Scene viewing and gaze analysis during phonetic segmentation tasks

Arif Khan[1–3], Ingmar Steiner[1,2], Ross Macdonald[1], Yusuke Sugano[1,4], and Andreas Bulling[1,4]

[1]Multimodal Computing and Interaction, Saarland University, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[3]Saarbrücken Graduate School of Computer Science, Germany
[4]Max Planck Institute for Informatics, Saarbrücken, Germany

Research question Phonetic segmentation is the process of splitting speech into individual sounds. Human experts perform this task manually by analyzing auditory and visual cues using analysis software, but one minute of speech can take an hour to segment.

To improve automatic segmentation, which cannot yet match human experts' accuracy, we analyzed the behavior of experts performing segmentation tasks, using a stationary eye-tracker.

Methods A 46 s recording of "The Northwind and the Sun" was segmented using standard phonetic software (Figure 1).[1] Gaze activity was captured using a Tobii TX300. The computer screen and user interaction were recorded as well. Data collection is ongoing, and we plan to record 12 experts.

Results During the task, experts zoom in to view short spans of audio; we analyzed the scene viewing behavior (fixation locations and durations), as well as the audio segments to which they listen. Moreover, activity over the entire task was analyzed within and across participants (Figure 2).

Interpretation Preliminary results provide new insight about experts' behavior during phonetic segmentation tasks. Identifying critical features of visible speech in this manner will allow us to model their importance for automatic segmentation. It also exposes behavioral differences across individual experts performing the same task.

[1] *Praat*, `http://www.praat.org/`

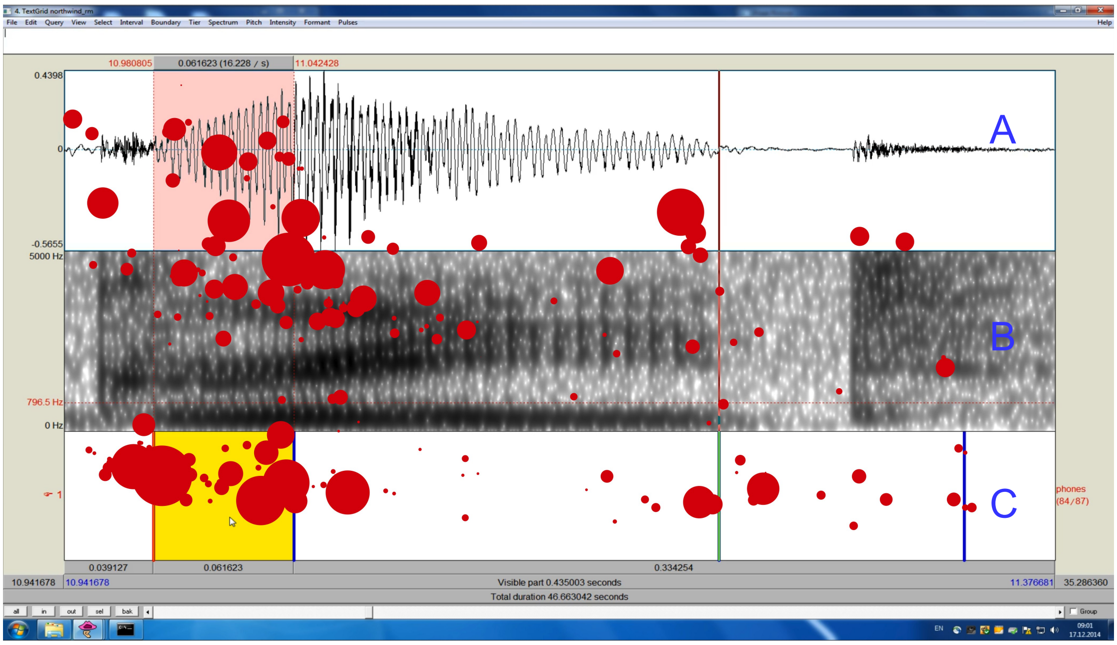

Figure 1: A screenshot showing one scene viewed by one subject, with gaze fixation events rendered as red circles. The phonetic software shows three panels (labeled **A**, **B**, and **C** in blue, at right); these are synchronized and in this figure display the span from 10.94 to 11.38 s in the speech data. The top panel (A) shows the audio waveform, the middle panel (B), a spectrogram (configured with a frequency range of 0 to 5 kHz), and the bottom panel (C) contains the speech unit boundaries manually placed by the experts in the phonetic segmentation task. Arbitrary selections of audio (pink in panel A, yellow in panel C) can be played back as desired.

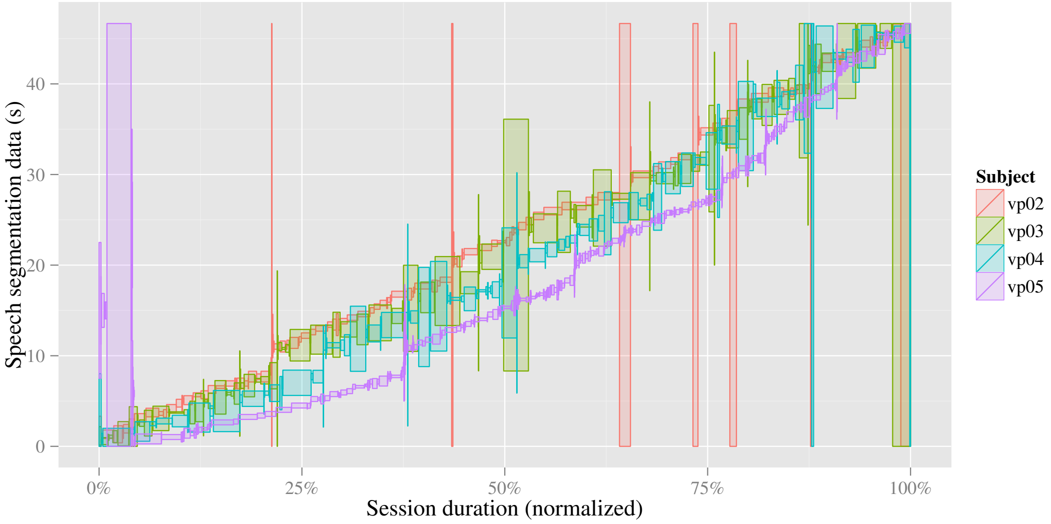

Figure 2: Speech segmentation data spans which were viewed as scenes over the (normalized) duration of the segmentation task. Each rectangle represents the portion of time (*width* of rectangle) spent segmenting a span of recorded speech data, where the rectangle *height* represents the duration of that span.

Individual differences between expert subjects are visible, e.g., subject vp05 zoomed in much further, viewing a significantly larger number of shorter spans of the speech data during segmentation than the other subjects. Certain spans of the recording consistently took more time to analyze, while pauses in the recorded speech (e.g., at 10 s and 20 s) were skipped by all subjects.