A Tool for Exploring Large Amounts of Found Audio Data

Per Fallgren, Zofia Malisz, Jens Edlund

KTH Royal Institute of Technology
perfall@kth.se, malisz@kth.se, edlund@speech.kth.se

Abstract. We demonstrate a method and a set of open source tools (beta) for non-sequential browsing of large amounts of audio data. The demonstration will contain versions of a set of functionalities in their first stages, and will provide a good insight in how the method can be used to browse through large quantities of audio data efficiently.

Keywords: Found data, visualization, machine learning, speech processing

1 Introduction

In many fields, the absence of data is no longer a pressing issue, instead there is a lack of methods that are able to handle the large collections of data that exists. To this end we present an early version a tool that lets professionals from different fields explore audio more efficiently. Our aim is to make it possible to utilize found data, meaning data that was not recorded with the purpose of being used in (speech) research. Examples of found data can include archive data, radio and television speech, and interviews. These data sets constitute speech found in the wild, and are often more valuable for research than manually constructed speech datasets as they are not constrained by a fabricated lab setting. Despite this, these kinds of data are rarely used, the reasons sometimes being legal and ethical issues or the simple fact that the existence of the data is unknown to many (Edlund & Gustafson, 2016).

Large collections of this kind of data abound. In Sweden, the Institute for Language and Folklore (ISOF) hosts 13 000 hours of digitized speech, and the National Library (KB) hosts a staggering 10 million hours of audiovisual data. To put it simply, there is not a lack of data, but a lack in generalizable tools and methods that can help make use of the data. While one in theory could listen through 13 000 hours of data sequentially it would take several years, not taking into account added time for annotation. Instead, novel methods that dramatically cut down on the time it takes to explore sound are required. With this in mind, we present a tool in its early stages that fits well in the context of digital humanities given its ability to let a user browse their data in a more efficient manner. It should prove useful to a wide variety of professionals, not only in speech and speech technology research, but also to archivists, scholars and others in the social sciences and humanities.

2 Method

Our approach removes the temporal dimension – the sequential layout of the acoustic data – and instead organizes the data along a low number (2-3) of spatial dimensions representing acoustic features of the audio.

For each audio file, we create a monochrome spectrogram, i.e. a visual representation of the sound frequencies and their intensity over time. The two representations (audio and spectrogram) are then segmented in parallel into equal-sized chunks of T duration (T should sensibly be in the range of 50 or so milliseconds to one or two seconds).

The spectrogram segments are then fed to a dimensionality reduction algorithm that projects each segment onto a 2D plane – a technique often used to find similarity in images. As of now, we have tested t-SNE (Van Der Maaten & Hinton, 2008) and Self-Organizing Maps (SOM; Kohonen, 1982) for the dimensionality reduction. The former is very efficient and mostly used for testing, while the latter better retains the topology of the high-dimensional data.

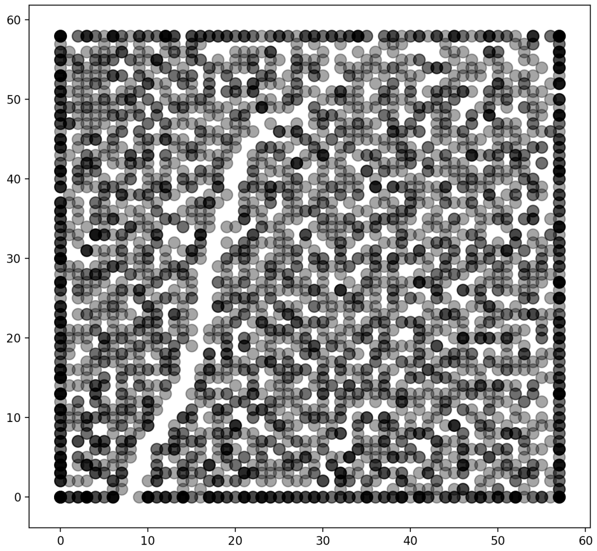

Figure 1: A clear separation of verbal chattering in a cafe. Each datapoint represents one second extracted from a one hour recording (downloaded from freesound.org) and its position corresponds to its acoustic properties, hence similar sounding segments are positioned in close proximity of each other. On listening, it's clear that the audio of the left region has a constant background noise that is not present in the larger right region.

3 Discussion

The processing will potentially result in a distribution of spectrogram chunks that have formed coherent regions in the sense that similar sounds are positioned in the same vicinity. Plotting the data points and adding advanced listening functionality results in a simple, interactive interface allowing its users to explore the initial audio at a considerably higher pace than the real-time duration of the initial audio.

Our proof-of-concept experiments show that the method captures characteristics of audio that are readily interpreted by a human listener. Among the audio types we have experimented with are speech, where we find crude regions of vowels, consonants and silence; animals noises, where the sounds of birds, cows and sheep are distinguished almost perfectly; music, where the algorithm differentiates between a singing voice and an instrument (guitar or piano); and café buzz (Fig. 1), for which verbal chattering with and without background noise is distinguished. One might argue that there already exist methods that potentially would outperform the proposed approach in the mentioned tasks, however these methods would all most likely lack generalizability. They are as such tuned for certain kinds of audio and are not usable when the nature of the data is unknown, which is often the case when dealing with found data.

4 Future work

The method is still under development and there are many directions to explore. Our vision is to turn the tools into a web based framework that can be used by anyone regardless of hardware and operating system. We also wish to consider that our target audience has a varied technology expertise, hence we want the tools to be easy to use for a wide range of people. The interaction between user and framework is an essential means to strengthen the link between users and their data. A prerequisite is a set of tools that provides a smooth and pleasant experience.

Sound can be represented in a wide variety of manners. Each of these captures some characteristics better and others worse. In our current version, the SOX library[1] is used to extract the spectrogram we use for clusters. Although we show promising results with this approach, there are other techniques that might capture acoustic features better. Regarding characteristics that are relevant to speech, window size adjustments are a primary candidate for tweaking. Analysis of small windows (e.g. 25ms) will create very different maps compared to larger (e.g. 1s) windows, something that will also greatly affect the listening experience.

Furthermore, we will add an annotation function, so that a user interested in a certain type of sound event can locate and tag a region for further exploration with ease. Additionally, the user will have the option to revert to the original audio with new information and be able to see when any observed sound events occurred in the sequence. This would also facilitate the incorporation of a human-in-the-loop, which is the idea of retraining a map based on human feedback.

The demonstration will contain first versions of these functionalities, and will provide a good insight in how the method can be used to browse through large quantities of audio data efficiently.

5 Acknowledgements

The project described here is funded in full by Riksbankens Jubileumsfond (SAF16-0917: 1). Its results will be made more widely accessible through the infrastructure supported by SWE-CLARIN (Swedish research Council 2013-02003).

[1] sox.sourceforge.net

References

Edlund, J., & Gustafson, J. (2016). Hidden Resources — Strategies to Acquire and Exploit Potential Spoken Language Resources in National Archives. In N. C. (Conference Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, … S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Van Der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605. https://doi.org/10.1007/s10479-011-0841-3