INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Human-in-the-Loop Efficiency Analysis for Binary Classification in Edyson

Per Fallgren, Jens Edlund

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

Abstract

Edyson is a human-in-the-loop (HITL) tool for browsing and annotating large amounts of audio data quickly. It builds on temporally disassembled audio and massively multi-component audio environments to overcome the cumbersome time constraints that come with linear exploration of large audio data. This study adds the following contributions to Edyson: 1) We add the new use case of HITL binary classification by sample; 2) We explore the new domain oceanic hydrophone recordings with whale song, along with speech activity detection in noisy audio; 3) We propose a repeatable method of analysing the efficiency of HITL in Edyson for binary classification, specifically designed to measure the return on human time spent in a given domain. We exemplify this method on two domains, and show that for a manageable initial cost in terms of HITL, it does differentiate between suitable and unsuitable domains for our new use case - a valuable insight when working with large collections of audio.

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

1. Introduction

Working with large collections of unannotated audio is challenging. Compared to other data types, such as text and images, sound is difficult to browse ease due to its transient nature. Images and words can be observed for as short or long as needed, at the whim of the observer, while sound recordings generally have to be perceived linearly over time, at or near the pace of the original sounds. Although one certainly can start, stop, fast forward, pause and replay recordings, or change the replay rate to faster or slower, there is no good way to linger on, say, a word or a phrase, nor is there any standard way to decouple listening from the temporal domain. The human-in-the-loop (HITL) audio browsing tool Edyson [1] remedies this through a blend of new and old techniques that together decouples recordings from the temporal domain and facilitates browsing of audio at paces well above real-time.

In this paper, we make three main contributions to the existing body of work around Edyson. Firstly, a new use case - HITL binary classification by sample. Note that Edyson is intended for exploratory, fast work on materials that are not well-known. In cases, where there is time and resources to train a task specific classifier, this will with all likelihood outperform the algorithmically simplistic but HITL assisted classification Edyson can provide. But Edyson shines in that it is fast [2] and can be used without training models, with limited knowledge of the data set, and with little prior knowledge on the user's part. We also believe that Edyson classification may help to bootstrap the training process for fully automatic classifiers, as it allows us to quickly acquire large albeit potentially noisy sets of training labels.

Secondly, we put Edyson to the test on a new acoustic domain: oceanic hydrophone recordings and whale song. The methods used in Edyson are still new and their strengths and weaknesses on different types of materials are still understudied, and this study provides another step in that direction.

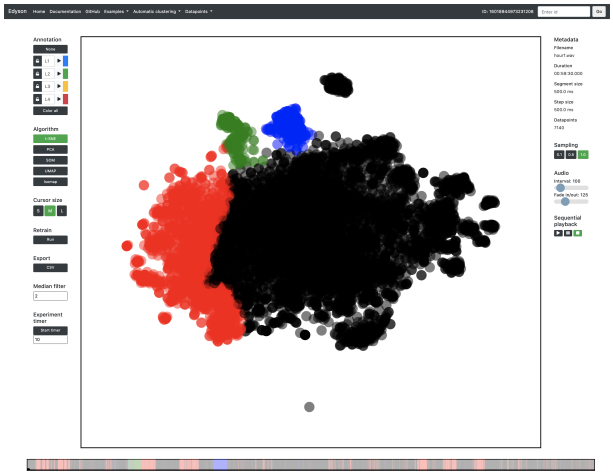


Figure 1: *An example of what the Edyson interface looks like during audio browsing. For a complete list of its functionalities see [2].*

Thirdly and most importantly, we propose a repeatable method to analyse, evaluate and visualise the efficiency of HITL in Edyson for specific binary classification tasks. As stated above, quick exploration and fast initial label acquisition are among Edyson's strengths. But we know that results will vary wildly depending on the data set and the task, so for the tool to be meaningful, we need the ability to gauge whether Edyson classification is at all meaningful in a particular domain, and if so, how much HITL time it is meaningful to spend.

2. Background and related work

Edyson[1] is a tool intended to facilitate exploratory work on large sets of unknown audio data, for example recordings found in digital archives [3]. It is developed within a project aiming to make speech archives more accessible [4]. Among its core components we find temporally disassembled audio (TDA), a method to seamlessly switch between the conventional (as far as human perception is concerned) temporally ordered view of sound and audio that has been chopped up into sound snippets that are liberated from their temporal anchors and instead arranged along other dimensions. Edyson provides a range of standard feature extractions and each feature is seen as a dimension that the sound snippets can be arranged along. In our work so far the multiple dimensions corresponding to acoustic features are projected down to two dimensions using one out

[1]Edyson is open source and available for download at github.com/perfall/Edyson.

http://dx.doi.org/10.21437/Interspeech.2021-45

of several dimensionality reduction algorithms, after which the snippets are visualised as a point cloud alongside their conventional temporally ordered visualisation.

Using the mouse pointer, it is possible to hover over different areas of this point cloud, and use the points under the pointer to create a massively multi-component audio environment (MMAE; [5]), allowing us to listen to a large number of sound snippets simultaneously. The effect is similar to the sound one perceives when attending a cocktail party.

Through MMAE, the TDA can be explored much more efficiently than linear listening to the original audio, which allows us to gain valuable insights that would otherwise be undiscovered. Furthermore the approach can also be used for rudimentary annotation tasks: the points under the pointer can be coloured, with the colours corresponding to labels. By zooming in and out and going back and forth between the temporally ordered audio and the TDA, large quantities of sound can be manually labeled in a very short time [2].

Note that comparison with machine learning frameworks for binary classification is misguided. In Edyson, binary classification is a peripheral functionality and a side effect of its exploratory functionality. Nevertheless, the tool has proven useful for several tasks, for example speech activity detection in noisy radio transmission data [2].

3. Method

Two experiments were conducted to quantify the efficiency of binary classification using samples and HITL in Edyson. In both cases, preexisting labels were added to a small portion of the data, such that a subset of the visual point cloud was coloured from the start. This primes the human and illustrates roughly how the classes are distributed. The human, a person familiar with Edyson[2], then spent time colouring (classifying) the remaining points. The results of this process was evaluated at regular intervals to gain an understanding of value of the human time spent in terms of classification accuracy.

3.1. Tasks and data

3.1.1. Task 1: Whale song

The Marinexplore and Cornell University Whale Detection Challenge [6] is an inactive Kaggle challenge that aimed to produce a whale detection algorithm distinguishing right whale up-calls from other noise in oceanic hydrophone recordings. 30,000 training samples were provided with a set of binary labels. Every sample is two seconds long, with 2KHz sample rate containing either a whale call (label=1) or not (label=0). Three independent hours were used in this experiment. 5,400 (3 x 1800) samples were selected according to an even distribution to avoid any potential labelling bias. 300 (3 x 100) additional files served as priming labels for the HITL.

3.1.2. Task 2: Speech activity

AVA-Speech is a publicly released dataset by Google with dense labels for speech activity in a collection of YouTube videos [7]. Sourced from the original AVA video dataset [8], it is audio-visual and tagged with 4 unique labels, namely CleanSpeech, Speech+Noise, Speech+Music and NoSpeech. For the purposes of this paper all speech labels were collapsed to Speech and NoSpeech was left intact. The dataset originally contained 15 minute segments extracted from 192 YouTube videos. Some videos have been taken down or made inaccessible for other reasons. Labels for 160 videos exist at the time of writing; 119 were attainable. This resulted in 29 hours and 45 minutes of material, of which 10% or approximately 3 hours (3 x 59:30) were used - 90 seconds from every video.

3.2. Process

3.2.1. Task 1: Whale song

Each of the whale song subsets was processed in a separate session using three different snippet durations for the temporal disassembly: 500ms (=8,000 datapoints), 1s (=4,000 datapoints), and 2s (=2,000 datapoints). Mel-frequency cepstral coefficients[3] (MFCCs) were then used as features that were projected onto 2 dimensions. Pilot testing also included sessions with linear spectrograms as features. These are omitted as they yielded poorer performance. 10% of the datapoints were coloured from the start, the rest of the datapoints were black and unknown. Several projection algorithms were available to dynamically change between, however tSNE [10] seemed to work best for this task. Five minutes and 30 seconds were spent on the first and second session, and 4 minutes on the third.

3.2.2. Task 2: Speech activity

The process was again divided into three sessions, each containing one hour of material. Temporal disassembly was set to 500ms snippets (7,140 datapoints), mel-frequency cepstral coefficients (MFCCs) were again used as features. Six minutes of the audio, or 714 evenly distributed snippets, were used for priming. The self-organizing maps (SOM) [11, 12] view was predominately used. Six minutes HITL time was spent for each of the first two sessions and seven minutes and 30 seconds for the third.

Edyson was set to generate an output file every 30 seconds for to allow us to follow how results progressed over time.

3.3. Analysis

For both tasks, all output files except the last one contained unknown labels - that is uncoloured datapoints. A simple centre of gravity (CoG) approach was used to automatically colour these: the euclidean distance in the 2D-space from each datapoint to the CoG of the prime labels, using tSNE for the task 1 and SOM for task 2. The datapoint was assigned a normalized continous value based on its proximity to both CoGs, a value close to 1 is near the true label CoG and vice versa. Since Task 1 used 0.5s audio snippets for the first session, every two-second long sample had four values. This was resolved by using the max value of the segment, and the same method was used in the second session (with 1s snippets). Task 2 contains data where speech segments are expected to be longer sequences. To utilize this expectation, we included an alternative analysis where a temporal median filter of three snippets was applied.

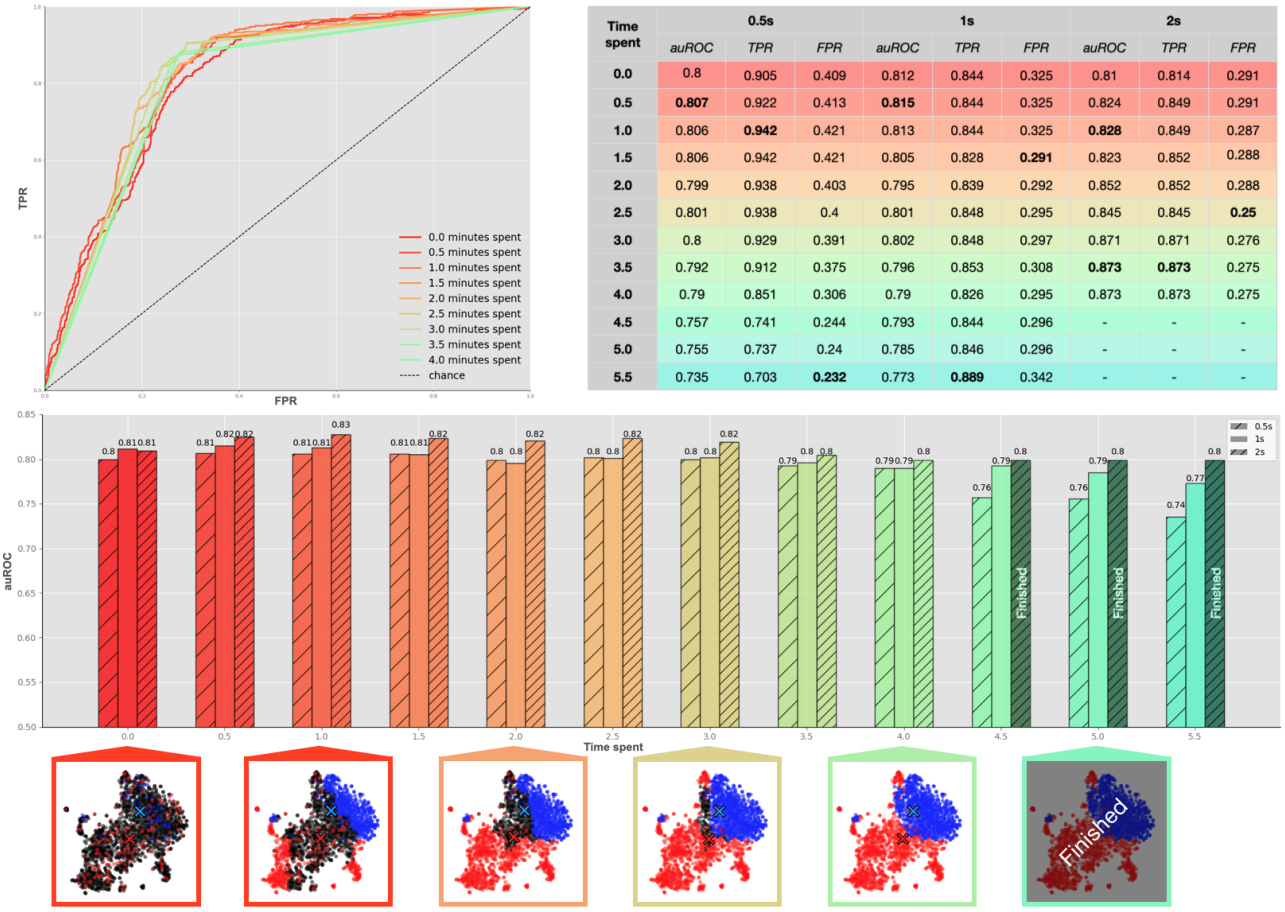
Receiver Operating Characteristic (ROC) curves were plotted and area under ROC (auROC) scores calculated for all outputs. True positive rates (TPR) and False positives (FPR) were also included.

4. Results

Figure 2 shows all metrics for task 1, including tabular data. Figure 3 shows the same information for task 2.

[2]The first author of this paper was the HITL for these experiments.

[3]Default MFCC settings from OpenSmile [9].



Time spent	0.5s auROC	0.5s TPR	0.5s FPR	1s auROC	1s TPR	1s FPR	2s auROC	2s TPR	2s FPR
0.0	0.8	0.905	0.409	0.812	0.844	0.325	0.81	0.814	0.291
0.5	**0.807**	0.922	0.413	**0.815**	0.844	0.325	0.824	0.849	0.291
1.0	0.806	**0.942**	0.421	0.813	0.844	0.325	**0.828**	0.849	0.287
1.5	0.806	0.942	0.421	0.805	0.828	**0.291**	0.823	0.852	0.288
2.0	0.799	0.938	0.403	0.795	0.839	0.292	0.852	0.852	0.288
2.5	0.801	0.938	0.4	0.801	0.848	0.295	0.845	0.845	**0.25**
3.0	0.8	0.929	0.391	0.802	0.848	0.297	0.871	0.871	0.276
3.5	0.792	0.912	0.375	0.796	0.853	0.308	**0.873**	**0.873**	0.275
4.0	0.79	0.851	0.306	0.79	0.826	0.295	0.873	0.873	0.275
4.5	0.757	0.741	0.244	0.793	0.844	0.296	-	-	-
5.0	0.755	0.737	0.24	0.785	0.846	0.296	-	-	-
5.5	0.735	0.703	**0.232**	0.773	**0.889**	0.342	-	-	-

Figure 2: *Analysis results from task 1; The top-left pane shows ROC curves for every output from start to finish for the 2s session; the top-right pane holds tabular scores for auROC, TPR and FPR (rows correspond to HITL time spent, best results per column are bold); centre bar chart shows the auROC over HITL time for all three sessions (0.5s, 1s and 2s); and the bottom row shows Edyson interface screenshots at matching HITL times (with a step duration of 60s) for the 2s session (CoGs are included as crosses). For support, everything is colour-coded in a gradient from red (start of classification) to blue (end) This information is also represented in the order of the elements.*

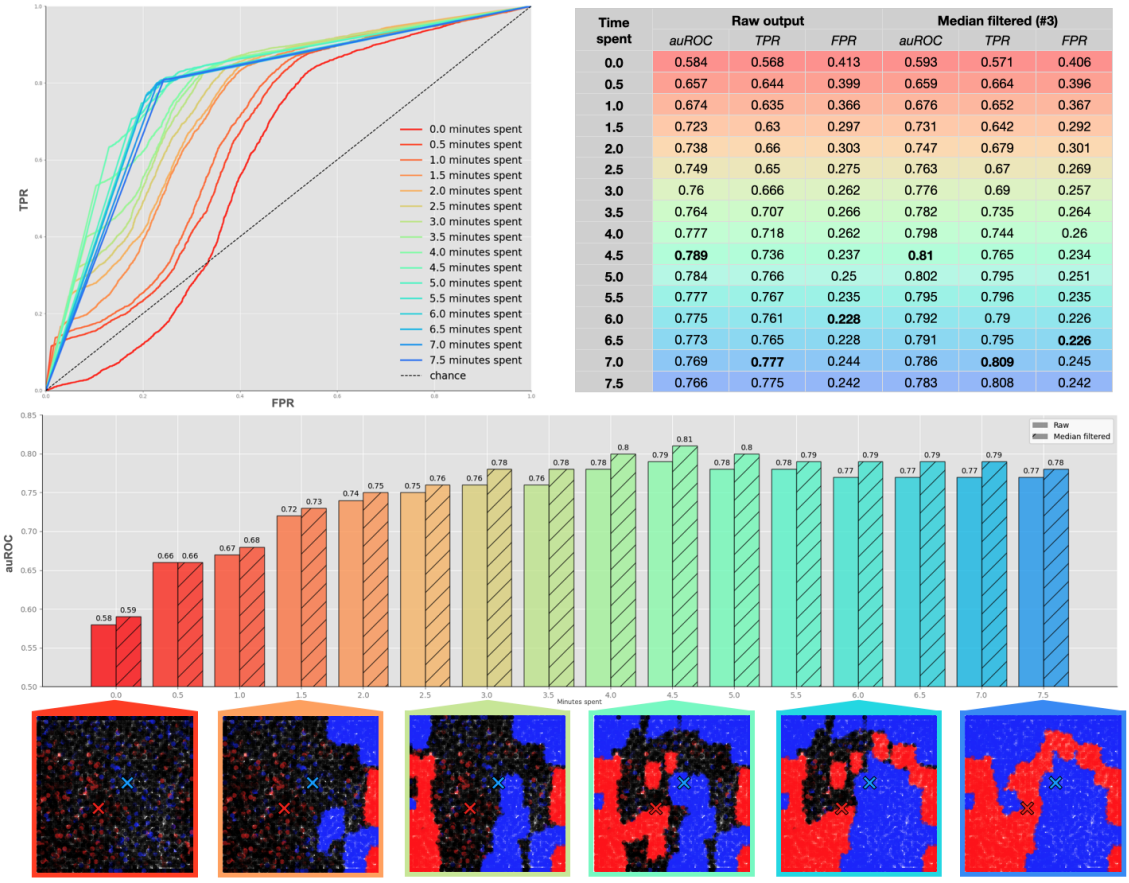
5. Discussion

An important thing to note is that these experiments assume a scenario where a labelled subset of data is available - our efficiency analysis expects access to ground truth.

The ROC curves in Figure 2 suggest that the Edyson output does not improve over time for task 1. auROC confirms this: for all three sessions the score increases only marginally initially and declines after less than a minute. The crude method of classifying points based on proximity to priming label CoG yields similar performance. TPR behaves similarly, but for the second and thirds setups with snippet durations of 1 and 2 TPR does improve over time, peaking at the end or near end of the session. The heterogeneous distribution of colouring in the Edyson screen dumps in Figure 3 make it clear that distance to CoGs is inappropriate for this data. The ROC curve of the initial output partly falls below the dashed random classifier curve, suggesting that the CoGs do not reflect an inclination towards either of the two classes. We see that in general, spending more HITL time achieves an auROC score increase, with a peak at around 4 minutes and 30 seconds, after which we observe a slight decline. Median filtering the decisions in the time domain yields a consistent improvement for all outputs. TPR (seen in the table) compare reasonably well to the results of the three baselines included in the AVA-speech paper [7]: the publicly available voice activity detector of WebRTC [13] achieved 0.722; a small version of a then state-of-the-art convolutional neural network speech detector trained on AudioSet [14] achieved 0.810; and the larger version of the same model achieved 0.917. They report a FPR of 0.315 for all models [4]. The median filtered Edyson output reaches a TPR of 0.809 and FPR of 0.245 after 7 minutes, which outperforms the WebRTC VAD and is comparable to the smaller version of the CNN-model in TPR.

We see from the results that the efficiency of HITL supported binary classification with Edyson is highly dependent on the classification task. This is to be expected - any classification method will be more or less suited for specific tasks. Our primary purpose here, however, is to demonstrate how well-defined evaluation metrics can provide early information on whether a specific HITL task is worthwhile, and how much time it is sensible to spend on it.

[4] It seems unlikely that all models share the same FPR, although this is what is reported in the paper.



Time spent	Raw output			Median filtered (#3)		
	auROC	*TPR*	*FPR*	*auROC*	*TPR*	*FPR*
0.0	0.584	0.568	0.413	0.593	0.571	0.406
0.5	0.657	0.644	0.399	0.659	0.664	0.396
1.0	0.674	0.635	0.366	0.676	0.652	0.367
1.5	0.723	0.63	0.297	0.731	0.642	0.292
2.0	0.738	0.66	0.303	0.747	0.679	0.301
2.5	0.749	0.65	0.275	0.763	0.67	0.269
3.0	0.76	0.666	0.262	0.776	0.69	0.257
3.5	0.764	0.707	0.266	0.782	0.735	0.264
4.0	0.777	0.718	0.262	0.798	0.744	0.26
4.5	**0.789**	0.736	0.237	**0.81**	0.765	0.234
5.0	0.784	0.766	0.25	0.802	0.795	0.251
5.5	0.777	0.767	0.235	0.795	0.796	0.235
6.0	0.775	0.761	**0.228**	0.792	0.79	0.226
6.5	0.773	0.765	0.228	0.791	0.795	**0.226**
7.0	0.769	**0.777**	0.244	0.786	**0.809**	0.245
7.5	0.766	0.775	0.242	0.783	0.808	0.242

Figure 3: *Analysis results from task 2. The panes are arranged in the same way as in Figure 3. The middle pane, here, shows auROC scores for each time step for both raw and temporal median filtered outputs;*

6. Conclusion

Our proposed HITL efficiency analysis for binary classification in Edyson does well: at the cost of a small amount of labelled data and less than an hour's work per task, it shows convincingly that task 1 is ill suited for this type of classification, and task 2 is well suited. It also suggests that a HITL should spend around 4-5 minutes on each batch of task 2 data, and not more. In addition, the analysis is able to provide useful feedback on which parameters provides the best HITL efficiency - snippet duration had little effect in task 1, post processing with a temporal median filter consistently helped in task 2.

For tools that have as their primary goal to provide fast methods to work on new and unknown tasks, such as data browsing or initial labelling, it is critical to be able to evaluate whether a tool actually achieves this goal - both in general and given specific tasks. For HITL supported binary classification, we hold that our analysis results are encouraging, and the analysis method will be included as a standard method in an upcoming release of Edyson. Further, an extension of this method will be added to include multi-classification and other tasks.

In addition to the efficiency analysis, we have added a domain (whale song) and expanded another (speech activity detection) to tested areas of application for Edyson, and a task (binary classification) to the range of HITL supported tasks Edyson can be used for. We have also added functionality to prime Edyson HITL work with known labels.

It's important to note again that the goal here is not to compete with pure machine learning approaches, and while it is comforting to see that the HITL output outperforms the baseline and is equal to the smaller state-of-the-art speech activity detection model provided by [7], our interest lies in providing an alternative when such approaches can not be applied. It is also worth noting that there is a wide range of ways to improve HITL results, for example through experimenting with feature sets or adding a curation process to refine the Edyson output (work in progress), or employing a HITL who is trained in distinguishing the sounds in the target domain, to name a few.

Finally, we would like to point out that although the tool can handle significantly more data in one session, a strong motivation for these experiments was to show that it is entirely possible to perform HITL binary classification on a standard personal computer in a standard environment.

7. 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).

8. References

[1] D. House, P. Fallgren, and J. Edlund, "Speech technology for swedish: Current impact areas for applications and edyson, an innovative tool for accessing speech data," 2019.

[2] P. Fallgren, Z. Malisz, and J. Edlund, "How to annotate 100 hours in 45 minutes." in *INTERSPEECH*, 2019, pp. 341–345.

[3] J. Edlund and J. Gustafson, "Hidden Resources â Strategies to Acquire and Exploit Potential Spoken Language Resources in National Archives," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris: European Language Resources Association (ELRA), may 2016.

[4] 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.

[5] J. Edlund, J. Gustafson, and J. Beskow, "Cocktail : a demonstration of massively multi-component audio environments for illustration and analysis," 2010. [Online]. Available: http://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A465457&dswid=3538

[6] Marinexplore and C. University. (2021) The marinexplore and cornell university whale detection challenge. [Online]. Available: https://www.kaggle.com/c/whale-detection-challenge/overview

[7] S. Chaudhuri, J. Roth, D. P. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. Wilson *et al.*, "Avaspeech: A densely labeled dataset of speech activity in movies," *arXiv preprint arXiv:1808.00606*, 2018.

[8] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[9] 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.

[10] 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

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

[12] T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds., *Self-Organizing Maps*, 3rd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[13] WEBRTC. (2011) The webrtc project. [Online]. Available: https://webrtc.org/

[14] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.