A Multimodal Corpus of Expert Gaze and Behavior during Phonetic Segmentation Tasks

Arif Khan[1–3], Ingmar Steiner[1,2], Yusuke Sugano[4], Andreas Bulling[1,5], Ross Macdonald[6]
[1]Multimodal Computing and Interaction, Saarland University, Germany,
[2]German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany,
[3]Saarbrücken Graduate School of Computer Science, Germany,
[4]Osaka University, Japan
[5]Max Planck Institute for Informatics, Saarbrücken, Germany
[6]University of Manchester, UK,
{arifkhan,steiner}@coli.uni-saarland.de

Abstract

Phonetic segmentation is the process of splitting speech into distinct phonetic units. Human experts routinely perform this task manually by analyzing auditory and visual cues using analysis software, which is an extremely time-consuming process. Methods exist for automatic segmentation, but these are not always accurate enough. In order to improve automatic segmentation, we need to model it as close to the manual segmentation as possible. This corpus is an effort to capture the human segmentation behavior by recording experts performing a segmentation task. We believe that this data will enable us to highlight the important aspects of manual segmentation, which can be used in automatic segmentation to improve its accuracy.

Keywords: eyetracking, gaze analysis, manual segmentation behavior

1. Introduction

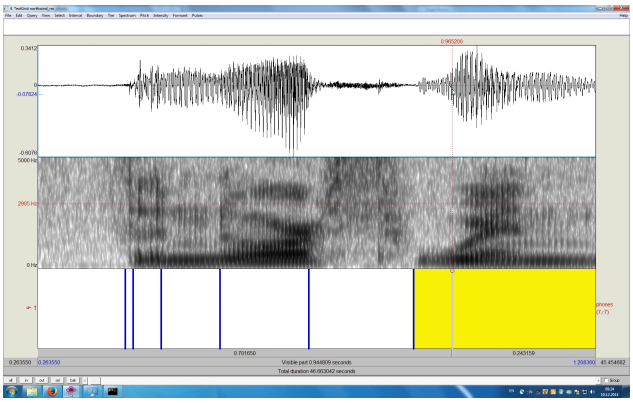
Speech segmentation is the process of splitting the acoustic speech signal into distinct units by placing timestamped boundaries. This forms a crucial data processing step for phonetic analysis, as well as speech technology applications such as text-to-speech synthesis and automatic speech recognition. The results and output quality depend on accurately segmented speech data.

Speech segmentation can be done manually, using specialized software, e.g., *Praat* (Boersma, 2001), *Wavesurfer* (Beskow and Sjölander, 2000), *ELAN* (Sloetjes and Wittenburg, 2008), and *EMU* (Winkelmann et al., 2017). In this workflow, a speech recording is displayed as a waveform and/or spectrogram, and boundaries are inserted using the mouse or keyboard (cf. Figure 1a). Short audio segments can be played back to validate the boundary placement. This process is repeated until the whole audio file is segmented. Manual segmentation by experts is considered to produce the best phonetic segmentation one can achieve for any given data (Svendsen and Soong, 1987; Wesenick and Kipp, 1996). One reason for this is that they combine experience with multiple sources of information. However, there are some critical drawbacks of manual segmentation which make it impractical for large speech corpora. The first is that it is very laborious and time consuming; on average, manual segmentation can take up to 30 s per phone (Leung and Zue, 1984; Stolcke et al., 2014) to segment. As a result, newly recorded speech data cannot be used quickly if manual segmentation is desired. Secondly, the exact placement of boundaries is subjective, and there may be disagreement between multiple experts.

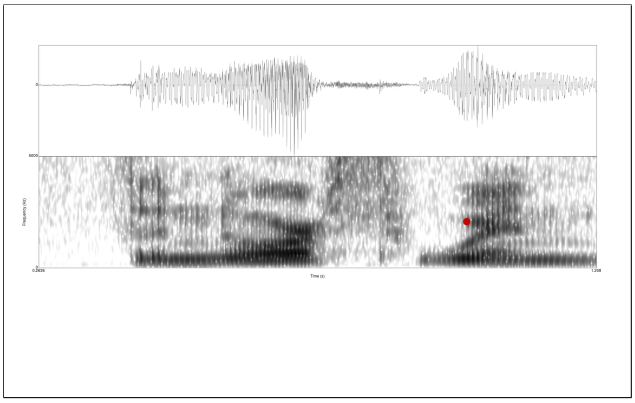
The second method of segmentation is doing it automatically, by training a model on the audio data, and then using it to segment speech. In this method, the accuracy of the segmented speech directly depends on the quality of the trained model, which itself depends on the quality of training data. Previous studies have used different approaches for automatic segmentation. For a long time, researchers have used hidden Markov models (HMMs) for automatic segmentation (Rabiner, 1989; Juang and Rabiner, 1991; Toledano et al., 2003; Brognaux and Drugman, 2016). Others have used neural networks for automatic segmentation (Karjalainen et al., 1998; Schwarz et al., 2006). One commonality of these approaches is the use of only audio as input features for training the model. The audio is processed to extract acoustic features, which are then used for training. Several techniques are available for extracting acoustic features from speech, but the most commonly used are mel-frequency cepstral coefficients (MFCCs) (Logan, 2000) and Perceptual Linear Prediction (PLP) (Hermansky, 1990). While the use of only audio as acoustic features produces acceptable results for most segmentation requirements, humans use more than audio for segmenting speech. To improve automatic segmentation, we therefore want to add more modalities to model it as closely as possible to the manual segmentation. We hope that modeling automatic segmentation in this way will produce better results.

To this end, we first need to analyze the human segmentation behavior and highlight the key information sources that humans experts use to segment speech. Our data includes gaze information, which shows where the experts look on the screen, the audio to which they listen during segmentation, video from a webcam attached to the monitor, and a screen recording of what they are viewing. To the best of our knowledge, this is the first corpus that records the human segmentation in such a setup.

The rest of this paper is organized as follows. Section 2 provides details of how the data was recorded, along with the format and structure. In Section 3, we present the results of some preliminary analysis conducted on the data. Finally,



(a) A screenshot of a sound recording and annotation in Praat. The GUI is split into three sections: waveform (top), spectrogram (middle), and annotation (bottom).



(b) Corresponding scene data reconstructed from recorded audio using Praat log and gaze information. Here, the subject is looking at a formant in the spectrogram; the fixation is rendered in red.

Figure 1: One frame from the screen capture video (left) and the corresponding reconstruction (right).

the conclusion and future use of the data is outlined in Section 4.

2. The Corpus

In order to study the behavior of human experts during speech segmentation tasks, we designed and recorded the multimodal corpus described in this section.

2.1. Preparation

We recorded a native speaker of Scottish English, reading aloud the standard passage, "The North Wind and the Sun" (International Phonetic Association, 1999). The recording was made in a sound-attenuated booth, with a close-talking microphone, sampling at 48 kHz with 24 bit quantization. The resulting file has a duration of 46 s.

2.2. Data Collection

We recorded seven subjects, with the instruction that they were to segment (but not label) the recording into phones using the Praat graphical user interface (GUI). All of the subjects who participated in the data collection are trained phoneticians with varying amounts of experience; details are given in Table 1.

The participants took different amounts of time (44 to 96 min) to complete the task. The normalized session duration for all the subjects in shown in Figure 2. We did not control the speed in which the participants completed the task, so each took time according to his or her preference, which resulted in different session durations.

We used a Tobii TX300 eyetracker,[1] to record the gaze movements and capture where the subject looked on the computer screen during the entire session, at a sampling rate of 120 Hz. For each subject, before the beginning of recording, we first calibrated the eyetracker. The calibration is done to adjust the height of head and seating position, which is different for each subject. Using the *TobiiStudio* software (v3.2.3), we also recorded the screen content itself (at a resolution of 1920×1200 pixels), as well as any audio the subjects played back from the recording during the segmentation task. In addition to the gaze information and screen recording, TobiiStudio also allowed us to log any keystrokes and mouse clicks during the recording session, as well as the video from a webcam facing the subject, at a resolution of 640×480 pixels. The screen capture and webcam were intended to validate the subjects' head movements and input device logging.

In addition to these modalities, we polled the application state of the Praat GUI, once per second, in order to log the zoom level of the audio recording shown and other application-specific data. Finally, the segmentation itself, produced by each subject over the course of the session, was saved in Praat's widely supported *TextGrid* annotation file format.

2.3. Data Processing

After each recording session, the logs from TobiiStudio and Praat were exported to ASCII text files and compressed. The screen recordings and webcam videos, as well as the audio playback recordings, were exported from TobiiStudio in ASF containers, in TechSmith Screen Capture Codec (TSCC), Microsoft Video 1, and MP3 format, respectively, the latter at 22 kHz and 16 bit quantization, at a bitrate of 128 kbit/s.

In order to manipulate the multimedia streams from each recording session more efficiently, we first converted the video to H.264 format (which allowed more robust seeking and reduced the file sizes – from 52 GB to 3 GB without noticeable loss in quality), transcoded the audio to FLAC format,[2] and multiplexed all three streams into a single Matroska video container,[3] using FFmpeg.[4]

Next, we parsed the Praat logs to identify time segments in each recording session during which the subject was viewing the same zoom level and interval of the audio recording; doing this allowed us to treat them as quasi-static *scenes* viewed by the subject. The session times as well as the audio recording times of each scene were collected into a

[1] https://www.tobiipro.com/product-listing/tobii-pro-tx300

[2] https://xiph.org/flac/

[3] https://matroska.org/

[4] https://ffmpeg.org/

Subject	Gender	Age (years)	Native Language	Experience (years)	Segmentation Time (min)
01	F	26	German	7	44
02	M	47	German	20	55
03	M	37	German	15	73
04	F	35	Polish	10	96
05	F	27	German	4	71
06	F	22	German	1.5	80
07	F	22	German	4	92

Table 1: Age, gender, native language, and segmentation experience of the subjects who participated in the data collection.

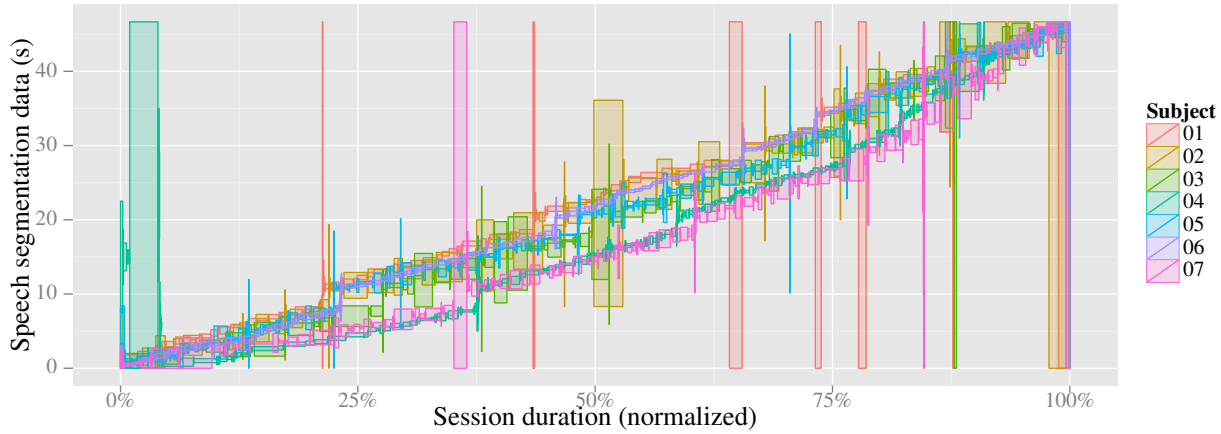


Figure 2: Speech segmentation data spans which were viewed as scenes over the (normalized) duration of the segmentation task. Each rectangle represents the portion of time (rectangle width) spent segmenting a span of recorded speech, while the rectangle height represents the duration of that span.

YAML file.[5]

After determining the constant time offset between the Praat and TobiiStudio logs, we could then select the gaze data related to each scene and store it in a structured format, validating it via the screen recording. The data is structured by scene and also includes the duration and location (absolute and classified by GUI region) of each fixation. Based on this information, we reconstructed the relevant information in each scene and synthesized it into a second video stream with the gaze location rendered as a red circle (cf. Figure 1b). We also extracted the signal time codes of each scene and added them as a subtitle track. The resulting YAML files and multimedia streams were finally packaged and provided as a data dependency for analysis.

3. Analysis

Our initial analysis concerns the eyetracking data. The main purpose of the corpus was to allow us to analyze the manual segmentation behavior and to identify modalities and features useful for modeling segmentation. For the analysis of the eyetracking data, it is important to understand the concepts of *fixation* and *saccades*. If the $\langle x, y \rangle$ location of the gaze on the screen does not change significantly within some time frame, then those gaze events are classified as a fixation. The movement of gaze between two fixations is referred to as a saccade. The actual time duration for which the $\langle x, y \rangle$ location movement should remain constant is subjective and device dependent. We used the default settings of Tobii to identify the fixations and saccades, the details of which are described by Ollson (2007). For analysis, the Praat GUI on the screen is divided horizontally into three sections, each representing a different portion of the screen. We refer to these sections as *waveform*, *spectrogram*, and *annotation*, as shown in Figure 1. The waveform represents the oscillogram of the audio recording in Praat. The spectrogram represents the time-frequency-energy representation of the signal; the x axis represents time, and the y axis, the frequency of the signal, while the grayscale value indicates the energy in each time-frequency bin. The annotation section is used by the subjects to place the boundaries. This is the only section which can be edited by the user for creating and manipulating time-aligned annotations (boundaries and labels).

3.1. Scenes

Further to the progress visualization in Figure 2, Table 2 summarizes the number of scenes the subjects viewed over the course of their session. As can be seen, subjects 01 to 03 and 05 to 07 used almost the same number of scenes for segmentation. Subject 04 viewed a larger number of scenes with the second lowest average scene length, indicating that this participant preferred to "zoom in" more than the others.

3.2. Fixations

One of the most important questions is, where the subjects look on the screen during the manual segmentation task. To

[5] http://yaml.org/

Subject	Total scenes	Total duration (s)	Average duration (s)
01	157	519.79	3.31
02	157	593.63	3.78
03	150	562.23	3.74
04	522	627.67	1.20
05	308	671.64	2.18
06	352	361.50	1.02
07	276	652.49	2.36

Table 2: The total number of scenes, sum of scenes length and average scene length the subjects used for segmenting the audio recording.

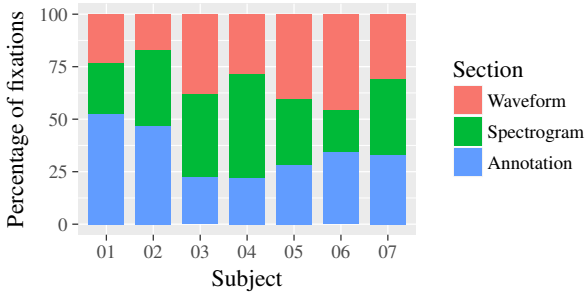


Figure 3: Average fixations for each subject in the three sections of the Praat GUI.

answer this question, we calculated the proportion of gaze events in the three sections of the screen. Figure 3 shows the percentage of fixations in each of the three screen sections for all subjects. The fixations in the *annotation* area can be disregarded, because in order to place the boundary, the subjects have to carefully "click" in the right location and during this process, a lot of gaze activity may occur in this section. The fixations in the *waveform* and *spectrogram* sections are important and have a mixed pattern. All subjects have a higher number of fixations in the *spectrogram* section than in the *waveform* section.[6]

4. Conclusion and Outlook

In this paper, we have presented a multimodal corpus of behavior data from expert phoneticians performing a manual speech segmentation task. All important information sources that are relevant to the segmentation task were recorded. This includes gaze, playback audio, video, and screen recording. The produced segmentation, as well as events logged from the keyboard, mouse, and Praat GUI are also provided. We believe that this data will prove valuable for research in observing and understanding manual segmentation.

This corpus can help identify critical information sources used by humans during manual segmentation, which can be modeled to improve the accuracy of automatic segmentation. In addition, this data can be useful in analyzing the interaction of phoneticians with speech segmentation software (Praat) and can be used to improve the usability of such a software. For example it might be possible to modify the way the boundaries are defined or to introduce a software feature to visualize the predicted complexity of speech regions while they are being segmented.

The processed data (cf. Section 2.3) has been released under a Creative Commons license (CC-BY-NC-SA) and published on GitHub,[7] along with the processing recipes. This public release *excludes* the webcam videos, in order to protect the privacy of our participants.

[6]The exception is subject 06; this may be because she had the least amount of segmentation experience (see Table 1) and relied more on the waveform section to segment.

5. Acknowledgements

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

This study was funded by the German Research Foundation (DFG) under grant number EXC 284.

6. Bibliography

Beskow, J. and Sjölander, K. (2000). WaveSurfer: An open source speech tool. In *International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Brognaux, S. and Drugman, T. (2016). HMM-based speech segmentation: Improvements of fully automatic approaches. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(1):5–15.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272.

Karjalainen, M., Altosaar, T., and Huttunen, M. (1998). An efficient labeling tool for the QuickSig speech database. In *International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia.

Leung, H. and Zue, V. (1984). A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 9, pages 73–76.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval (ISMIR)*, Plymouth, MA, USA.

Ollson, P. (2007). Real-time and offline filters for eye tracking. Master's thesis, KTH.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[7]https://github.com/m2ci-msp/eyetracking-data

Schwarz, P., Matějka, P., and Černocký, J. (2006). Hierarchical structures of neural networks for phoneme recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume I, pages 325–328, Toulouse, France.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Stolcke, A., Ryant, N., Mitra, V., Yuan, J., Wang, W., and Liberman, M. (2014). Highly accurate phonetic segmentation using boundary correction models and system fusion. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5552–5556, Florence, Italy.

Svendsen, T. and Soong, F. K. (1987). On the automatic segmentation of speech signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 77–80, Dallas, TX, USA.

Toledano, D. T., Gómez, L. A. H., and Grande, L. V. (2003). Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625.

Wesenick, M.-B. and Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. In *International Conference on Spoken Language Processing (ICSLP)*, pages 129–132, Philadelphia, PA, USA.

Winkelmann, R., Harrington, J., and Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, 45:392–410.