Using Audio Transformations to Improve Comprehension in Voice Question Answering*

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

[1] Google Research {chuklin, severyn, ealfonseca, silen}@google.com
[2] RMIT University {johanne.trippas, damiano.spina}@rmit.edu.au

Abstract. Many popular form factors of digital assistants—such as Amazon Echo or Google Home—enable users to converse with speech-based systems. The lack of screens presents unique challenges. To satisfy users' information needs, the presentation of answers has to be optimized for voice-only interactions. We evaluate the usefulness of audio transformations (i.e., prosodic modifications) for voice-only question answering. We introduce a crowdsourcing setup evaluating the quality of our proposed modifications along multiple dimensions corresponding to the informativeness, naturalness, and ability of users to identify key parts of the answer. We offer a set of prosodic modifications that highlight potentially important parts of the answer using various acoustic cues. Our experiments show that different modifications lead to better comprehension at the expense of slightly degraded naturalness of the audio.

Keywords: speech generation · question answering · crowdsourcing

1 Introduction

Recent advances in technology have transformed the ways we access information. With the rise of voice-only digital assistant devices, such as Amazon Echo, Apple Homepod, or Google Home users can express information needs verbally and receive answers exclusively 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.

Searching is traditionally considered as a visual task since reading information-dense sections such as search snippets is already a cognitively demanding undertaking. Thus, screen-based systems typically provide visual cues to *highlight* key parts of text responses (e.g., boldfacing key parts in passages) which helps to identify answers while skimming a results page. However, the serial nature of audio-only communication channels hampers "skimming" the information as can be done in a visual interface.

In this paper, we explore different prosody modifications—such as insertion of pauses, decreasing of speaking rate, and increase in pitch—to highlight key answer parts in audio responses. While these features of prosody in natural speech have been associated with positive effects, to our knowledge they have not been analysed empirically for presenting answers in voice-only channels.

* For extended version of this paper, please refer to Chuklin et al. [2].

Moreover, it remains unclear which effects it would have when incorporated in a voice Question Answering (QA) system and how these effects can be evaluated at scale. We propose to address the problem by asking the following questions:

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

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

Related Work. Most of the related work on QA systems with speech interfaces focuses on the problem of spoken language recognition and understanding of voice-based questions [5, 6, 14]. The scope of our work is to better understand how to *present* answers when delivered via the audio-channel.

In contrast to traditional desktop search, there are no commonly agreed task and evaluation guidelines for assessing *voice-only* QA. Filippova et al. [4] proposed to evaluate sentence compression techniques in terms of readability and informativeness using human raters. In contrast to that work, we propose an evaluation setup where raters are asked to *listen* and assess the voice answers across *multiple* dimensions, as well as to *extract* the key answer part, which we check for correctness.

The audio modifications presented in this paper alter the *prosody* of the spoken answer (i.e., the patterns of stress and intonation in speech). Prosody has an essential cognitive role in speech perception [13]. Sentence stress seems to ease comprehension of stressed words and has been shown to lower reaction time independent of a word's syntactic function [3]. Simultaneously, pauses in speech convey information about intonational boundaries [9].

2 Methodology

Assume that for a user's *question* we have an *answer sentence* where we identify the *answer key* (key answer part) with the help of some algorithm. Example:

- *Question*: Which guitarist inspired Queen?
- *Answer Sentence*: Queen drew artistic influence from British rock acts of the 60s and early 1970s [...] in addition to American guitarist **Jimi Hendrix**, with Mercury also inspired by the gospel singer Aretha Franklin.
- *Answer Key*: Jimi Hendrix

This pattern is used by commercial search engines, e.g., Google's featured snippets[3] or Bing Distill answers [7], where the most important parts of the answer are highlighted or called-out separately. Additionally, there are datasets available for researchers to study text-based QA, such as MS MARCO [8] or the Stanford Question-Answering Dataset (SQuAD) [11], which we use here.

We hypothesize that highlighting the key answer part by modifying the prosody during the audio generation step, makes it easier for the worker to understand the answer, potentially at the expense of naturalness of the audio.

The problem of identifying key parts is an active area of research in QA and is beyond the scope of the current work. Note that, unlike human-curated datasets

[3] https://blog.google/products/search/reintroduction-googles-featured-snippets

mentioned above, the quality of the automatically extracted answer keys may not be high enough for them to be surfaced as stand-alone answers in a production system. Without the context of the entire sentence, the risk of misleading the user by a low-quality short answer is high. This potential risk motivates our work on how to *emphasize* the key part of the answer when presented via voice.

We propose to ask *crowd workers* to evaluate the prosody modifications. Given a *question* and verbalization of a corresponding *answer sentence*, crowd workers need to give feedback on the quality of the audio response as well as identify the phrase in the audio that corresponds to the answer key. The judgments were collected using the Figure Eight crowdsourcing platform.[4] Tasks were randomly assigned to paid workers residing in English-speaking countries. The dataset and crowd judgments can be accessed at https://github.com/varepsilon/clef2019-prosody.

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

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

Evaluation. We study the following four explicit dimensions to evaluate the utility of highlighting via prosody modifications and naturalness of the audio response: *informativeness* (how satisfactorily the audio-response answers the user's question on the scale of 0 to 4), *elocution* (whether the words in the full answer sentence were pronounced correctly, 0 to 2), presence of unwarranted *interruptions* (0 or 1), appropriateness of the audio *length* (-1 to 1). These dimensions are based on the guidelines for evaluating speech in the Google Assistant.[5]

In addition to collecting the aforementioned judgments, we also calculate one objective measure, the *correctness* of the workers' typed answer key. To compute correctness, we compare the answer key typed by the worker against the given short answer from the dataset (what we treat as the gold answer key for highlighting). We convert both into a Metaphone representation [10] to account for typos and misheard words, and then compute the difference using the Ratcliff-Obershelp algorithm [12]. The *correctness* value ranges from 0 to 1.

Quality control. 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 then 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 worker to give the lowest rating on the *informativeness* scale.

[4] Experiments performed under Ethics Application BSEH 10-14 at RMIT University.

[5] https://ai.googleblog.com/2017/12/evaluation-of-speech-for-google.html

Table 1. 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	Wavenet-F	`strong`	`slow`	`+2st`	`strong`

3 Experimental Setup

In our study, we use question/answer pairs from the widely used SQuAD [11]. In our experiments, whole SQuAD paragraph were fed to a TTS generating the audio response and the ground truth answers are used to highlight the key part of it. For our set of experiments, we used the first 300 Wikipedia articles and their corresponding question/audio pairs. We further split these articles into four groups of 75 question/audio pairs (one group per modification: pause, rate, pitch, and emphasis). Different articles were used for each prosody modification to reduce the chance that a crowd worker is exposed to the same question multiple times. We then generated the audio of the baseline (no modifications) and modified versions of the answer sentence. Three crowd workers rated each of the resulting question/audio pairs.

We use two TTS platforms in our experiments: IBM Watson (`https://ibm.com/watson/services/text-to-speech`) and Google Wavenet (`https://cloud.google.com/text-to-speech`). The settings are summarized in Table 1.[6] Note that these settings are chosen ad-hoc based on the subjective perception and test runs. The perceived size of the effect depends on the TTS engine and voice used, as well as on the sentence being modified. We leave the optimization of these settings for future work.

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 (pause, rate, pitch), plus an equal number of baseline pairs); 1,820 rows of judgments for the Google TTS engine from 85 workers for 600 question-audio pairs (four modification plus baseline).

Agreement between crowd workers is rather low when measured by the Krippendorff's alpha. For *informativeness* and *length* the scores are low (ranging from 0.27 to 0.37 for the IBM engine, and from 0.06 to 0.43 for Google TTS), but are comparable with similar crowdsourcing judgment collections [1]. The agreement is even lower for *elocution* and *interruption* scores. When it comes to majority agreement (two out of three workers), however, it was substantially high across all dimensions/modifications/voices (above 0.79 for the lowest slice).

Judgments are treated as Likert scale and, in case of *length*, the absolute value is taken, making it binary ("OK" vs. "too short/too long"). We use the median to aggregate judgments per item. Wilcoxon signed-rank test on a per-item level was used to report statistical significance. We use * (**) to indicate

[6] The `emphasis` feature is currently only available in the Google TTS and the implementation details are not specified in the SSML standard nor the documentation.

statistical significance with $p < 0.05$ ($p < 0.01$ respectively). Equivalent results were obtained when the t-test and/or average instead of median was used.

4 Results and Discussion

Table 2 shows the result. We only report absolute difference in the score given by the raters to avoid a direct comparison between two commercial systems. Note also that the results are not comparable across two systems because the prosody modifications with the same name have a noticeably different effect on them.

Table 2. Absolute difference relative to the baseline. The higher the better for *inform.*, *correctness*, and *elocution* (↑); the lower the better for *interruption* and *length* (↓).

		inform.↑	*correctness*↑	*elocution*↑	*interruption*↓	*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 through naturalness, as captured by *elocution* or *interruption* ratings. Interestingly, only **rate** modification was deemed to significantly hurt elocution and no significant *length* change were reported. As expected, workers identified more unexpected *interruptions* when **pauses** are used to highlight the answer keys in both TTS engines. There were also *interruptions* reported for other modifications in the Google TTS engine, which is due to the peculiarity of that engine, which always adds sentence breaks—and therefore small pauses—around `<prosody>` and `<emphasis>` tags. We expect that once that issue is resolved, no interruptions will be reported.

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 answer (*informativeness*).

5 Conclusions

We investigate how prosody modifications can help users to identify answers from audio responses in a QA setting. To answer our first research question (**RQ1**) we conclude that, yes, the proposed crowdsourcing 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 QA.

Answering our second research question (**RQ2**), we show that emphasizing the answer—via lowering speaking rate and simultaneously 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), which is an artefact of a particular TTS implementation.

The near future work includes further studies to find the optimal combination of prosody modification to highlight answers in a given audio response depending on the different answer features (and possibly on the user features). Another open question for future work is to better understand how modifying the prosody impacts the users' comprehension and satisfaction in a more general context, such as when users are not asked to extract answers and converse naturally.

References

1. Chuklin, A., de Rijke, M.: Incorporating Clicks, Attention and Satisfaction into a Search Engine Result Page Evaluation Model. In: CIKM (2016)
2. Chuklin, A., Severyn, A., Trippas, J.R., Alfonseca, E., Silen, H., Spina, D.: Prosody modifications for question-answering in voice-only settings. CoRR abs/1806.03957 (2018), http://arxiv.org/abs/1806.03957
3. Cutler, A., Foss, D.J.: On the Role of Sentence Stress in Sentence Processing. Language and Speech 20, 1–10 (1977)
4. Filippova, K., Alfonseca, E., Colmenares, C.A., Kaiser, L., Vinyals, O.: Sentence Compression by Deletion with LSTMs. In: EMNLP (2015)
5. Kumar, A.J., Schmidt, C., Köhler, J.: A Knowledge Graph-Based Speech Interface for Question Answering systems. Speech Communication (2017)
6. Mishra, T., Bangalore, S.: Qme!: A Speech-based Question-answering System on Mobile Devices. In: Proceedings of NAACL'10. pp. 55–63 (2010)
7. Mitra, B., Simon, G., Gao, J., Craswell, N., Deng, L.: A Proposal for Evaluating Answer Distillation from Web Data. In: Proceedings of the SIGIR 2016 WebQA Workshop (2016)
8. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A Human Generated MAchine Reading Comprehension Dataset. arXiv preprint arXiv:1611.09268 (2016)
9. Pannekamp, A., Toepel, U., Alter, K., Hahne, A., Friederici, A.D.: Prosody-driven Sentence Processing: An Event-related Brain Potential Study. Journal of Cognitive Neuroscience 17, 407–421 (2005)
10. Philips, L.: The Double Metaphone Search Algorithm. C/C++ Users Journal 18(6), 38–43 (2000)
11. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: EMNLP (2016)
12. Ratcliff, J.W., Metzener, D.E.: Pattern Matching: the Gestalt Approach. Dr. Dobb's Journal 13(7), 46 (1988)
13. Sanderman, A.A., Collier, R.: Prosodic phrasing and comprehension. Language and Speech 40(4), 391–409 (1997)
14. Whittaker, E.W.D., Mrozinski, J., Furui, S.: Factoid Question Answering with Web, Mobile and Speech Interfaces. In: NAACL (2006)