INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

How to annotate 100 hours in 45 minutes

Per Fallgren, Zofia Malisz, Jens Edlund

KTH Royal Institute of Technology

perfall@kth.se, malisz@kth.se, edlund@speech.kth.se

Abstract

Speech data found in the wild hold many advantages over artificially constructed speech corpora in terms of ecological validity and cultural worth. Perhaps most importantly, there is a lot of it. However, the combination of great quantity, noisiness and variation poses a challenge for its access and processing. Generally speaking, automatic approaches to tackle the problem require good labels for training, while manual approaches require time. In this study, we provide further evidence for a semi-supervised, human-in-the-loop framework that previously has shown promising results for browsing and annotating large quantities of found audio data quickly. The findings of this study show that a 100-hour long subset of the Fearless Steps corpus can be annotated for speech activity in less than 45 minutes, a fraction of the time it would take traditional annotation methods, without a loss in performance.

Index Terms: Found data, audio browsing, speech, annotation, human-in-the-loop, dimensionality reduction, speech processing

1. Introduction

Found audio data - audio that was recorded for some purpose other than research - are oftentimes so large that it would take a person many years, maybe lifetimes, just to listen through. The newly released Fearless Steps corpus is a case in point. Recorded in 1968 during the Apollo-11 mission, the corpus consists of 19 000 hours of conversational speech, spanning over thirty channels containing over six hundred speakers[1]. Similar quantities are observed elsewhere, such as in SpeechFind [2] that served as an audio index and search engine for spoken word collections from the 20th century, operating on 60 000 hours of audio, and the MALACH (Multilingual Access to Large Spoken Archives) project [3] that comprises 116 000 hours of multilingual interviews with Holocaust survivors. The collections of national archives dwarfs these corpora: the National Library of Sweden[1] audio-visual archives alone exceeds 20 million hours of audio-visual data.

These are only a handful of examples of large, natural datasets that possess high ecological validity and as such, high value for many research endeavours. Not least in speech and language, but culture overall. Their value is further enhanced if one considers that such large datasets are typically found at public institutions and in archives, and could provide much needed resources for publicly-funded machine learning endeavours. However, delivering the edge relative to commercial players, can occur only once rapid and improved access to found data is warranted. To access and utilise the data, firstly, one needs to tackle their size, as well as their noisy and heterogeneous nature. To this end, we present work built upon previously presented methods designed to explore and annotate large quantities of found audio data.

Annotation is a task that is often performed under strict time and budget constraints [4, 5]. It is difficult to provide exact numbers on how much time a given annotation task demands, but it is usually the case that it takes at least as much time as the duration of the audio that is to be annotated. In this work, we argue that the idea of listening to an audio file sequentially from start to finish is just one alternative. In principle, one could explore and arrange the sound in any order imaginable and furthermore remove the limitations that may be inevitable in a one-dimensional arrangement. We use this idea to circumvent the problem of listening and annotating in linear time by segmenting an audio file into short snippets of equal length and rearranging them in terms of a given acoustic quality in two dimensions. This method, we call Temporally Disassembled Audio (TDA), provides an informative overview of one's data that is more amenable to browsing and quick labelling.

The notion of arranging sounds on a 2-dimensional listening plot was inspired by [6] where they used t-SNE[7] to project spectrograms of 14 000 bird sounds onto a grid. The result was an interface that provided the user with a compact, easily browseable overview of the data with specific bird sounds grouped into regions corresponding to particular species.

Previous versions of the TDA approach, developed further in the present paper, have shown promising results with regard to efficient exploration and delivery of simple annotations on large quantities of audio. Proof-of-concept studies have been conducted on how non-sequential listening methods could be used to get insight into different kinds of audio, such as speech, music and animal sounds[8]. In [9], evidence was provided for the use of the method in accessing and annotating large quantities of audio, lacking any kind of description or transcript. This was achieved by letting participants explore and provide subjective labels on 9 hours of data containing presidential speeches they were not familiar with. The results showed that the entire corpus could be annotated very quickly with adequate accuracy based on the most frequent labels: speech and applause.

In this study we present previous, and novel functionality embedded into an interactive audio browsing interface, we use this tool to annotate speech activity in a subset of Apollo-11 mission data. We evaluate the performance of the tool and proposed human-in-the-loop method by comparing the annotation output with 40 hours of humanly verified labels.

2. Method

2.1. Data

We use the annotated portion of the newly released Fearless Steps corpus[1], consisting of a subset of ~100 hours the 19 000 hours of real-world data that was captured during the Apollo 11 mission. The data was made public for *The Fearless Steps Challenge: Massive Naturalistic Audio*, a special session at Interspeech 2019, which consists of 5 tasks: Speech Activity Detection (SAD), Speaker Diarization, Speaker Identification,

[1] kb.se/kb-in-english.html

http://dx.doi.org/10.21437/Interspeech.2019-1648

Automatic Speech Recognition and Keyword Spotting for Joint Topic and Sentiment. The material provides us with the opportunity to compare the SAD annotation results we can achieve with SAD labels provided with the data.

The data has a sample rate of 8 kHz and is divided into train-dev-eval sets with durations of 58:32:24, 20:10:08 and 20:27:53 respectively, totalling 99 hours and 10 minutes. It should be noted that only the development set has publicly available ground truth transcripts.

2.2. Audio processing

The audio was divided into ten ~10-hour segments, six containing the train data, two the development data and two the evaluation data. Each segment was then individually processed according to the pipeline described below. The entire process can be replicated using the tool which is openly available for download[2].

1. **Temporal disassembly**
   The file was segmented into snippets with a duration of 500 milliseconds and a step size of 500 milliseconds, resulting in a list of roughly 70 000 snippets.
2. **Feature extraction**
   MFCCs were extracted in 50 millisecond long subsnippets: 13 MFCCs (0-12) were computed from 26 Mel-frequency bands, cepstral liftering filter was applied with a weight parameter of 22. Thirteen delta and 13 acceleration coefficients were appended to the MFCC resulting in ten 39-dimensional vectors. These were then concatenated into a flat 390-dimensional feature vector. The openSMILE feature extraction toolkit[10] was used for feature extraction.
3. **Projection to 2D**
   The feature vector set was independently processed by four dimensionality reduction algorithms, specifically t-SNE[7], UMAP[11], PCA[12] and self-organizing maps (SOM)[13], projecting each snippet onto four independent 2-dimensional grids.

2.3. Audio browsing interface

The datapoints, coupled with the 2-dimensional representation of the audio snippets, were plotted in a zoomable and draggable browsing interface, initiated with t-SNE coordinates (see Figure 1). While this paper is not serving as a full documentation of the tool, the following list covers the functionalities relevant for this study.

1. **Listening functionality**
   Every point corresponds to a 500 millisecond long snippet from the original recording and it is possible to listen to a given point, or preferably many simultaneously. The cursor transforms to a circle with adjustable radius when in contact with the plot. By pressing and holding a listening button and positioning the cursor over an area of points, the system samples randomly from the sounds that are in contact with the cursor and fires them with a short and adjustable launch rate. With a short enough launch rate (50-100 milliseconds was used for this study) the system produces a blend of sounds that sound coherent to the human ear. With this listening functionality, the user can move around in certain regions and patterns over the plot to quickly get a sense of what kind of sound the specific region consists of.
2. **Dynamic change of dimensionality reduction**
   With the 2-dimensional vector sets produced by the four algorithms, it is possible for the user to change view. When the user clicks the button for an algorithm change, every point re-positions itself without changing colour and a new distribution emerges which may uncover new information about the data, see Figure 2.
3. **Annotate and export**
   Should the user hear something of interest in a certain region, in the case of the present study this would be speech or the absence of speech, they can annotate it. This is done by selecting a colour and providing a label for it, then pressing the colouring button while hovering over a group of points. As a results the target points are labelled and coloured accordingly. When finished, there is an export functionality that outputs a temporally ordered list of each snippet and their respective label.
4. **Timeline and sequential listening**
   In the lower section of the interface, there is a horizontal representation of the timeline. When a point is coloured, the corresponding timestep in the timeline is also coloured. This, in combination with the possibility to click on any timestep and play the original audio sequentially, allows the user to get quick feedback of where the annotated interval occurs in the original audio.

2.4. Annotation

Using the interface and data described above, every 10-hour segment from the 100-hour subset of the Fearless Corpus was annotated for speech activity[3]. All segments were annotated in one consecutive session during which time spent was recorded.

A random sampling was performed to only present 10% of the datapoints to the annotator, reducing the original N from 70 000 to 7 000 in the downsampled view. After annotation, every skipped point was assigned a label identical to the point with the shortest Euclidean distance (in the t-SNE plot) in the downsampled view. During this process, labels were also assigned to any points that were in the downsampled view but were skipped during annotation. A second set of labels was produced by applying a median filter of 3 on the system output, i.e. a sequence of labels corresponding to 500 millisecond snippets.

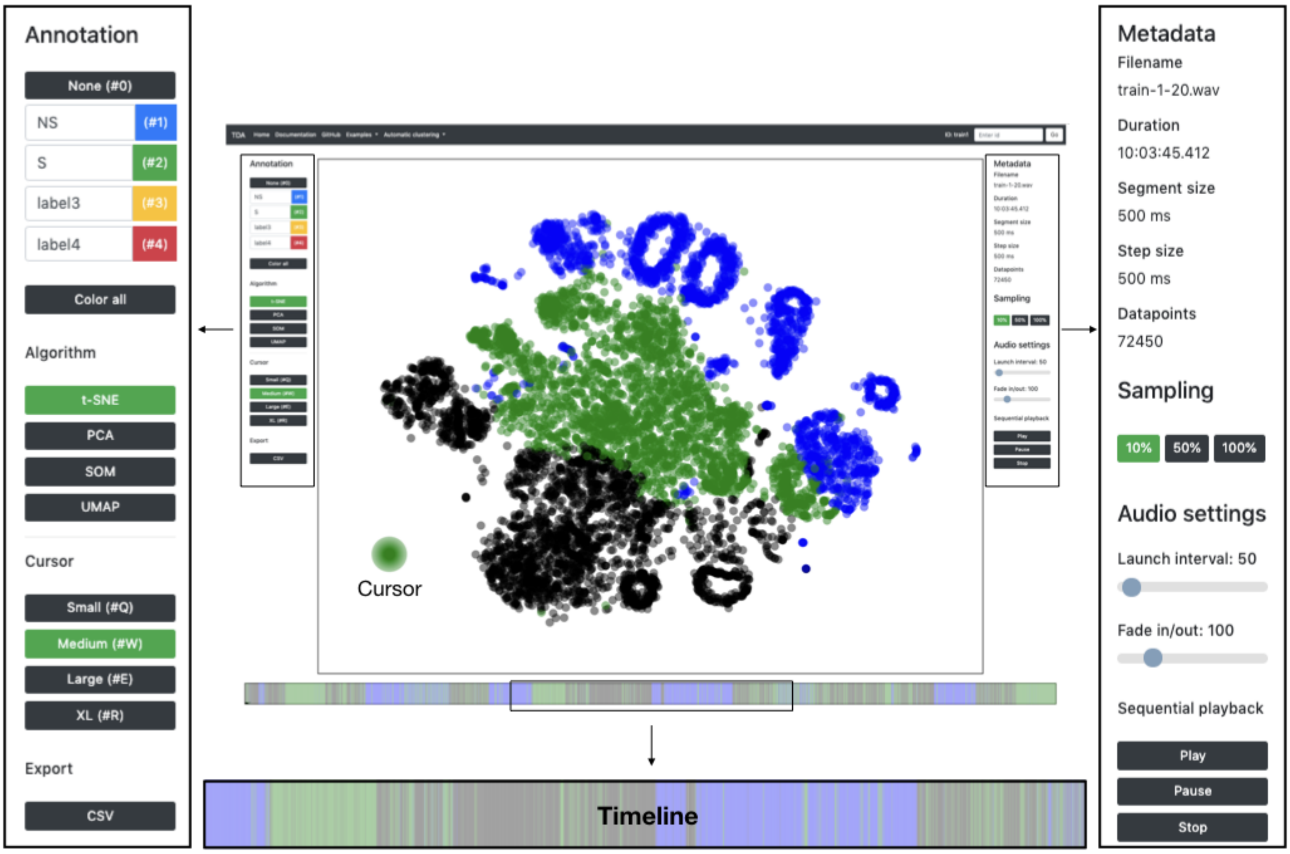
3. Results

The produced transcripts were then compared to the transcripts included in the Fearless Steps corpus using DCF (detection cost function) which is the main reference for comparison in the Fearless Steps SAD-challenge. Additionally the constituents of DCF, namely, the probability of false alarms (P$_{fp}$) and missed detections of speech (P$_{fn}$) are included. Precision and recall for the speech class is also calculated on every tenth millisecond which was the resolution of the provided annotations. As ground truth (humanly annotated) transcripts were only provided for the development set, this is the basis for the main result analysis. For completeness, we include the results of our labels on the evaluation data as well, produced by the Fearless Steps team.

The results are also compared to the baseline provided by the Fearless Steps team, as described in [14]. As this software

[2] github.com/perfall/Temporally-Disassembled-Audio

[3] The main author of this paper performed the annotation.



Filename
train-1-20.wav

Duration
10:03:45.412

Segment size
500 ms

Step size
500 ms

Datapoints
72450

Launch interval: 50

Fade in/out: 100

Figure 1: *Screenshot of the tool during the process of annotating speech activity in ten hours of Fearless Steps training data. The medium-sized cursor in the bottom-left part of the plot is ready to deploy green colouring to the datapoints, corresponding to the label S (speech). As seen in the left toolbar, t-SNE is chosen for the active view. The timeline (bottom) is also automatically coloured accordingly, giving the annotator quick feedback on when the target segment occurs in the original audio. The toolbar to the right contains metadata and a functionality for sampling used for plotting a subset of the data. Audio settings are used for the listening function and sequential playback, the latter can also be controlled by clicking on a certain position within the timeline itself.*

Table 1: *Annotation results in percentages compared to the provided labels from the Fearless Steps corpus. TDA and TDA$_{mf3}$ is the main output of our tool (the latter with an applied median filter of 3), FS$_{bs}$ is the baseline provided by Fearless Steps, and kM$_2$ is the two cluster k-means baseline. Best results for each metric is in bold.*

Data	Metric	TDA	TDA$_{mf3}$	FS$_{bs}$	kM$_2$
Dev	**DCF**	6.35	**5.9**	8.6	17.52
	P$_{fp}$	2.08	1.79	-	**0.44**
	P$_{fn}$	7.77	**7.3**	-	23.21
	Precision	79.54	77.88	-	**91.24**
	Recall	92.22	**92.73**	-	76.8
	F-score	**85.41**	84.66	-	83.4
Eval	**DCF**	8.24	**7.99**	11.7	-

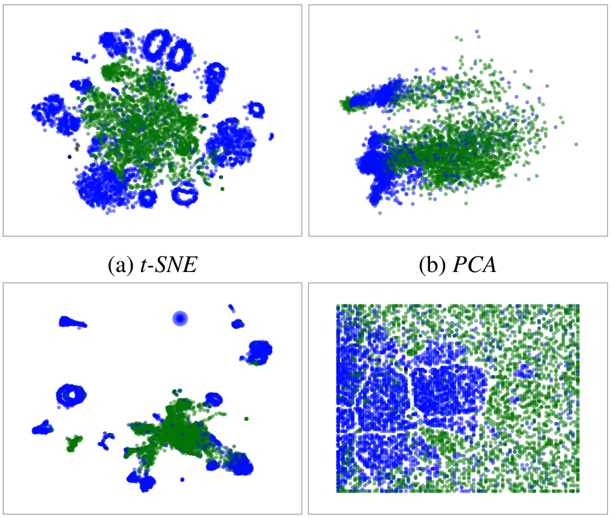
was not released in time for submission deadline only the publicly reported DCF results will be used for comparison with the baseline. For further comparison, a second baseline using k-means clustering was calculated on the 390-dimensional MFCC space within the annotation interface using a k of 2 and assigning the speech label to the cluster that seemed to contain more speech based on quick listening.

The results (Table 1) show that for the main reference metric DCF, the system output (TDA) and the median filtered output of 3 (TDA$_m f3$) outperforms both baselines. This is the case for the development set as well as the evaluation set. The kM$_2$ baseline produces the best results for P$_{fp}$ and precision while TDA$_m f3$ performs best for P$_{fn}$ and recall. TDA achieves highest F-score. The average duration for an annotation session of 10 hours was 261 seconds (SD = 34.7) totalling 43 minutes and 31 seconds.

4. Discussion

The results show that it is possible to use the concept of temporally disassembled audio to get a quick overview of large datasets and provide labels in a short period of time that are better than the two presented baselines in terms of the standardised detection cost function. These findings corroborate previous results and provide an incentive for further development and research.

The two cluster k-means was included in the evaluation as a second baseline, conveniently built into the browsing interface. This clustering technique has also shown adequate performance in earlier experiments. While it did not perform well on the metrics DCF, P$_{fn}$ and recall, it achieved high results on precision.


(a) *t-SNE* (b) *PCA*

(c) *UMAP* (d) *SOM*

Figure 2: *Visualising the different dimensionality reduction algorithms used in the annotation interface using ten hours of the Fearless Steps training data. Datapoints and colour assignment was not changed between plots. While there are clear similarities between them in terms of homogeneously coloured regions, it is evident that the four methods emphasise different aspects of the data.*

The likely reason is that it managed to isolate a homogeneous cluster of prototypical speech snippets, which in certain scenarios is advantageous. However, the difficulty in SAD is, arguably, to distinguish segments that are not clear speech candidates. This is captured by the DCF metric that weighs false negatives higher. As such, including a non-speech segment rather than excluding a speech segment, leads to more optimal results.

Regarding precision and recall, the balance between the two metrics is easily controlled using the presented interface: finding the most evident speech region of the plot and labelling that area as speech allows for achieving higher precision. Alternatively, one could label regions with extensive overlap between speech and non-speech to maximise recall. This flexibility could be put to use in many other tasks and further motivates the flexibility of human-in-the-loop frameworks.

We need to stress that the aim of the tool is not to provide a general-purpose annotation framework, neither should it be considered an alternative to automatic annotation. Rather, the goal is to provide researchers and scholars that work on large audio datasets with an interface that gives them a quick and intuitive insight into their data.

For instance, in certain scenarios where transcripts of large audio files and automatic classifiers do not exist, the method would be deployed to give the user a first set of labels used to bootstrap a classifier. Alternatively one could use it to simply browse and explore one's data. The latter could bring important insight into the data that could be lost otherwise. As an example, during the annotation process of this study, isolated regions of high-pitched signals were observed that seemed to coincide with speech in the timeline. It was found that in certain segments of the audio, the recordings have picked a high-pitched noise occurring whenever people pressed the transmission button. It was also instantly evident that the background noise of the audio was inconsistent between recordings, additionally, no silent segments were observed. These are just some examples of observations that could be made with a human-in-the-loop framework such as the one in this work. Conclusively, in contrast to the trend of learning from large amounts of data via a blind classifier with the aim of getting high numbers as output, we argue that the combination of a human and technology is fruitful for many speech processing tasks.

4.1. Future work

The tool at the time of writing is available for download. Further work on making the tool more robust and documenting the installation process and functionality is currently underway. For this paper, the audio was divided into segments with a duration of 10 hours, additional functionality to browse or sample even more data in one session is of interest. Furthermore, letting sound snippets temporally overlap is also worth exploring, currently the temporal disassembly was performed with a resolution of 500 milliseconds with no overlap. As such, it is possible that a given snippet contains both speech and non-speech but as the annotation process is binary and does not distinguish between these categories some performance limitations are expected.

5. Conclusion

This work shows that the TDA based audio browsing technique can be used to annotate large quantities of audio quickly. Specifically, we have shown that using the described annotation interface, it is possible to annotate speech activity in 100 hours of real-world audio in under 45 minutes without suffering a loss in quality.

6. Acknowledgements

The project is funded in full by the Riksbankens Jubileumsfond funded project TillTal (SAF16-0917: 1). Its results will be made more widely accessible through the national infrastructure Nationella Språkbanken and Swe-Clarin (Swedish Research Council 2017-00626).

7. References

[1] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," *Proc. Interspeech 2018*, pp. 2758–2762, 2018.

[2] J. H. L. Hansen, R. Huang, B. Zhou, M. S. Seadle, J. R. D. Jr., A. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5-1, pp. 712–730, 2005. [Online]. Available: https://doi.org/10.1109/TSA.2005.852088

[3] J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. J. Byrne, J. Hajič, S. Gustman, and B. Ramabhadran, "Automatic transcription of Czech language oral history in the MALACH project: Resources and initial experiments," in *Text, Speech and Dialogue*, P. Sojka, I. Kopeček, and K. Pala, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 253–260.

[4] D. Suendermann, J. Liscombe, and R. Pieraccini, "How to drink from a fire hose: One person can annoscribe 693 thousand utterances in one month," in *Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue*. Association for Computational Linguistics, 2010, pp. 257–260.

[5] C. Draxler, "Webtranscribe–an extensible web-based speech annotation framework," in *International Conference on Text, Speech and Dialogue*. Springer, 2005, pp. 61–68.

[6] K. Tan, M. McDonald. (2017) Google ai experiments: Bird sounds. [Online]. Available: https://experiments.withgoogle.com/bird-sounds

[7] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[8] P. Fallgren, Z. Malisz, and J. Edlund, "Bringing order to chaos: a non-sequential approach for browsing large sets of found audio data," in *Proc. of the 12th International Conference on Language Resources (LREC2018)*, Miyazaki, 2018.

[9] P. Fallgren, Z. Malisz, and J. Edlun*d*, "Towards fast browsing of found audio data: 11 presidents," in *DHN2019*, Copenhagen, 2019.

[10] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[11] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: uniform manifold approximation and projection," *J. Open Source Software*, vol. 3, no. 29, p. 861, 2018. [Online]. Available: https://doi.org/10.21105/joss.00861

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[14] A. Ziaei, L. Kaushik, A. Sangwan, J. H. L. Hansen, and D. W. Oard, "Speech activity detection for nasa apollo space missions: challenges and solutions," in *INTERSPEECH*, 2014.