

Using Audio Transformations to Improve Comprehension in Voice Question Answering

Aleksandr Chuklin¹ Aliaksei Severyn¹ Johanne R. Trippas² Enrique Alfonseca¹ Hanna Silen¹ Damiano Spina²

¹Google Research (Zürich, Switzerland & London, UK)

²RMIT 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

Finding Answers – Eyes-Free Voice Assistant

Instructions

An eyes-free voice assistant is an electronic device that can understand voice queries from a user and provide audio responses to that query. Eyes-free voice assistants that you may know include Amazon Alexa or Google Home

In this task, you will first read a query and then you will listen to a corresponding answer in an audio file from the eyes-free voice assistant. Your job is to provide feedback about:

1. the answer to a query in audio,
2. the quality of the audio.

User's query: [Which guitarist inspired Queen?]

Eyes-free voice assistant's response:



[1a] Does the audio response from the eyes-free voice assistant answer the user's query? (required)

No, it doesn't Slightly Moderately Highly Yes, fully

[1b] Type the short answer which you hear in the audio response as you hear it (or n/a if not available) (required)

- Keep the answer as short as possible
- Use lowercase letters

[2a] Is the length of the full audio response appropriate? (required)

Too short OK Too long

[2b] Were the words in the full audio response pronounced correctly with no awkward speech artefacts? (required)

Bad Moderate Good

[2c] Are there any unexpected interruptions in the full audio response? (required)

No Yes

[2d] What would you change to the audio to better identify the answer.

Tell us how you would like to hear the response from the eyes-free voice assistant.

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* (↓).

	<i>inform.</i> ↑	<i>correctness</i> ↑	<i>elocution</i> ↑	<i>interruption</i> ↓	<i>length</i> ↓
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

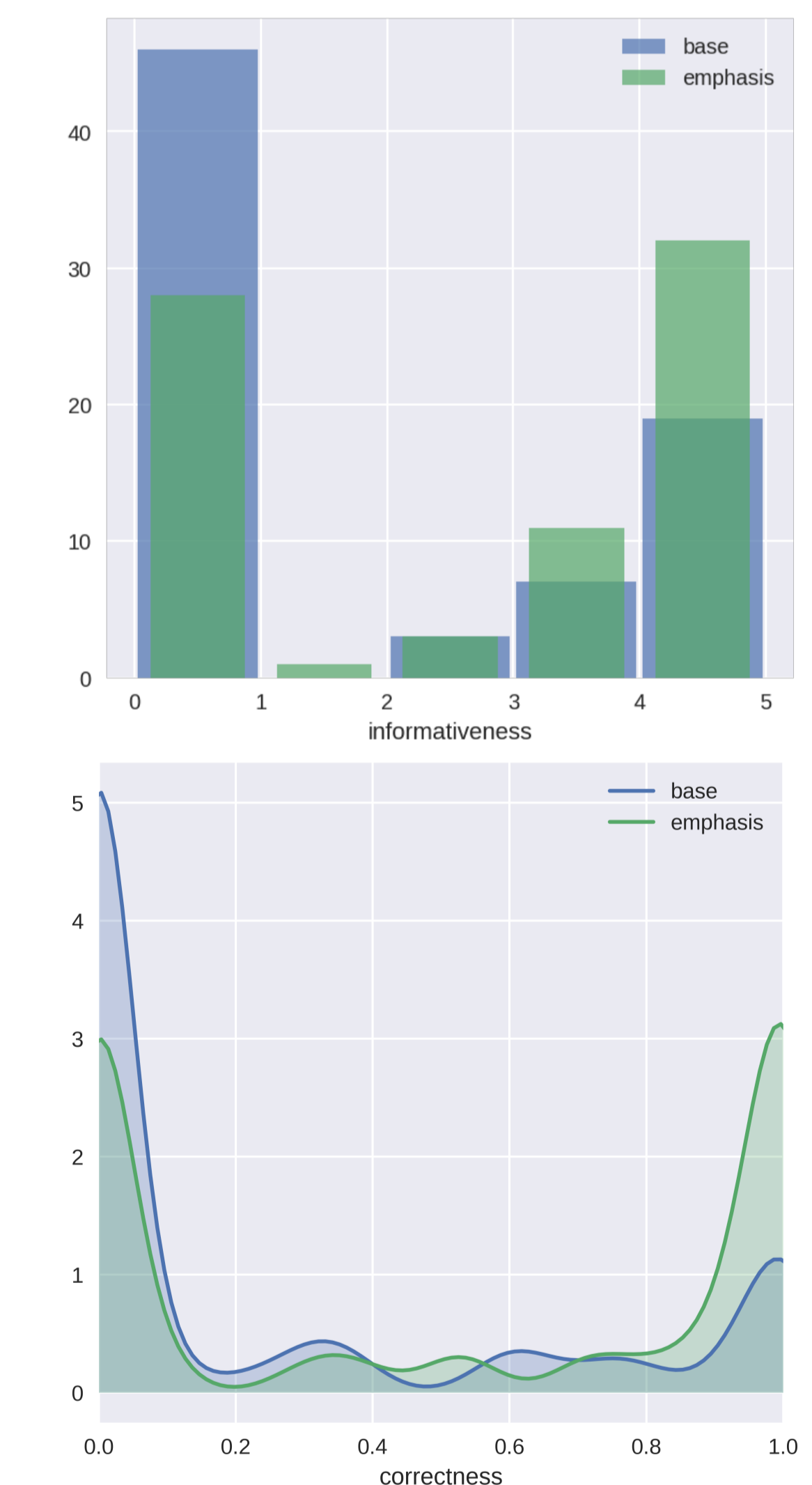


Figure: Distributions of the *informativeness* (left) and *correctness* (right) of the audio with emphasis (red narrow bars) vs. the baseline (blue wider bars). Higher scores are better.

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



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)