Towards fast browsing of found audio data: 11 presidents

Per Fallgren[0000-0003-1262-4876], Zofia Malisz[0000-0001-5953-7310], and Jens Edlund[0000-0001-9327-9482]

KTH Royal Institute of Technology, Speech, Music & Hearing, Stockholm
{perfall, malisz}@kth.se, edlund@speech.kth.se

Abstract. Our aim is to rapidly explore prohibitively large audio collections by exploiting the insight that people are able to make fast judgments about lengthy recordings by listening to temporally disassembled audio (TDA) segments played simultaneously. We have previously shown the proof-of-concept; here we develop the method and corroborate its usefulness. We conduct an experiment with untrained human annotators, and show that they are able to place meaningful annotation on a completely unknown 8 hour corpus in a matter of minutes. The audio is temporally disassembled and spread out over a 2-dimensional map. Participants explore the resulting soundscape by hovering over different regions with a mouse. We used a collection of 11 State of the Union addresses given by 11 different US presidents, spread over half a century in time, as a corpus.

The results confirm that (a) participants can distinguish between different regions and are able to describe the general contents of these regions; (b) the regions identified serve as labels describing the contents of the original audio collection; and (c) that the regions and labels can be used to segment the temporally reassembled audio into categories. We include an evaluation of the last step for completeness.

Keywords: found data, speech processing, self-organizing maps, dimensionality reduction, visualisation

1 Introduction

This paper presents a method to explore prohibitively large audio collections rapidly using a combination of techniques. The work is motivated by the prevalence of audio archives that remain unused and unexplored because of their large size.

For example, the audiovisual archives of the Swedish National Library currently contain more than ten million hours of data - an amount that would take 100 people spending 40-hour weeks throughout their entire working lives just to listen through. The scope of the project motivating the present work is slightly less daunting: 13000 hours of mixed audio recordings gathered over a period of more than a century.

Previous results show that our method can help differentiate audio segments [1]. It facilitates browsing and annotation of lengthy recordings in little time by combining a number of techniques designed to present large quantities of audio simultaneously and by building experimental setups that allow listeners to judge what they hear quickly.

The present work extends these findings and corroborates their usefulness through a study in which 8 participants are used as annotators of an 8 hour corpus containing material unfamiliar to them. The study simulates a situation of an archivist faced with a set of audio recordings they do not know anything about and do not have labelling for - a situation that takes place often enough in current archive work. The goal for the archivist is to explore the data, get a sense of what it contains, and possibly provide some crude annotations. Rather than listening through the recordings from start to finish, which is a time consuming and tiresome task, even if the data is sampled sparsely, the intention is that they use the proposed method to facilitate the process.

2 Background

Found data comprises data collections that were not recorded with the specific purpose of being used in research. As such, typical examples of found resources are political speeches, radio segments, interviews, audiovisual data such as television and movies, music, recordings and transcriptions of the NASA's Apollo missions [2] and archived material in general. In many cases, these data collections are of higher value compared to artificially constructed data sets with regard to ecological validity. As such, there is no risk for any unnatural properties in the data that may have been transferred from a poorly constructed lab setting. Additionally, they possess significant cultural value, their full potential is, however, not realisable without the help of automatic processing. Furthermore, the sheer size of found data collections demonstrates that there is not a shortage of data out there, rather a lack of methods that are able to handle the huge quantities at hand. Unsupervised machine learning methods are very helpful in this domain, but to perform conventional supervised classification tasks, one needs labels.

To tackle this, many initiatives have been started with a focus on national collections of archive data. In Sweden, the project TillTal [3] aims to organize Swedish archives that have collections exceeding 10 million hours of audiovisual data, a number that is increasing significantly every day. [4] describes a software platform for automatic transcription and indexation of Czech and former Czechoslovakian radio archives, containing more than 100,000 hours of audio. Furthermore, there are older initiatives with similar aims: SpeechFind [5] had the purpose to serve as an audio index and search engine for spoken word collections from the 20th century containing 60,000 hours of audio; [6] considers automatic transcriptions of the INA (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 [7] addressed the problems in analysis and processing of large multilingual spoken archives containing 116,000 hours of interviews from Holocaust survivors while CHoral [8] considered audio indexing tools for Dutch audiovisual cultural heritage collections.

3 Method

In an effort to corroborate the method, we set up an experiment with people with no prior knowledge of the project or the method.

3.1 Data

The data set was chosen to be manageable from an experimental point of view, yet representative of archive (found) data. It consists of 11 American State of the Union addresses recorded over a span of half a century, with the oldest being delivered by John F. Kennedy the 25th of May in 1961 and the newest by Donald Trump in 2017. Albeit a restricted domain, the data set holds the type of variability one might expect from unknown archival data: different speakers, different venues, different equipment, different post processing, different audio quality different content, and different times. As such, we deemed this data to be suitable for corroborating the method.

The speeches range from 32 minutes and 38 seconds to 1 hours 1 minute and 56 seconds. Each audio file was converted to one channel with a sample rate of 16kHz. Table 1 shows the details of the recording[1].

3.2 Stimuli generation

The method we are exploiting starts with temporally disassembling the audio signal by chopping it up in small segments that are then reorganized without consideration of their original temporal organization. We have used the technique (TDA) in several applications where we insert humans-in-the-loop of audio processing. In this case we used SOMs, self-organizing maps [9], to organize the sound segments in two-dimensional

Table 1. Details of the 11 state of the union speeches

President	Date	Duration
John Fitzgerald Kennedy	May 25, 1961	0:45:37
Lyndon Baines Johnson	Jan 8, 1964	0:40:17
Richard Milhous Nixon	Jan 30, 1974	0:48:49
Gerald Rudolph Ford	Jan 19, 1976	0:49:30
James Earl Carter, Jr.	Jan 23, 1980	0:32:38
Ronald Wilson Reagan	Jan 26, 1982	0:43:27
George Herbert Walker Bush	Jan 29, 1991	0:47:21
William Jefferson Clinton	Jan 23, 1996	1:01:56
George Walker Bush	Sep 20, 2001	0:38:05
Barack Hussein Obama	Jan 12, 2016	0:58:50
Donald John Trump	Feb 28, 2017	1:00:16

[1] The recordings were downloaded from americanrhetoric.com and youtube.com

maps. We then take advantage of a proof-of-concept technique, massively multi-component audio environments [10], to present a multitude of sound snippets simultaneously. The soundscape created in this manner is what was explored by our participants.

Each recording was converted into a greyscale spectrogram using the Sound EXchange Library (SOX)[2]. Apart from adopting a temporal resolution of 1000 pixels per second and a height of 65 pixels, default settings were used. Both audio and spectrograms were then segmented into equal sized chunks resulting in 100 ms long segment-pairs of audio and spectrogram - disassembled audio with connected spectrograms. Each spectrogram was then converted into a 6500 dimensional vector where each element corresponds to the greyscale value of the given pixel. This results in a matrix where each row corresponds to a feature vector of its original audio segment. We sampled over this matrix extracting one segment per second which gave us a training data set of ~32000 datapoints.

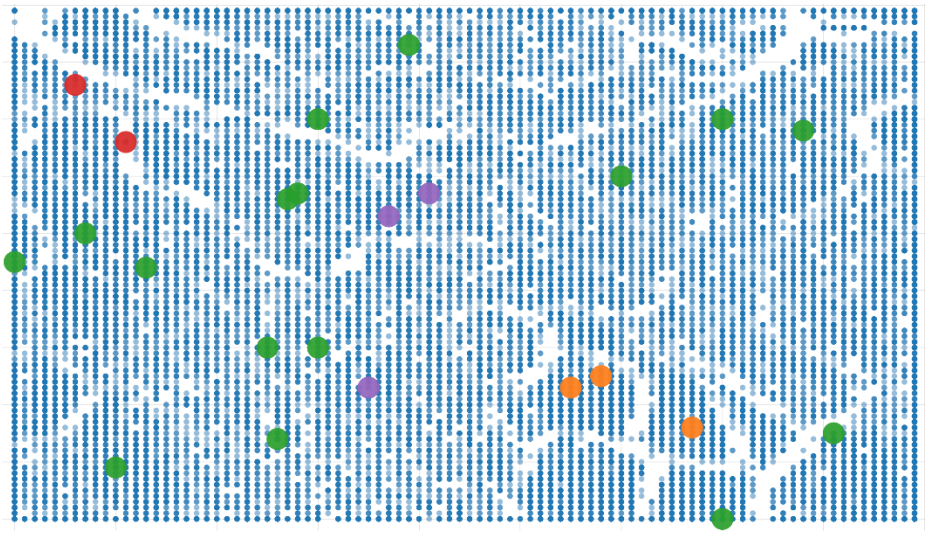
The data was then used as input into a SOM projecting each data point onto a 90x90 2-D grid with the goal of forming regions based on the audio characteristics of each segment. We then visualise the grid as a plot where each audio segment is represented by a point in a certain position. By assigning the corresponding audio to each point, so that the target audio segment is played when a user hovers over it with the mouse, we get a simple interactive interface that can be used to browse the original audio in a more efficient manner. The technical details of the proposed method, and motivation to our decision on using self-organizing maps, is described in more detail in [1].

3.3 Subjects

8 participants (M = 30.38 years, SD = 7.24, 22 - 46 years, Male = 6, Female = 2), all without known hearing impairments, were given the same task. One participant misunderstood the task and did not provide labels, and is therefore excluded from the results by necessity.

[2] sox.sourceforge.net

Fig. 1. Centroids proposed by the participants (orange=applause; purple=speech; red=silence, green=other)



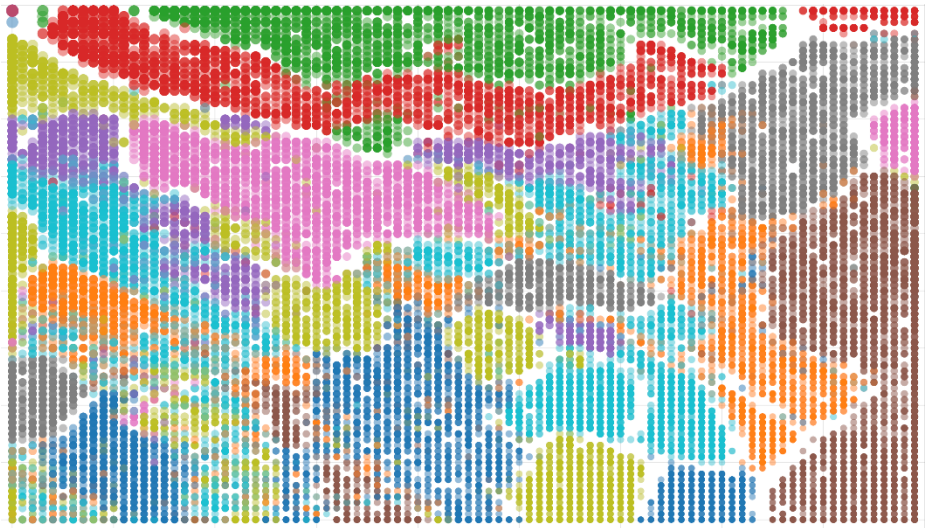
3.4 Exploration and annotation process

Participants were told to take the role of an archivist with the task of exploring large amounts of audio recordings. They did not receive any details on the data, only that it could be anything recorded at any time, hence the nature of the recordings was not known to them. Their instructions were to interact and explore the plot and see if they could find different types of regions based on what type of sound they heard. They were then asked to give the centroid coordinates and labels for the three most distinctly sounding regions they found. The participants had to assign labels without any guidance. This was done so that we would not influence the participants choice of labels. The top two labels (based on frequency) were used for further analysis.

3.5 Analysis

For each coordinate pair and each label, selected as described above, the Euclidian distance to each point - each audio segment - on the map was calculated. The audio was then temporally reassembled, and smoothed graphs showing the relative closeness to a label centroid for each segment in time were constructed. This was done by adding each inverted distance to a Kernel Density Estimate (KDE) for each label and speech. The KDEs were then sampled, and the value of the two top ranked labels deducted from each other at each point in time. The result is a rough, smoothed estimate of which of the two labels any given segment belongs to.

Fig. 2. Each colour represents the speech of one president. Note that this information was not presented to the participants, nor was it provided as input for the clustering. Blue=Obama; orange=Ford; green=Johnson; red=Nixon; purple=Carter; brown=Trump; pink=Reagan; grey=Kennedy; yellow=Clinton; cyan=Bush.



For further validation one person (not participating in the previously mentioned annotation process) performed a crude manual labelling of 5 minutes from each speech (10:00 - 15:00). Anything that was applause was labelled as such, and anything else was labelled as speech. This means that silent segments, for example, were labelled as speech. Temporal granularity was limited to 100 ms, identical to the automatic temporal disassembly rate used in the preprocessing for the experiment. The resulting number of segments was 3000 in each recording. The manual labels provided by the expert and provided by the participant in each recording were compared. Albeit a crude comparison, it reflects the purpose of the method: to get a rough idea of the general contents of unknown audio quickly.

4 Results

4.1 Salient areas

The participants selected coordinates spread out over the entire map (Figure 1). Without asking them for labels, it would be hard to find any patterns or clusters in their results. With the labels provided by the participants, however, a clear pattern appears.

4.2 Labels of disassembled audio

Inspection of the labels given by our informants revealed that 2 labels had been supplied on 3 occasions, 1 label on 2 occasions, and the remaining 13 labels only once each (see Table 2). The labels with 3 mentions were "applause" and "speech", and we contrast these in our analysis. The label mentioned twice was "silence". Among the remaining labels with a solitary mention, many seem to point to something very similar to "speech" or "applause", but for our purposes, the two highest ranking labels will be sufficient, so we leave the rest out of the discussion. The spatial distribution of the labels is shown in Fig. 1. Fig. 2 shows the same data, but now colour coded according to what audio it's originating from.

Table 2. Labels proposed by the participants. The first column holds considers the count of the label in the second column. Only the most commonly occurring labels, applause and speech, were used for further analysis

#	**Label**
3	applause, speech
2	silence
1	sports commentary, human speech, news segment, nothing much going on in big hall, in street, radio talk - tv commercial, people speaking, low pitch environmental sounds, in plane, background noise, synthesized speech, high pitch high volume environment, some news or political report

Table 3. Overlap between automatic annotation and manual annotation. Accuracy: 76.33%; Prec: 58.60%; Rec: 70.95% (applause)

		Manual annotation	
		Speech	Applause
Automatic segmentation	Speech	**7636**	5394
	Applause	3127	**19843**

4.3 Segmentation of reassembled audio

Fig. 3 shows spectrograms of minutes 10 to 25 from each of the speeches, with an overlay of the collected data. The orange graph shows areas that are likely "applause", whereas the purple graph shows areas more likely to be "speech", according to our method. In other words, orange peaks should be more likely to contain applause. At face value, it seems that the purple overlay is correlating with lighter, more speech-like spectrogram sections, and the orange with darker areas that resembles white noise more than speech (as applause would).

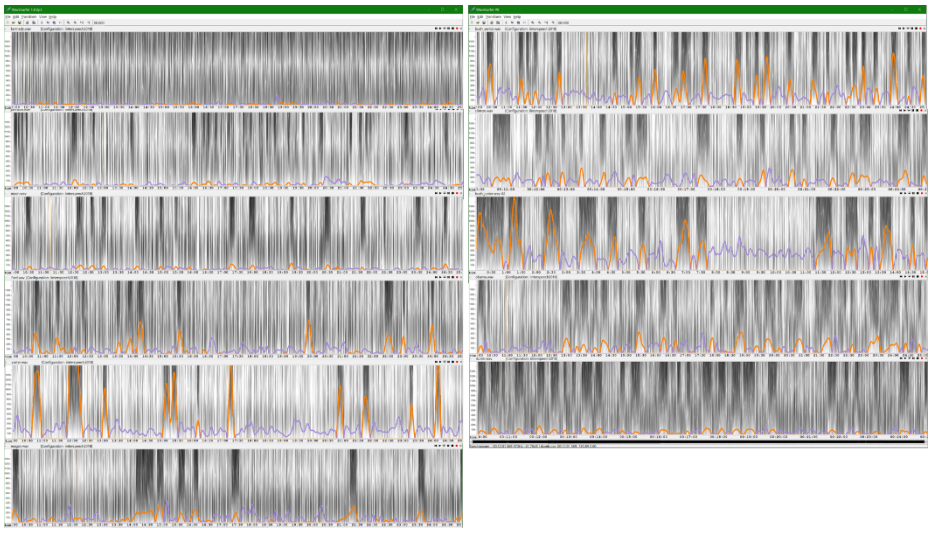


Fig. 3. A spectrogram sample (10:00 - 25:00) from each president. The orange and purple line represents how close the current timestamp is to the previously collected applause and speech centroids respectively. The darker areas are, as a rule, applause. The speeches are, top to bottom, left to right Kennedy, Johnson, Nixon, Ford, Carter, Reagan, Bush Sr, Clinton, Bush Jr, Obama, Trump.

4.4 Validation

The seemingly good match between visual areas in the spectrogram and the automatic segmentation is corroborated by results presented in Table 3. The table shows how well the automatic annotation compares to the manual annotation of 5*11 minutes of the data.

5 Discussion

The experiment we presented provides good evidence that the combination of disassembling audio temporally, reorganizing according to a similarity measure, and displaying interactively to people using massively multi-component audio replay does provide a window into recordings that would otherwise remain unexplored. Our participants spent only a few minutes on the task - 10 at most. The participant that took the longest later stated that "it took much longer because I found listening to the sounds intriguing". Out of 21 labels collected, 2 groups of 3 identical words were found. A number of similar words could have been clustered with these easily.

The validation of segmentation/categorization results showed that we beat the majority class. That is hardly a feat, however. There are obvious optimizations that could be made, such as selecting a better threshold for "speech"/"applause" (the optimal threshold on this data achieves 80+% precision at 40+% recall) or discarding the poorest quality audio sources. But good results on a specific, known classification task is

not what we look to achieve here. Rather, the key is that the system did not know what to look for, nor did the participants providing the labels. We believe that this is a good starting point when it comes to investigating large quantities of completely unknown found audio data.

6 Future work

We pursue this line of inquiry along with optimizing usability and improving robustness. As to the former, as it stands, the framework presented here involves a preliminary experimental setup that does not offer full comfort of use to an actual investigator of archives. In parallel with the work presented here, which is aimed at validation and development, we place considerable effort into creating a freely available software resource for anyone interested in employing these techniques.

Regarding robustness, the more generalizable framework we aim for, with added functionality and better facilities for optimization, will help exploit the hidden resources found in archives and freely available collections.

On the technological side, this goes hand in hand with exploring alternative methods for dimensionality reduction, such as auto-encoders, and for feature extraction - conventional spectrograms are clearly just one of many methods to represent sound.

7 Acknowledgements

The work 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).

References

[1] 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)*, 2018.

[2] A. Sangwan, L. Kaushik, C. Yu, J. H. L. Hansen, and D. W. Oard, "'Houston, We have a solution' : Using NASA Apollo program to advance speech and language processing technology," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 1135–1139.

[3] J. Berg *et al.*, "TillTal – making cultural heritage accessible for speech research," in *CLARIN Annual Conference*, 2016.

[4] J. Nouza *et al.*, "Making Czech historical Radio archive accessible and searchable for wide public," *J. Multimed.*, vol. 7, no. 2, pp. 159–169, 2012.

[5] J. H. L. Hansen *et al.*, "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, 2005.

[6] C. Barras, A. Allauzen, L. Lamel, and J. Gauvain, "Transcribing audio-video archives," *Language (Baltim).*, vol. 1, pp. 1–13, 2002.

[7] J. Psutka *et al.*, "Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments," in *Proceedings of TSD 2002*, 2002, pp. 253–260.

[8] R. Ordelman, F. De Jong, and W. Heeren, "Exploration of Audiovisual Heritage Using Audio Indexing Technology," in *In Proceedings of the 1st ECAI Workshop on Intelligent Technologies for Cultural Heritage Exploitation*, 2006, pp. 36–39.

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

[10] J. Edlund, J. Gustafson, and J. Beskow, "Cocktail--a demonstration of massively multi-component audio environments for illustration and analysis," in *SLTC 2010*, 2010.