effect and no effort was put into optimizing those.

Table 4: Difference between **base** and various prosody modifications. The higher the better for *informativeness*, *elocution* and *correctness*; the lower the better for *interruption*, and *length*.

	inform.↑	*correct.*↑	*elocution*↑	*interrupt.*↓	*length*↓
IBM					
pauses	-0.21	$+0.04$	-0.03	$+0.37^{**}$	$+0.08$
rate	$+0.26$	$+0.02$	-0.24^{**}	$+0.03$	$+0.03$
pitch	$+0.02$	-0.03	-0.11	$+0.01$	-0.03
Google					
pauses	$+0.21$	$+0.09$	-0.04	$+0.15^{**}$	$+0.00$
rate	$+0.22$	$+0.07$	-0.18^{**}	$+0.18^{**}$	$+0.03$
pitch	-0.03	$+0.08$	$+0.08$	$+0.13^{**}$	$+0.07$
emphasis	$+0.87^{**}$	$+0.28^{**}$	-0.07	$+0.13^{**}$	-0.07

The main pattern that emerges from the data is an increase in informativeness and correctness, and a decrease in speech quality.

In general, all modifications compromise the naturalness of the speech as captured by *elocution* or *interruption* ratings. Interestingly, only **rate** modification was deemed to significantly hurt elocution and no significant *length* change was reported. As expected, workers identified more unexpected *interruptions* when **pauses** are used to highlight the short answer in both TTS engines. There are also *interruptions* reported for other modifications in the Google engine. This is due to the implementation details which lead to sentence breaks—and therefore small pauses—around `<prosody>` and `<emphasis>` tags.

We also observe that our prosody modifications either improve or leave the *correctness* score unchanged, and most of them—although not all—are perceived by workers as more useful for the job of identifying the short answer.

Next, we look at whether particular prosody modifications are especially effective (or not) on certain slices of the data. By splitting the data by median length or median position of the short answer results in roughly halving the data into two similarly-sized slices in each case.

Table 5 demonstrates how different prosody modifications strategies perform depending on the length of the short answer. Table 6 shows how results change depending on the short answer offset from the end of the support sentence. Both offset and length were measured in number of words, but similar results were obtained when measuring the offset in characters.

As we can see from Table 5, **rate** change in the IBM engine is perceived to increase *informativeness* for shorter answers, while **pitch** appears to work better for longer ones. For the Google engine we observe a similar pattern for **pauses** which are very effective for shorter answers, but even hurt for longer ones as measured by both subjective (*informativeness*) and objective (*correctness*) scores. On the other hand, **emphasis**, have positive effect on both slices. We also see that the effect of undesirable interruptions in the Google engine is not noticeable for longer answers. Note that implementations differ between TTS engines, especially

Table 5: Difference between base and the various prosody modification and various answer length. "Short" answers are equal in length or shorter than the median answer (two words), while "long" is the rest.

	inform.↑	*correct.*↑	*elocution*↑	*interrupt.*↓	*length*↓
IBM (short)					
pauses	−0.12	+0.09	−0.05	+0.37**	+0.16
rate	+0.48	+0.07	−0.16	+0.00	−0.10
pitch	−0.14	−0.08	−0.06	+0.02	−0.05
IBM (long)					
pauses	−0.33	−0.03	+0.00	+0.36**	−0.03
rate	−0.02	−0.05	−0.35*	+0.06	+0.20
pitch	+0.24	+0.05	−0.18	+0.00	−0.02
Google (short)					
pauses	+1.09**	+0.27**	−0.07	+0.19**	−0.05
rate	+0.11	+0.08	−0.19**	+0.25**	+0.03
pitch	−0.05	+0.10	+0.02	+0.16**	+0.07
emphasis	+0.54	+0.21*	−0.12	+0.17**	−0.05
Google (long)					
pauses	−0.97*	−0.16*	+0.00	+0.09	+0.06
rate	+0.38	+0.06	−0.16*	+0.09	+0.03
pitch	+0.00	+0.05	+0.16	+0.10	+0.06
emphasis	+1.26**	+0.36**	+0.00	+0.09	−0.09

for **rate** and **pitch** settings where we were unable to exactly match the perceived strength of the modifications in two engines.[9]

Table 6: Difference between base and the various prosody modifications and answer offsets from the end. "Easy" answers have equal or less than median offset to the end, while "hard" are the rest.

	inform.↑	*correct.*↑	*elocution*↑	*interrupt.*↓	*length*↓
IBM (easy)					
pauses	+0.02	+0.08	−0.05	+0.31**	+0.06
rate	+0.60	+0.12*	−0.26*	+0.06	−0.03
pitch	−0.20	−0.08	−0.10	+0.02	+0.01
IBM (hard)					
pauses	−0.53	−0.02	+0.00	+0.45**	+0.11
rate	−0.02	−0.07	−0.23*	+0.00	+0.07
pitch	+0.28	+0.04	−0.12	+0.00	−0.09
Google (easy)					
pauses	+0.61	+0.15*	−0.07	+0.15**	−0.04
rate	+1.01**	+0.24**	−0.15*	+0.19**	−0.07
pitch	+0.09	+0.08	+0.07	+0.14*	+0.12
emphasis	+1.23**	+0.35**	−0.16	+0.13*	−0.03
Google (hard)					
pauses	−0.41	−0.02	+0.00	+0.14*	+0.07
rate	−0.49	−0.07	−0.20**	+0.17**	+0.13
pitch	−0.19	+0.07	+0.09	+0.12*	+0.00
emphasis	+0.61	+0.23*	+0.00	+0.14*	−0.09

Next, Table 6 shows a breakdown by the offset to the end of the audio. Intuitively, the answers that are closer to the end of the audio should be easier to understand. For these answers we observed only weak significant improvemend of the **rate** modification for the IBM engine as measured by *correctness* and, to a certain extent, by *informativeness*. On the same slice the Google engine benefits from both **rate** and **emphasis** modifications as measured by both *informativeness* and *correctness*.

The **emphasis** modification in the Google engine obtains the highest (and statistically significant) gain in terms of *informativeness* and *correctness* (see Figure 2 for a visual representation of the distribution shift). That means that the combination of prosody modifications developed by the engineers of this feature outperforms modifications of just one dimension. Further exploration of how to combine these modifications needs to be done.

[9] We selected maximum strength for **rate** and **pitch** in the IBM TTS engine, but the settings with the same labels in the Google TTS were subjectively too strong and unnatural, so we chose to tune them down.

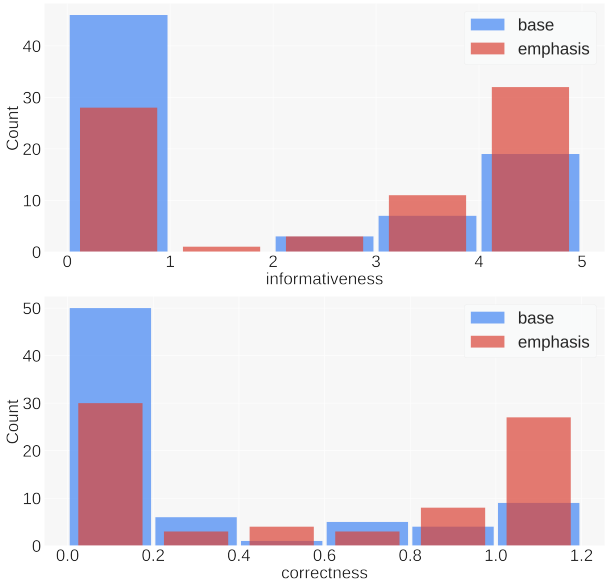

Figure 2: Distributions of the *informativeness* (above) and *correctness* (below) of the audio with emphasis (red narrow bars) vs. the baseline (blue wider bars). Higher scores are better.

5. Conclusions

We conducted a crowdsourcing experiment to investigate how prosody modifications can help users to identify answers from audio responses in a question answering setting. To answer our first research question outlined in the introduction, we conclude that, yes, this setup is viable and gives an actionable breakdown of quality dimensions. To our knowledge, this is the first experiment that validates the use of a crowdsourcing methodology to analyze prosody modification in voice-only question answering.

Answering our second research question, we showed that emphasizing the answer—via lowering speaking rate and increasing pitch—provides subjectively more informative responses and makes workers more effective in identifying the answers, at the expense of the naturalness in the audio (interruptions).

Results suggest that further studies are needed to better understand the optimal combination of prosody modification to highlight answers in a given audio response. An open question for future work is to better understand how modifying the prosody impacts the users' satisfaction in a more general context, such as when users are not asked to extract answers.

6. References

[1] J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson, "Informing the design of spoken conversational search," in

CHIIR'18 Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval. ACM, 2018, pp. 32–41.

[2] K. Filippova, E. Alfonseca, C. A. Colmenares, L. Kaiser, and O. Vinyals, "Sentence compression by deletion with lstms," in *EMNLP'15 Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 360–368.

[3] F. Hinterleitner, C. Norrenbrock, and S. Möller, "Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech," in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[4] A. Cutler and D. J. Foss, "On the role of sentence stress in sentence processing," *Language and Speech*, vol. 20, pp. 1–10, 1977.

[5] L. B. Astheimer and L. D. Sanders, "Predictability affects early perceptual processing of word onsets in continuous speech," *Neuropsychologia*, vol. 49, pp. 3512–3516, 2011.

[6] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, pp. 425–452, 2010.

[7] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive Science*, vol. 40, no. 7, pp. 1739–1774, 2016.

[8] A. Pannekamp, U. Toepel, K. Alter, A. Hahne, and A. D. Friederici, "Prosody-driven sentence processing: An event-related brain potential study," *Journal of Cognitive Neuroscience*, vol. 17, pp. 407–421, 2005.

[9] K. G. Bailey and F. Ferreira, "Disfluencies affect the parsing of garden-path sentences," *Journal of Memory and Language*, vol. 49, no. 2, pp. 183–200, 2003.

[10] M. Corley and R. J. Hartsuiker, "Why um helps auditory word recognition: The temporal delay hypothesis," *PLOS ONE*, vol. 6, no. 5, pp. 1–6, 05 2011.

[11] "Google's Featured Snippets," https://www.blog.google/products/search/reintroduction-googles-featured-snippets/, accessed: 2018-03-21.

[12] B. Mitra, G. Simon, J. Gao, N. Craswell, and L. Deng, "A proposal for evaluating answer distillation from web data," in *Proceedings of the SIGIR 2016 WebQA Workshop*, 2016.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *EMNLP'16 Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 2383–2392.

[14] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," *arXiv preprint arXiv:1611.09268*, 2016.

[15] "Evaluation of Speech for the Google Assistant," https://research.googleblog.com/2017/12/evaluation-of-speech-for-google.html, accessed: 2018-03-21.

[16] L. Philips, "The double metaphone search algorithm," *C/C++ users journal*, vol. 18, no. 6, pp. 38–43, 2000.

[17] J. W. Ratcliff and D. E. Metzener, "Pattern matching: the gestalt approach," *Dr. Dobb's Journal*, vol. 13, no. 7, p. 46, 1988.

[18] A. Sorin, S. Shechtman, and A. Rendeli, "Semi parametric concatenative tts with instant voice modification capabilities," *Proc. Interspeech 2017*, pp. 1373–1377, 2017.

[19] "IBM Watson Text to Speech," https://www.ibm.com/watson/services/text-to-speech/, accessed: 2018-03-21.

[20] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[21] "Cloud Text-to-Speech - Google Cloud," https://cloud.google.com/text-to-speech/, accessed: 2018-03-27.

[22] "Speech Synthesis Markup Language (SSML) Version 1.1," https://www.w3.org/TR/speech-synthesis11/, accessed: 2018-03-21.

[23] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[24] A. Chuklin and M. de Rijke, "Incorporating clicks, attention and satisfaction into a search engine result page evaluation model," in *CIKM'15 Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 175–184.

[25] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.