Abstract

Pause-internal phonetic particles (PINTs) comprise a variety of phenomena including: phonetic-acoustic silence, inhalation and exhalation breath noises, filler particles "uh" and "um" in English, tongue clicks, and many others. These particles are omni-present in spontaneous speech, however, they are under-researched in both natural speech and synthetic speech. The present work explores the influence of PINTs in small-context recall experiments, develops a bespoke speech synthesis system that incorporates the PINTs pattern of a single speaker, and evaluates the influence of PINTs on recall for larger material lengths, namely university lectures.

The benefit of PINTs on recall has been documented in natural speech in small-context laboratory settings, however, this area of research has been under-explored for synthetic speech. We devised two experiments to evaluate if PINTs have the same recall benefit for synthetic material that is found with natural material. In the first experiment, we evaluated the recollection of consecutive missing digits for a randomized 7-digit number. Results indicated that an inserted silence improved recall accuracy for digits immediately following. In the second experiment, we evaluated sentence recollection. Results indicated that sentences preceded by an inhalation breath noise were better recalled than those with no inhalation. Together, these results reveal that in single-sentence laboratory settings PINTs can improve recall for synthesized speech.

The speech synthesis systems used in the small-context recall experiments did not provide much freedom in terms of controlling PINT type or location. Therefore, we endeavoured to develop bespoke speech synthesis systems. Two neural text-to-speech (TTS) systems were created: one that used PINTs annotation labels in the training data, and another that did not include any PINTs labeling in the training material. The first system allowed fine-tuned control for inserting PINTs material into the rendered material. The second system produced PINTs probabilistally. To the best of our knowledge, these are the first TTS systems to render tongue clicks.

Equipped with greater control of synthesized PINTs, we returned to evaluating the recall benefit of PINTs. This time we evaluated the influence of PINTs on the recollection of key information in lectures, an ecologically valid task that focused on larger material lengths. Results indicated that key information that followed PINTs material was less likely to be recalled. We were unable to replicate the the benefits of PINTs found in the small-context laboratory settings.

This body of work showcases that PINTs improve recall for TTS in small-context environments just like previous work had indicated for natural speech. Additionally, we've provided a technological contribution via a neural TTS system that exerts finer control over PINT type and placement. Lastly, we've shown the importance of using material rendered by speech synthesis systems in perceptual studies.