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] Damiano Spina[2]

[1]Google Research (Zürich, Switzerland & London, UK)
[2]RMIT University (Melbourne, Australia)

Highlighting Answers in Voice

Question	Which guitarist inspired Queen?
Answer 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.
Answer Key	Jimi Hendrix

Research Questions

- Can crowdsourcing measure the effect of voice highlighting?
- Which prosody modifications work best?
- What type of answers benefit the most?

Highlighting Techniques

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

SSML Markup

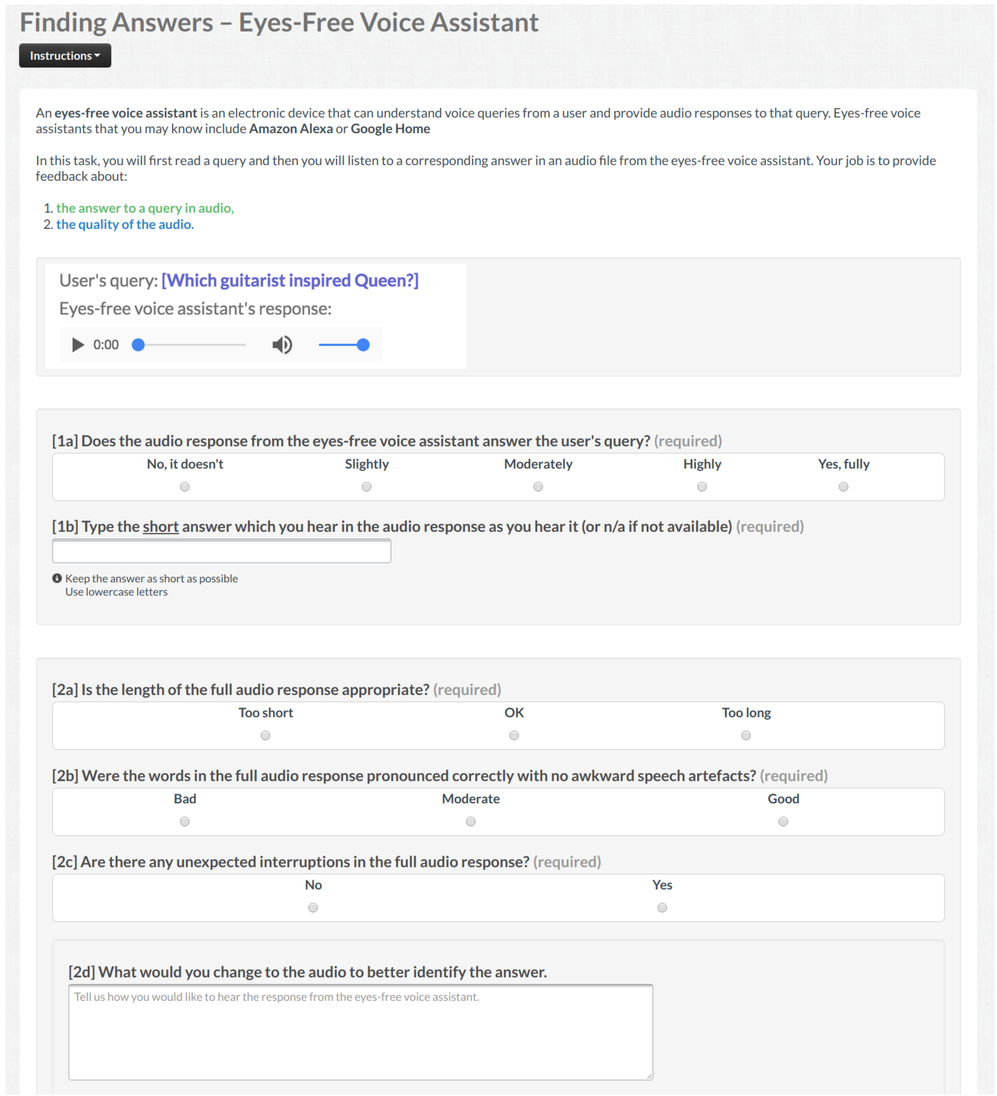
```
<speak>Queen drew artistic influence from British
rock acts of the 60s [...]  in addition to
American guitarist <prosody rate="slow">Jimi
Hendrix</prosody> with Mercury also inspired by
the gospel singer Aretha Franklin.</speak>
```

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`

Crowdsourcing Setup



https://github.com/rmit-ir/clef2019-prosody

Results

Difference relative to the baseline (no modifications); absolute value. The higher the better for *informativeness*, *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

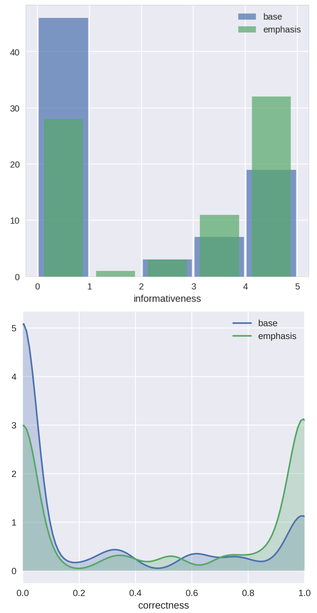


Figure: Distributions of the *informativeness* (left) and *correctness* (right) of the audio with emphasis (red narrow bars) vs. the baseline (blue wider bars). Higher scores are better.

Conclusions

- Crowdsourcing is viable if correctness is verified
- A combination of rate/pitch (emphasis) works best
- Sentence where the highlighted part is closer to the end benefit the most (see the ArXiv version of the paper)

chuklin@google.com