Vocal activity detection and speaker diarization in speech databases: a feasibility study

Fredrik Karlsson[1, 2]
[1] *Department of clinical science, Umeå University, Sweden*
[2] *Humlab, Umeå university, Sweden*
fredrik.k.karlsson@umu.se

Abstract

The task of creating speech corpora for phonetic research is time-consuming and could be alleviated by automatic algorithms to provide draft indexing of speech acts. The present investigation assessed the feasibility of applying speech segmentation and speaker diarization models across a collection of recordings to produce a draft indexing that could be utilised by speech management systems to help the researcher to navigate a corpus. The results show that a readily available model for speech segmentation is very likely to contribute to the effectiveness of speech annotation workflows in phonetic research. Speaker diarization models may require specific training to manage consistent speaker separation across a speech corpus, and the evaluated model currently offers no clear advantage to the effectiveness of a speech corpus creation process.

Introduction

Research into how speech is organised would receive a substantial benefit from the collection of larger and more ecologically valid speech samples on which to build theories and models of speech production may be built. Phonetic analysis in most cases, however, requires a preprocessing step in which units of importance for the analysis are identified and linked in time with the digital speech recording. This indexing task precedes phonetic or orthographic transcription but may in itself be cumbersome enough to constitute a barrier for more extensive speech recording efforts due to the fact that human transcribers currently perform it. The task of indexing a speech corpus is further increasingly more resource-demanding with the complexity of the speech task and more difficulty to alleviate with automation in the recording stage; simple productions embedded into a carrier sentence may be efficiently segmented directly in software applications that allow scripted recording instructed by a visual prompt (e.g. Draxler & Jänsch, 2004). With more ecologically valid tasks, such as continuous speech, the start and end times of utterances and phrases need to be identified by a human, and the resource intensity of this task increases in direct relation to the ecological validity of the speech task. In multi-party conversation, where speech by different speakers may overlap, and utterances may be interrupted by other parties, the task of separating speech acts into segments with starts and end times may even go from merely being time-consuming to perform to becoming a complex task even for humans.

Developments in speech processing technologies suggest that the task of indexing speech corpora may be alleviated by providing automated draft indexing. The draft segmentation of the speech signal into utterances may then be used in a speech database management system as a base for navigating between candidate utterances in supportive speech database management systems, such as the Emu-SDMS (Winkelmann, Harrington, & Jänsch, 2017). The open-source effort `pyannote-audio` (Yin, Bredin, & Barras, 2018) offer a PyTorch (Paszke et al., 2019)-based python library and pretrained models for speech activity detection, overlap detection, speaker segmentation, and overlap aware speaker diarization (Bullock, Bredin, & Garcia-Perera, 2020). For phonetic speech research, the application of speech activity detection to identify portions of a long recording session in which speech is most likely to have occurred would offer a substantial benefit in the initial stage of corpus annotation work. In multi-party recordings, such as recordings of parent-child interactions, a process of automatic speaker diarization may likely substantially facilitate analysis. The human transcriber would be able to use these indices and the functionality of speech database management systems to navigate efficiently between only the speech produced by the speaker of interest (i.e. the child in a speech acquisition study or the parent in a child-directed speech study).

Set against the background of a desire to expand speech corpora efficiently, it appears advantageous to use automatic speech indexing techniques to bootstrap subsequent manual phonetic analysis procedures. However, two points need to be considered before fully adopting these techniques. The indexing provided by the automatic system needs to be accurate enough so that the time taken to run the analysis and adjust the resulting draft indexing is substantially smaller than the time it would take to perform the indexing manually. The outcome of an analysis procedure may not be expected to capture discipline-specific definitions of what should be considered an utterance or a phrase without substantial parameter tuning of segmentation models. Instead, what comes out of an automatic segmentation should agree with some definition of the top-level unit of speech used in a manual indexing process.

In larger speech databases, the management of identified speakers constitutes separate consideration. Models performing speaker diarization cannot be assumed to be designed to handle conversations between hundreds of speakers and label these correctly, but rather that they are directed towards separating a handful of speakers in a recording. This potential limitation requires some consideration when a speaker diarization procedure is applied in a speech corpus context. As the output when processing one file, the procedure will identify a specific speaker as, for example, SPEAKER_01. Across recording sessions, however, the label SPEAKER_01 will likely refer to different speakers, and it may also be that SPEAKER_01 in one recording refers to the same speaker as SPEAKER_03 in another recording. Suppose consistent identification of an individual across the database is a desired feature, which is likely the case in most corpora. In that case, the speaker diarization needs to identify speakers consistently across the database.

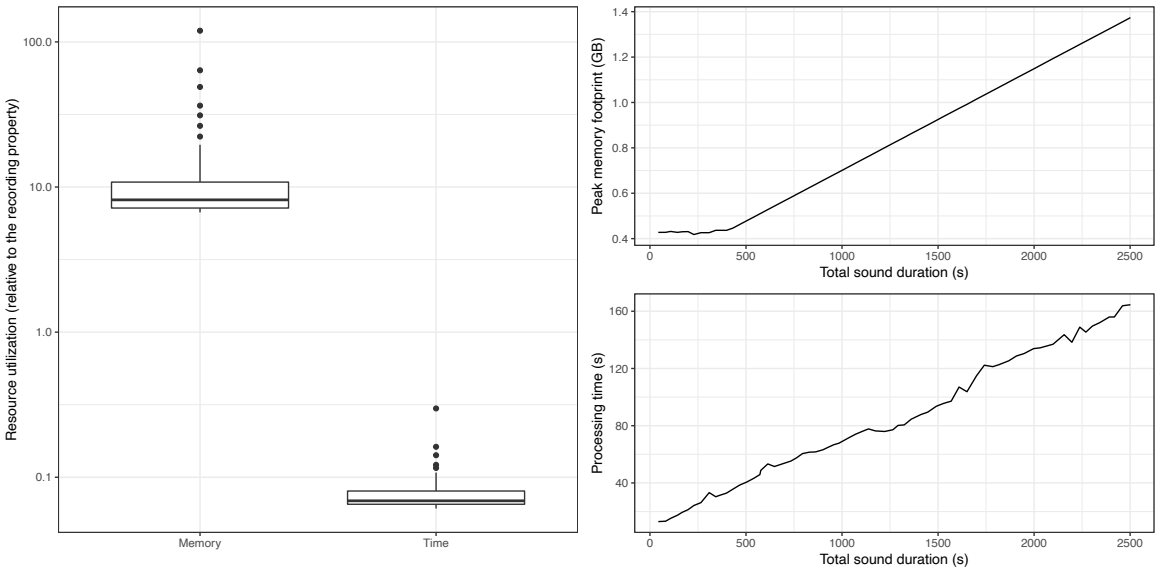

Figure 1. Peak memory consumption (top right) and time taken (bottom right) to identify portions of a recording with speech in recordings of increasing durations. The processing resources in terms of time and computer memory required for speaker segmentation per respective property in the recording is also indicated (left).

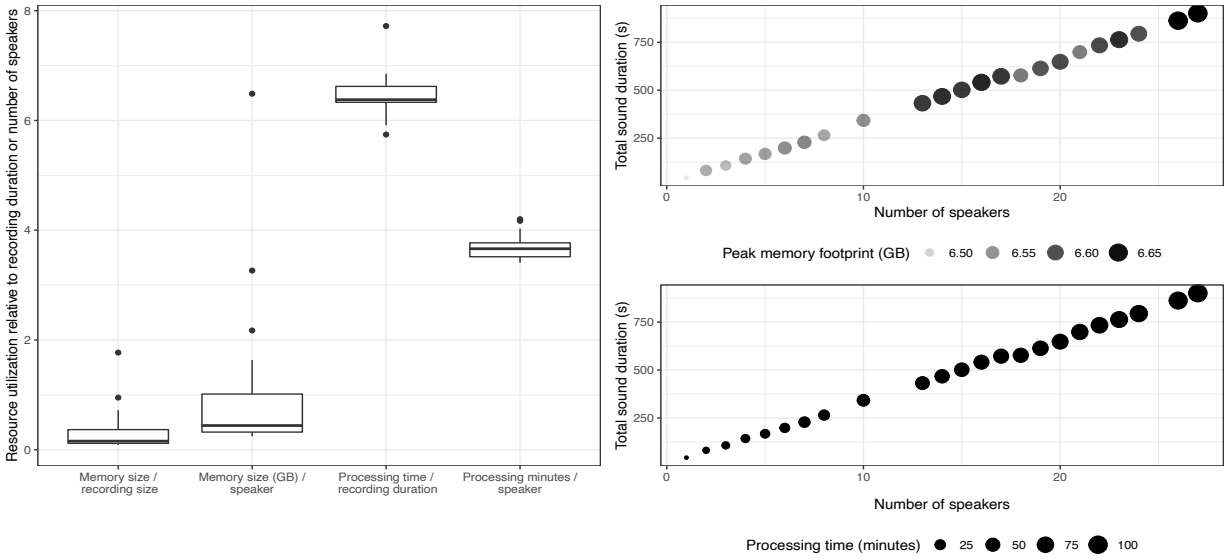

Figure 2. Peak memory consumption (top right) and time taken (bottom right) to identify speech and speakers in recordings of an increasing number of speakers and with increasing duration. The processing resources (time and memory) required for speaker diarization per speaker in the recording or minute recording is also indicated (left).

The easiest way to achieve a consistent specification of speakers is to apply the procedure the entire database simultaneously. However, speaker diarization is a resource-intensive task, and the task of keeping all speakers in even a smaller database may be too cumbersome to manage in implementations not purposely developed for this task. In this study I explored the utility of speech segmentation and speaker diarization models for producing draft indexing of multi-speaker corpora.

Materials and methods

A test corpus with varying speech recording lengths was simulated by iteratively concatenating recordings of speakers reading a standard text. Speaker diarization with consistent identification of individuals across a corpus was assumed to be expected by a researcher and enforced in the simulation by concatenating recordings of an increasing number of speakers together into a single file before diarization. The pre-trained models for segmentation (Bredin & Laurent, 2021) and diarization (Bredin et al., 2019) provided by the `pyannote-audio` community was consequently applied to speech recordings of 44s to 41 minutes of speech. The read speech passaged had been produced by 1 to 67 (44 male and 23 female) healthy adult speakers who had agreed to be recorded to assess their speech using digital signal processing methods in an ethically approved project (Regional Ethical Review Boards of Umeå, Case number 2012-368-31M). All sound files were recorded in a sound-treated recording booth and were sampled in the original wav format and digitised with a 16bit quantisation and either a 44.1kHz sampling frequency or a 48 kHz sampling frequency and then downsampled to 44.1 kHz using Praat (Boersma & Weenink, 2021).

The efficacy of speech segmentation and speaker diarization models to alleviate analysis work effort were assessed qualitatively (speech segmentation) and quantitatively (segmentation and diarization). The qualitative assessment of speech segmentation was based on the agreement of identified segments with any valid definition of an overarching unit of analysis used in speech research (e.g. an utterance, a sentence, or an intonational phrase). Further, whether portions of speech would be missed if human transcribers were to use the automatic segmentation as markers of regions of potential interest when transcribing was considered. The qualitative assessment was performed within the largest sound file, in which the recordings of all 67 speakers were included. The qualitative assessment of speaker diarization was performed holistically in terms of how well the diarization procedure recognised that the recordings were actually built up of utterances produced by singleton speakers concatenated together into one file to assign speakers' portions of speech consistently to the same speaker.

The quantitative assessments of the feasibility of applying segmentation and diarization procedures in a multi-speaker speech corpora context were assessed by the processing and total memory required to process recordings of increasing length and with an increasing number of speakers in them. The processing time and peak memory utilisation was quantified by the "`peak memory footprint`" and "`real`" time used for the computation as reported by the OSX Monterey `/usr/bin/time -l -hp` command. All computations were performed on a 2 GHz quad-core i5 MacBook Pro with 16GB RAM.

Results

Speech segmentation

The resource utilisation (time and memory) used while segmenting speech recordings of increasing lengths is indicated in Figure 1. For the small sample of recordings assessed here, the peak memory consumption is increased approximately linearly with 440 kilobytes per second of recording or 6.8 megabytes per megabyte in the sound file. The time taken complete the segmentation process was, on average, (with standard deviation) 8±3% (6-30% range) of the duration of the recording. The qualitative evaluation of the resulting segmentation showed that utterances made in the readings of the standard text were in all cases marked as speech, and sections marked as non-speech were never incorrectly marked as speech. The positioning of speech segment boundaries was generally correct and narrow to the actual speech task. However, instances of a delayed marking of the end of an utterance was observed where the speaker had an audible exhale or breathing pattern at the end of an utterance. In the read speech material assessed here, a human-perceived end of utterances was sometimes missed, and adjacent utterances were marked as one by the segmentation model, if read in a more fluent manner than what is usual.

Speaker diarization

The processing time and system memory required for speaker diarization is summarised in Figure 2. The increased resource (time and memory) requirements with increasing duration and with an increasing number of speakers are presented separately. On average (with standard deviation), diarization took 6.5±0.4 times the duration of the recording (range 5.7-7.7) and 3.7±0.23 (4.3-4.2) minutes per speaker in the recording to perform. System memory use remained stable within the 6.5-6.7 gigabyte range across recordings of different durations and with a different number of speakers in the sounds file. Figure 3 presents an overview of wherein a particular file specific speakers were identified by the diarization model in sound files with 2, 5, 15 and 25 induced speakers, respectively.

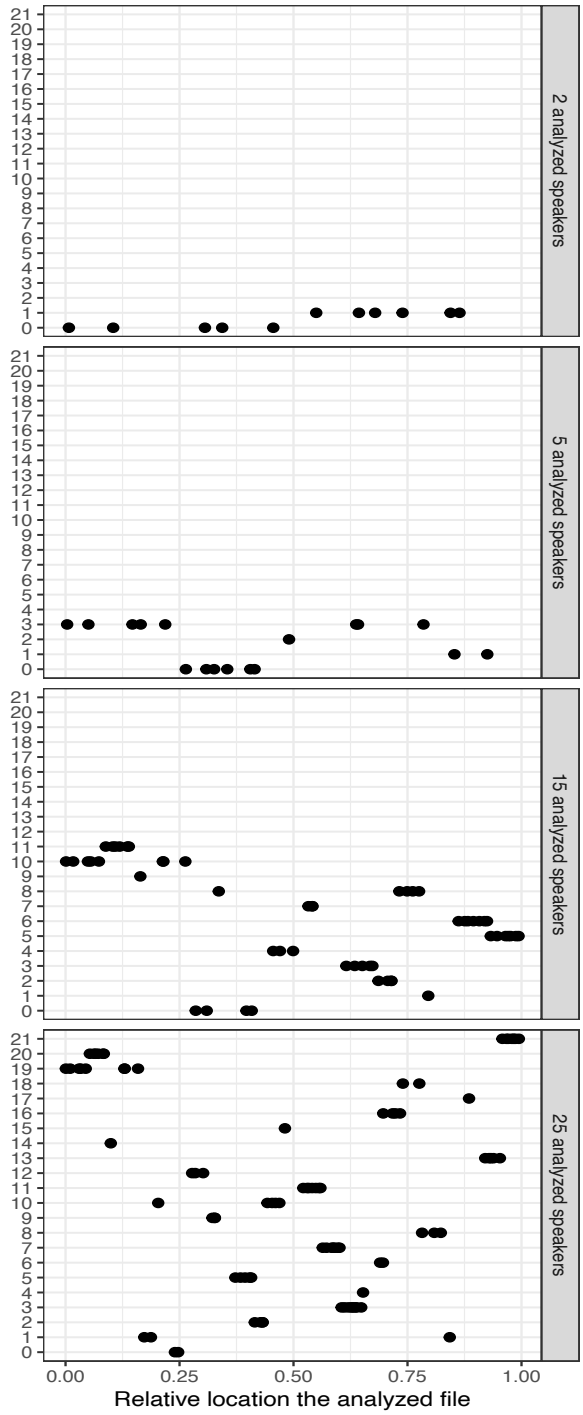

Figure 3. The speaker (identified by the per recording speaker sequence id) is attached to a portion of the automatically identified speech in relation to the relative start time of the speech portion within the recording. Individual speakers are expected to be identified in sequence and within the same part of the recording. The sound file containing read speech of 5, 15 and 25 speakers are visualised separately.

Discussion

The present investigation explored the efficacy of the `pyannote-audio` speech segmentation and speaker diarization in providing draft indexing and separation of speech acts of different speakers across speech corpora. Corpora of varying sizes and with a varying number of speakers were simulated by joining multiple recordings of a text read by different speakers into one file, which was then submitted for automatic analysis. The output was assessed in terms of resources required and the utility of the results for phonetic research.

The results indicated that the application of the automatic speech segmentation model of `pyannote-audio` to produce a draft indexing would provide a valuable addition to the processing workflow of speech corpora. The resulting segmentation included all portions of speech that would be of interest for phonetic researchers and, in that context, correctly excluded, e.g. coughs and portions with silences. The model was reasonably performant in terms of time and memory required and increased linearly in resource use at a reasonable pace within the corpus simulated here. The portions of the recording indexed as potentially having speech in it agreed well with possible definitions of overarching units of speech, which argues for the procedure's utility. The read speech samples investigated here did, however, include portions of speech where the speaker can be perceived to join utterances together more closely than might be the case in spontaneous speech, and manual intervention by a human transcriber will be required to separate utterances that have erroneously been joined into one. Utterance endpoints may require some adjustment in cases of heavy exhales or other speech production artefacts that follow the actual utterance. Speech materials recorded in a less ideal milieu than the samples investigated here may undoubtedly present a more significant challenge for the algorithm, but the overall benefit of a draft indexing of a corpus to identify portion with speech support the inclusion of the procedure into the workflow of speech corpus analysis.

The application of speaker diarization onto the same simulated set of speech corpora produced more mixed results compared to speech segmentation. The application of the model used a consistent amount of memory independent of the number of speakers that the model was tasked to handle, but it can be assumed to take between 6 and 8 times longer than real-time playback of the speech recording to complete. Segmentation of speakers worked well for recordings with few speakers in it but was not able to correctly identify multiple speakers consistently. Therefore, while speaker diarization of singleton recordings even of multi-party conversations may produce a result that requires relatively moderate manual adjustment to separate speech acts of different speakers, and may therefore be worthwhile to apply for indexing, application of the model cannot be expected to consistently separate speakers' speech acts across a corpus. It may be that the relatively poor performance of speaker diarization of multiple speakers was aggravated by my use of read speech samples here, which may be argued to dampen the vocal expressiveness of speakers and therefore limit the basis on which individual speakers could be consistently identified. The question of whether many speakers are more successfully identified and kept separate in spontaneous speech by the speaker diarization model evaluated here requires more research to address. However, as speech collection efforts in many areas of speech research include controlled, including reads of a standard text, as well as spontaneous speech tasks, the results presented here suggest that the utility of speaker diarization using current models may is limited if not applied to speech within a single conversation. Conversations between two parties will likely be more successfully diarised, especially if the speakers are acoustically very different such as in parent-child interactions. Within file application of the model will lead to issues of the same label/identity likely being assigned to different actual speakers and different labels being assigned to the same speaker across different recordings. Manual editing or special facilities in the speech management system to map speakers to consistent identities will therefore likely be required if current models are to be used in a speech database context.

Conclusions

Preprocessing a speech corpus using the speech segmentation models of `pyannote-audio` may provide draft identification of speech acts that are likely to benefit the effectiveness of annotated speech corpora creation. An additional speaker diarization processing step is expected to create issues that need to be manually resolved and therefore offer substantially reduced benefits to a corpus creation process is currently implemented in speech database management systems.

Acknowledgements

This work was conducted as part of the development of the Visible Speech (VISP) platform which is a part of the Swedish national research infrastructure Språkbanken and Swe-Clarin, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

References

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program] (Version 6.2).

Bredin, H., & Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. *arXiv*.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., … Gill, M.-P. (2019). pyannote.audio: neural building blocks for speaker diarization. *arXiv:1911.01255*.

Bullock, L., Bredin, H., & Garcia-Perera, L. P. (2020). Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7114–7118

Draxler, C., & Jänsch, K. (2004). SpeechRecorder-a Universal Platform Independent Multi-Channel Audio Recording Software. *LREC*. 559-562.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32..

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

Yin, R., Bredin, H., & Barras, C. (2018). Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization. *Interspeech 2018*, 1393–1397.