GECKO - A Tool for Effective Annotation of Human Conversations

Golan Levy[1], *Raquel Sitman*[1], *Ido Amir*[1], *Eduard Goldstein*[1], *Ran Mochary*[1], *Eilon Reshef*[1], *Roi Reichardt*[1 2], *Omri Allouche*[1]

[1]Gong.io
[2]Faculty of Industrial Engineering and Management, Technion, Israel
{golan.levy|rocky.sitman|ido.amir|eduard.goldstein|ran.mochary |eilon.reshef|roi.recihardt|omri.allouche}@gong.io

Abstract

With the dramatic improvement in automated speech recognition (ASR) accuracy, a variety of machine learning (ML) and natural language processing (NLP) algorithms are designed for human conversation data. Supervised machine learning and particularly deep neural networks (DNNs) require large annotated datasets in order to train high quality models. In this paper we describe Gecko, a tool for annotation of speech and language features of conversations. Gecko allows efficient and effective segmentation of the voice signal by speaker as well as annotation of the linguistic content of the conversation. A key feature of Gecko is the presentation of the output of automatic segmentation and transcription systems in an intuitive user interface for editing. Gecko allows annotation of Voice Activity Detection (VAD), Diarization, Speaker Identification and ASR outputs on a large scale. Both annotators and data scientists have reported improvement in the speed and accuracy of work. Gecko is publicly available for the benefit of the community at `https://github.com/gong-io/gecko`.

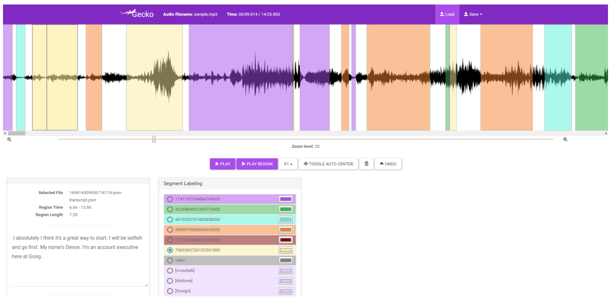
Index Terms:Annotation, Labeling, Speaker segmentation, Diarization, Speech recognition, VAD

1. Introduction

Automatic speech recognition (ASR) has dramatically improved in the last decade due to the excellent performance of deep neural networks (DNNs) [1, 2]. Consequently, it is now possible to develop ML and NLP algorithms that can effectively process human conversations. However, these algorithms, and particularly those based on DNNs, require large amounts of annotated data. To meet this need, in this paper we present Gecko, a tool designed to provide a convenient user interface for effective and efficient annotation of conversation speech signal, and discuss how it is used to annotate Voice Activity Detection (VAD), diarization, speaker identification and speech recognition. Such a tool, to the best of our knowledge, does not currently exist.

2. Tool Description

Gecko is a standalone web-based application written in Javascript. It runs in web browsers on both desktop and mobile devices, and does not require a server, making its deployment and update very simple. It provides an interactive interface for annotating audio files, with a special emphasis on conversations. The clean interface (Figure 1) integrates media player capabilities with an intuitive interface for interactive editing.



Selected File 1694149099956714174-json-transcript.json
Region Time 6.66 - 13.86
Region Length 7.20

I absolutely I think it's a great way to start. I will be selfish and go first. My name's Devon. I'm an account executive here at Gong.

Figure 1: *A simple view of one annotation file with an audio file. Notice that the speakers are color coded and the word heard in the audio is highlighted in the transcript.*

2.1. Input \Output

Gecko requires an audio file and one or more files with annotations of segments. Optionally, the text of each segment can be provided. Gecko supports annotation files in a variety of formats: RTTM, CTM, JSON, TSV. These input files are typically generated by leading Speech Recognition frameworks like Kaldi [3]. Multiple annotation files with segments and transcripts can be uploaded to Gecko, allowing comparison of multiple algorithms or evaluation of performance versus ground truth. The input annotation files can be edited and saved as RTTM/JSON/CTM files.

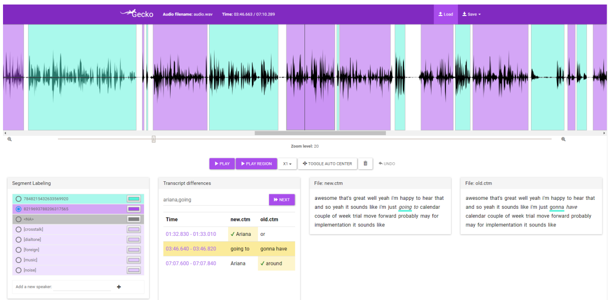
2.2. User Interface

The main view (see Figure 1) includes a waveform display, with speaker segments overlaid on it. The color of each segment indicates its label (often speaker identity, but this can be used for any label of interest). A table allows setting the label of segments, as well as the creation of new labels. If the transcript is available, it is also displayed and synced with the audio being played.

For segmentation tasks, such as VAD, diarization and speaker identification, the user can set the start and end times of each segment, delete a segment or create a new one. The user can further mark the speaker identity of each segment or annotate it as containing music, cross-talk or any other label.

2.3. Annotation

As noted above, multiple annotations can be uploaded to Gecko. This allows investigation of the changes between different models. For example, the diarization error rate (DER) is often used as a metric to evaluate diarization performance, although more careful evaluation requires a dive into specific recordings to rec-



Time	new.ctm	old.ctm
01:32.830 - 01:33.010	✓ Ariana	or
03:46.640 - 03:46.820	going to	gonna have
07:07.600 - 07:07.840	Ariana	✓ around

File: new.ctm

awesome that's great well yeah i'm happy to hear that and so yeah it sounds like i'm just *going to* calendar couple of week trial move forward probably may for implementation it sounds like

File: old.ctm

awesome that's great well yeah i'm happy to hear that and so yeah it sounds like i'm just gonna *have* calendar couple of week trial move forward probably may for implementation it sounds like

Figure 2: *Comparing two different system annotation files and solving the discrepancies.*

ognize the performance of the system in different scenarios. With Gecko, one can compare the ground truth to an output of the diarization system (or other systems), and quickly assess the difference between the automatically generated, the gold segments and the identification of the speakers. Gecko can also be used to compare results of multiple diarization algorithms, displaying them side-by-side.

While comparing the annotations of multiple files, users can also edit the input annotations in order to refine the ground truth or create a new one from scratch. Elements of the input files such as speaker segments in the audio files and words in the transcripts are editable. An ”undo” mechanism is also provided.

2.4. Transcript Evaluation

Given the high performance of modern ASR systems, it is often more efficient to create a manually labeled dataset by refining the results of an ASR system rather than manually transcribing it from scratch. During playback, Gecko highlights the current word, and allows transcript edits in order to improve transcription quality. When the transcript file contains a per-word confidence score, the confidence is visually reflected in the transcript by the color of the words.

2.5. Transcript Comparison

Gecko allows comparison of two alternative transcripts, presenting the differences between the two transcripts in a tabular format, indicating the places where insertions, deletions and substitutions are identified. For each such discrepancy, the user can listen to the corresponding audio segment while both transcripts are presented. The user can choose one of the presented alternatives or type in the correct word. When comparing two alternative transcripts, a report can also be generated highlighting the correct terms, allowing evaluation of the transcripts.

Figure 2 depicts the interface for transcript comparison, with a table for handling discrepancies. The user can also search for specific discrepancies or display only unresolved discrepancies using the search box. In this way irrelevant discrepancies (e.g., ”in” instead of ”on”) can be ignored and the number of discrepancies to resolve may be easily monitored.

This technique is ideal for correcting high quality yet imperfect transcripts of advanced ASR systems. Instead of listening to several hours of conversations attempting to detect errors, the user can compare the results of multiple ASR systems, assuming these are the places errors are most likely to happen, and manually fix those segments.

3. Conclusions

We have presented Gecko, a web-based tool that allows joint annotation of speech and language aspects of conversations. Gecko is also an effective tool for the analysis of the output of a speech recognition model when compared to a gold standard. Gecko has been used to improve performance of VAD, diarization, speaker identification and ASR systems on a large scale. Both annotators and data scientists have reported improvement in the speed and accuracy of their work. We have released Gecko under an open source license, hoping it can benefit the general community.

4. Acknowledgements

We wish to thank Zeev Rannon, Ruth Aloni-Lavi, Igal Grinish, Eran Marom, Inbal Horev, Yonathan Guttel, Vitaliy Rybnikov and Dana Shimoni for their helpful suggestions.

5. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal processing magazine*, vol. 29, 2012.

[2] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8599–8603.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.