Using Audio Transformations to Improve Comprehension in Voice Question Answering

Aleksandr Chuklin\textsuperscript{1} \quad Aliaksei Severyn\textsuperscript{1} \quad Johanne R. Trippas\textsuperscript{2} \quad Enrique Alfonseca\textsuperscript{1} \quad Hanna Silen\textsuperscript{1} \quad Damiano Spina\textsuperscript{2}

\textsuperscript{1}Google Research (Zürich, Switzerland & London, UK) \quad \textsuperscript{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

\textlt<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 \textlt<break>SSML tag, rate / pitch parameters of <prosody>, and level parameter of <emphasis>.

\begin{tabular}{|c|c|c|c|c|c|}
\hline
TTS engine & Voice & pause & rate & pitch & emphasis \\
\hline
IBM & Lisa & strong & x-slow & x-high & n/a \\
Google & Wavenet-F & strong slow & +2st & strong & \\
\hline
\end{tabular}

Reson娘 Setup

Differences 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 (↓).

\begin{tabular}{|c|c|c|c|c|c|}
\hline
TTS engine & pause & rate & pitch & emphasis & result \\
\hline
IBM & -0.21 & +0.04 & -0.03 & +0.37 & +0.08 \\
Google & +0.21 & +0.09 & -0.04 & +0.15 & +0.00 \\
\hline
\end{tabular}

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)