Utilising large quantities of found speech data

Per Fallgren

KTH Royal Institute of Technology

perfall@kth.se

Abstract

Collections of found speech data - data that was not recorded for research - have many benefits in a number of scenarios, not least because of their great size. There are also challenges to tackle when it comes to utilising found speech, the main one arguably being that there are no adequate methods that are able to handle huge quantities of noisy and heterogeneous data efficiently. With this in mind, my research is about investigating, producing and evaluating methods that aim to properly make use of the valuable information that is embedded in found data. To this end, my main contribution is an original approach that revolves around a semi-supervised human-in-the-loop framework for utilising large quantities of audio. I will in this Doctoral Consortium paper discuss my work so far in combination with plans for the future.

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

1. Introduction

National archives and similar organisations contain extremely large quantities of audio data. In Sweden the largest archives, ISOF (Institute for Language and Folklore) and the KB (National Library of Sweden), hold over 20,000 hours and 10,000,000 hours of audio or audio-visual recordings respectively. Similar numbers can be observed in many other countries' archives, as discussed in [1]: In Sweden, project TillTal [2] aims to organise the quantities of the mentioned Swedish archives. [3] presents a software platform for automatic transcription and indexation of Czech and former Czechoslovakian radio archives (100,000+ hours of audio). There are older initiatives with similar aims: SpeechFind [4] is an audio index and search engine for spoken word collections from the 20th century containing 60,000 hours of audio; [5] considers automatic transcriptions of the Institut National de l'Audiovisuel archives in France containing 1.5 million hours of radio and television programs dating back to 1933 and 1949 respectively; The MALACH (Multilingual Access to Large Spoken Archives) project [6] addressed the large multilingual spoken archives containing 116,000 hours of interviews from holocaust survivors; CHoral [7] considers audio indexing tools for the Dutch audiovisual cultural heritage collections.

Data collections like the ones mentioned above are often referred to as found data[1] - data that was not recorded with the specific purpose of being used in research. Examples of found audio data are radio and TV recordings, interviews, podcasts, audiobooks and archived data in general (to name a few). There are a number of reasons to favour found data over other datasets. Compared to data collected in a controlled lab setting found data has a much higher ecological validity given the naturalness of the data. Cultural worth is another aspect to consider as there is a lot of historical and cultural information embedded in recordings ranging from the dawn of audio recording to today. Finally, the sheer size of found data collections is an argument in itself - today data is interesting in its own right given the big data boom and the constant need for more data for machine learning. As such, found data is in many cases advantageous to alternative data collections. This also a challenge however, as there are no tools that can handle the large, diverse, often noisy, quantities at hand. This is where my research comes in, in which I aim to produce and evaluate methods and tools to utilise large quantities of speech data.

[1] Although I use the term found data freely, my research is almost fully regarding audio data.

2. Temporally disassembled audio

With the aim of producing methods for exploring, and to some extent annotating, large quantities of audio we came up with a concept that we have come to call Temporally disassembled audio (TDA). The notion is based on the unintuitive idea that one does not necessarily need to listen to a sample of audio sequentially from start to finish to get a grasp of what the sample contains. Say you have an hour a recorded material that you want to explore. By for instance segmenting the audio into 3600 short snippets of 1 second in length you could choose to organise them in whatever order you want. Organising them based on their temporal positioning might be the most natural, but you could also choose to sort them based on certain acoustic features, and even distribute them in two, or even three dimensions. TDA is based on these ideas - by temporally disassembling audio into short snippets, organising them by projecting an extracted set of feature vectors on a 2D plot, and adding a listening functionality we have shown that one can discover certain aspects of the audio very efficiently. Different versions of a tool that is built upon these ideas have been implemented, the latest is available for download at github.com/perfall/Edyson[2]. In short process of the method has the following steps:

1. Segment a given audio file into short snippets of equal length (50ms - 1000ms).
2. Extract some acoustic features for every snippets, MFCCs are used by default.
3. Plot every snippets as a point by projecting each feature vector onto a 2D grid using a dimensionality reduction method, e.g. t-SNE[8], UMAP[9], PCA[10] or self-organizing maps (SOM)[11].
4. Use a looping listening function that lets the user listen to regions of the plot by navigating via the cursor.
5. Use an annotation function that lets the user label regions of interest.
6. Revert to the temporal domain with potential labels, resulting with some crude annotations for the audio file.

[2] Included in the repository is a link to an online demo which I do recommend the reader trying out to get a better sense of how the tool works.

3. Previous results

The following section briefly presents the results of three published papers produced during the first two years of my PhD studies. They are presented in chronological order and revolve around the notion of TDA (as previously described), which is arguably the major contribution of my work.

3.1. Paper A - Bringing order to chaos: A non-sequential approach for browsing large sets of found audio data

[12] can be viewed as a proof-of-concept study that presented early versions of the method and did as such provide an incentive for further work. Specifically, it was shown that TDA could be used to get an insight into several different kinds of audio, including Swedish speech, animal sounds and music.

3.2. Paper B - Towards fast browsing of found audio data: 11 presidents

Using similar implementations of TDA [1] presented a user study where 8 participants were asked to explore and label 10 hours of (to them) unknown data using the proposed approach. The data was a concatenation of 11 presidential speeches, ranging from Kennedy in 1961 to Trump in 2017. It was shown that participants were able to distinguish, and furthermore annotate, three salient aspects of the audio in a matter of minutes - namely speech, applause and silence.

3.3. Paper C - How to annotate 100 hours in 45 minutes

The most recent addition with regard to my research is [13], where we provide evidence for the annotation potential for the method. In addition, this paper present the tool (now named Edyson) which is open source and available for anyone that is interested (see Figure 1). It was shown that 100 hours of the Fearless Steps corpus [14], a collection of the Apollo-11 radio transmission data, could be annotated for speech activity in the fraction of the time it would take using a traditional annotation method.

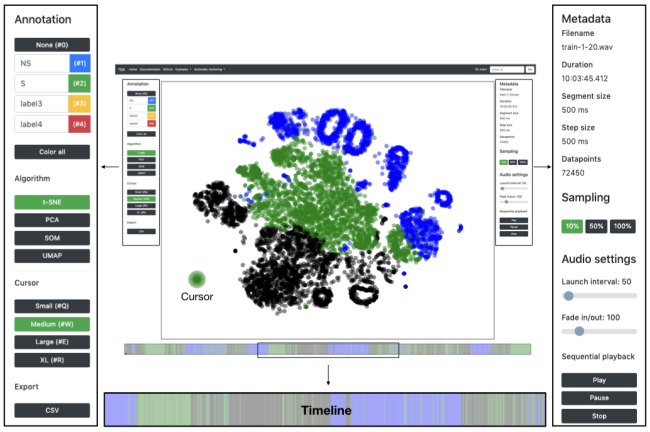

Figure 1: *Screenshot of the tool during the process of annotating speech activity in ten hours of Fearless Steps training data. For a more elaborate description see [13].*

4. Discussion and future work

The mentioned results along with separate pilot studies shows that the method has potential regarding utilising large quantities of noisy speech data. Both in terms of exploring audio efficiently, with the purpose of simply getting some insight in one's data, and for performing some rudimentary annotation. There are however some points that need to be mentioned, that mostly revolve around *what* and *who* the audio browsing tool is for. Starting with the *what* a common question I get is why one should use this tool when there are automatic classifiers for fields such as speech activity and applause detection. To that I usually say that 1, if you do not know what your audio contains (which is a common scenario in archives) how would you know what classifier to use. Also it should not be compared to automatic tools, nor manual annotators, but viewed in its own right; and 2, automatic methods seldom perform well in noisy and heterogeneous conditions. Regarding the *who* the tool is for I'm not certain, although there is evidence for the potential of the approach this evidence is fully dependent on a user that at least has some rudimentary understanding of feature extraction and dimensionality reduction. This can however not be a requirement if the aim is to let *anyone* that is wanting to explore their audio utilise the method. The problem essentially comes down to whether or not our findings should be utilised anywhere, or only by "experts". As such, there are some challenges to consider that I would like to get some feedback on.

4.1. Future work

Depending upon the above discussion there are many potential branches to pursue in the future. Further empirical studies for what the method can and cannot handle is certainly of interest. One concrete example, that there has been some experimenting with already, is the idea of extracting vowels in noisy data using the tool. On the topic of noisy archive data it would also be of interest to randomly select a number of samples from an audio archive and in some systematical manner categorise these according to one's findings. This would be of great interest for archivists and other individuals with large quantities of unstructured audio on hand, should the results be decent. Finally, an extensive literature study on found speech and methods to utilise it would be valuable, not least for me personally and my future thesis but for the field as a whole.

5. Acknowledgements

Thank you to my great supervisors Jens Edlund and Zofia Malisz. 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).

6. References

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

[2] J. Berg, R. Domeij, J. Edlund, G. Eriksson, D. House, Z. Malisz, S. Nylund Skog, and J. Öqvist, "TillTal – making cultural heritage accessible for speech research," in *CLARIN Annual Conference*, Aix-en-Provence, France, 2016.

[3] J. Nouza, K. Blavka, P. Cerva, J. Zdansky, J. Silovsky, M. Bohac, and J. Prazak, "Making Czech historical Radio archive accessible

and searchable for wide public," *Journal of Multimedia*, vol. 7, no. 2, pp. 159–169, 2012.

[4] 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

[5] C. Barras, A. Allauzen, L. Lamel, and J.-L. Gauvain, "Transcribing Audio-Video Archives," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002. [Online]. Available: ftp://tlp.limsi.fr/public/ica02cb.pdf

[6] 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.

[7] R. J. Ordelman, F. M. de Jong, and W. Heeren, "Exploration of audiovisual heritage using audio indexing technology," in *Proceedings of the First European Workshop on Intelligent Technologies for Cultural Heritage Exploitation*. Università di Trento, 2006.

[8] 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

[9] 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

[10] 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.

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

[12] 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.

[13] P. Fallgren, Z. Malisz, and J. Edlun*d*, "How to annotate 100 hours in 45 minutes," in *Interspeech*, 2019, p. TBA.

[14] 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.