Scene Viewing and Gaze Analysis during Phonetic Segmentation Tasks

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

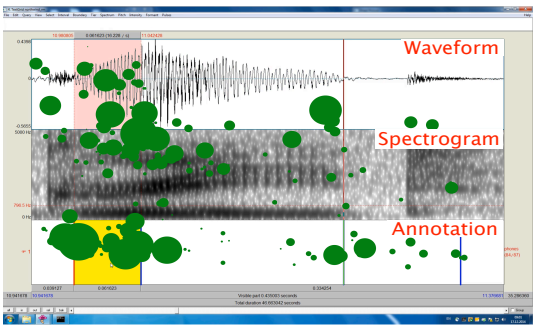
[1]*Cluster of Excellence MMCI, Saarland University*
[2]*DFKI Language Technology Lab, Saarbrücken*
[3]*Saarbrücken Graduate School of Computer Science*
[4]*Max Planck Institute for Informatics, Saarbrücken*

1. Introduction

Phonetic segmentation is the process of splitting speech into distinct sounds. Human experts perform this task manually by analyzing auditory and visual cues using analysis software, which takes a large amount of time. Methods exist for automatic segmentation (AS) but are not accurate enough for certain applications. For improving AS, we analyzed the behavior of experts performing segmentation tasks, in an effort to incorporate the same knowledge in AS.

2. Methodology

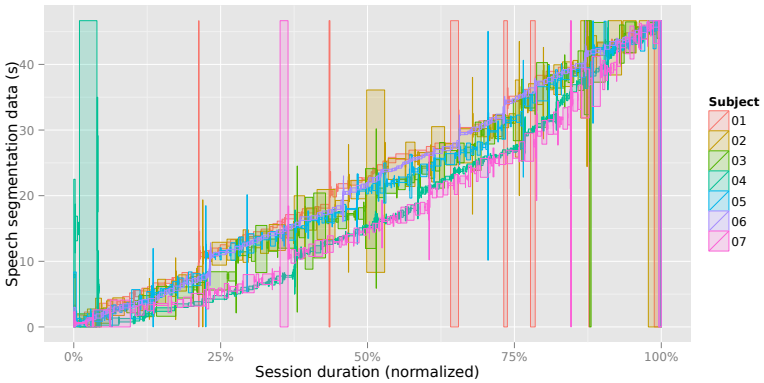
- Study participants segmented 46 seconds of recorded speech using the standard phonetic software, Praat (`http://praat.org`).
- During segmentation, gaze activity was captured using a Tobii TX300 eye tracker.
- The computer screen, user input (mouse and keyboard), and software output (audio) were recorded.
- The average number of fixations, fixation locations, and the task activity were analyzed within and across seven participants.



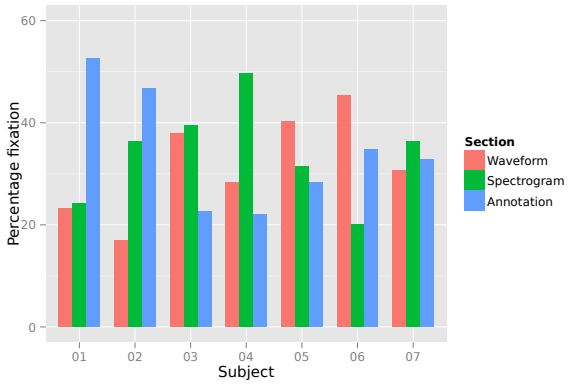
A screenshot of a scene viewed by one subject, with fixations rendered as green circles. The diameter of the circle represents the fixation duration.

3. Results

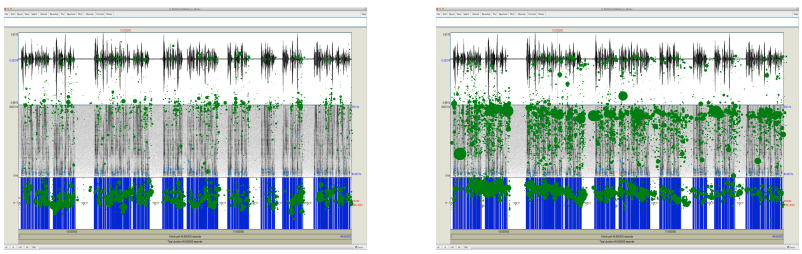
- For analysis, the screen is divided vertically into three sections, viz. waveform, spectrogram, and annotation.
- Data show differences in behavior across participants performing the same task.
- Task completion time varies from 45 min to 99 min across participants.
- Total number of scenes selected for segmentation also differed for all participants.



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



Average fixation for each participant in the three sections.



Fixations transformed into the time domain of the speech recording are shown in green for subject01 (left) and subject07 (right). They fixate on different regions in the screen for the same segmentation task.

4. Conclusion

- We analyzed the gaze movement behavior of phoneticians during speech segmentation.
- We have seen that humans use several modalities to segment speech, the focus on a specific modality is user specific.
- In the future, we plan to collect more data and model the phonetician behavior to improve AS.

5. Acknowledgments

We are extremely grateful to all the participants who gave their time for the segmentation task and who provided feedback.

German Research Center for Artificial Intelligence
CLUSTER OF EXCELLENCE
max planck institut informatik
UNIVERSITÄT DES SAARLANDES