INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Crowdsource-based validation of the audio cocktail as a sound browsing tool

Per Fallgren, Jens Edlund

Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden

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

Abstract

We conduct two crowdsourcing experiments designed to examine the usefulness of *audio cocktails* to quickly find out information on the contents of large audio data. Several thousand crowd workers were engaged to listen to audio cocktails with systematically varied composition. They were then asked to state either which sound out of four categories (Children, Women, Men, Orchestra) they heard the most of, or if they heard anything of a specific category at all. The results show that their responses have high reliability and provide information as to whether a specific task can be performed using audio cocktails. We also propose that the combination of crowd workers and audio cocktails can be used directly as a tool to investigate the contents of large audio data.

Index Terms: human-in-the-loop; found speech; hearing; annotation; exploration

1. Introduction

In data-driven speech-centric research, the main challenge is no longer to find data, it is to find reliable data. When it comes to the many huge data collections that are available in archives and on the Internet, even listening through the data to get a rough sense of what is in there is an almost insurmountable challenge. Audio cocktails have previously been proposed [1], a cocktail party buzz-like blend of sounds, as a technique to browse large audio data collections. Here, we propose a method in which we use crowd workers to evaluate their feasibility for a specific task before engaging in costly large-scale endeavours. As a side effect, these studies also give a hint of whether the combination of audio cocktails and crowd workers can be used directly to investigate the contents of large audio data.

2. Background

2.1. Found data

The experiments conducted here validate audio cocktails as a means to quickly acquire information about the general contents of large audio data - specifically *found data*. That is the many large collections of audio that were not collected specifically with science, and speech technology in particular, in mind. In Sweden alone, a mainly text based archive such as the Language Council of Sweden [2] has more than 20 000 hours worth of digitised recordings. Spotify recently released 50 000 hours worth of podcast episodes [3] and at the National Library of Sweden [4], the rapidly growing audiovisual collections exceeded 10 million hours several years ago. If we calculate the work effort involved in listening through these collections with full Swedish work weeks, and no holidays, it takes 10, 25 and 5000 person years, respectively. The latter is the equivalent of the entire work lives of 100 employees.

2.2. Audio cocktails

Audio cocktails are created using a combination of *Temporally Disassembled Audio* (TDA) and *Massively Multi-component Audio Environments* (MMAE). TDA [5] is the division of the audio signal into snippets of equal duration, potentially with a step length that is different than the snippet duration. In this, it is the same as any framed audio processing. The goal of TDA is partial removal of the temporal ordering of snippets, and they have the added requirement that the snippets be long enough to be perceptually meaningful (> 0.05s) and short enough to be perceived as near-instantaneous (< 1s). We use TDA to pick apart lengthy audio, rearrange it using for example a combination of feature extraction and visualisation methods, and then listen to it in order to acquire human-in-the-loop (HITL) feedback or labels. MMAE is used to build immersive listening experiences by playing large amounts of short, overlapping sounds in a continuous stream. It was originally used to create dynamic soundscapes that can be be changed on the spur of the moment [1], for example controllable applause built from a few recorded claps, ocean sound based on seagull and wave snippets, and the buzz of mingling people that is the background for the cocktail party effect [6] from a range of speech recordings. The audio cocktails used here are MMAEs based on specific selections of TDA with the purpose of providing a means to quickly browse the audio underlying the TDA. This technique has previously been used in several experiments: to test whether speech with specific voice characteristics can be found [7], and it is the main browsing method in the Edyson tool [1] where it has proved effective in quick labelling of speech vs applause [8] and speech vs non-speech [9].

2.3. Crowdsourcing and human-in-the-loop

Crowd workers have become a common resource for data-intensive science. There are numerous studies of their reliability, from the early findings of [10] that show that crowd workers results do not differ significantly from those of others, to more speech-centric studies with similar conclusions, such as [11]. At Interspeech 2011, a special session on crowdsourcing for speech processing ultimately resulted in [12], a hands-on manual that remains highly useful today. A systematic review of crowdsourcing is given by [13], who concludes that crowd workers provide "high-quality data consistent with previous speech-assessment standards in a rapid, cost-effective manner" and that the methodology "incorporates a lay perspective".

[1] github.com/perfall/Edyson

10.21437/Interspeech.2023-2473

This latter part is of particular interest to us: we are not looking for a method that calls for annotators with several years of training to make highly formalised technical distinctions. We seek a means to tap into language users' immediate and unreflected perception of speech.

2.4. Just noticeable difference

Like [14], which also addressed crowdsourcing of speech assessments, we take inspiration from perception studies based on the concept of *just noticeable difference* (JND). A common practice in such studies is to fit a sigmoid to the observed data and report model fit and the point at which 50% of the participants report that they perceive a difference. This has also been used in a speech technology context (e.g. [15]) and we apply a modified version here.

3. Method

3.1. Tasks

Two experiments were designed as methods to investigate crowdsourced perceptions of audio cocktails as a means to learn about the contents of their source audio. In the first, **HearMost**, crowd workers were told that "the attached audio file contains a blend of overlapping sounds" and that their task was to "listen and say which of the two categories you hear most of". In the second, **HearSome**, they were given the same introduction, then asked to "listen and say if you can hear any **ContrastSound**" and was told that the amount of the sound may be very small. In both experiments, crowd workers were presented with audio cocktails in which the amount of **ContrastSound** and **BackgroundSound** had been varied systematically.

3.2. Data source and selection

AudioSetCurated is a subset of AudioSet [16], the latter consists of over two millions human-labelled 10-second clips extracted from YouTube videos. Every clip corresponds to one of several hundred categories, ranging from speech and animal sounds to music and common environmental sounds. As speech is central to this study, the subset was created containing three speech categories - **Male**, **Female** and **Child**. A fourth contrastive category of **Orchestra** was also selected for comparison to the speech categories. It was mainly chosen on the basis of homogeneity (the general **Music** category would introduce too much random variation for our purposes), recognisability, and quantity.

3.3. Curation

The main goal was to propose and validate a method to quantify the usefulness of audio cocktails as a tool for discovering the contents of large audio sets. This was done by systematically varying the contents of the audio cocktails used in the experiments, which made it paramount that the source sounds were good representations of their labels. While AudioSet is human-labeled, a large proportion of the data contains a variety of undesirable elements such as silence, noise and overlapping categories. In the **Child** category a majority of the samples were mixed with adult speech or shouting. A curation step was as such added in which 100 10-second long samples from each category were extracted manually, resulting in the **AudioSet-Curated** subset.

A simple annotation tool (Figure 1) was implemented for this purpose. The tool takes a directory of audio files as input and plays them in a sequence to the annotator, who then decides whether or not the given file fulfils the requirements. By pressing space bar the file is rejected and the consecutive file is instantly played, by not doing anything the file is approved, as such all approved files has been listened to in full. To aid the annotator a spectrogram is included in the interface, as is a timer showing how many seconds are left. By rejecting files before they finish playing - either when hearing unwanted elements or when seeing noise in the spectrogram - the annotator can curate the data in significantly less time than realtime. Sound files for every category were randomly selected from AudioSet and then curated by the authors of this paper. Any noise, overlapping category, or other element that did not reflect the category label lead to a rejection, as did cumulative silence longer than half a second or so. In total, 1 543 sound files were annotated to generate 100 per category. Rejection rates were 73% for **Male**, 58% for **Female**, 87% for **Children** and 32% for **Orchestra**.

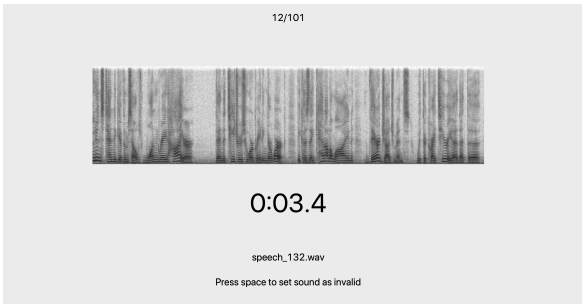

Figure 1: *Curation interface*

3.4. Stimuli generation

Systematically varied audio cocktails were created from **AudioSetCurated**. The following parameters were varied: **SnippetDuration** = {0.1s, 0.2s, 0.4s}, the duration of each snippet; **LaunchInterval**= {0.005s, 0.01s, 0.02s}, the temporal interval between the start of each snippet; and **FadeDuration**, the duration of an intensity fade (in and out) of each snippet, designed to minimise perceptual artefacts when building audio cocktails with very short snippets[2]. This sums to 9 possible different parameter setting combinations for each cocktail. Each cocktail consisted of a specific blend of **ContrastSound** and **BackgroundSound**, each of which could be either of **Children**, **Women**, **Men** and **Orchestra** (but not the same), adding another 6 possible combinations, or 12 permutations. These parameters were kept the same for both experiments. Finally, 4 different versions were created of every cocktail to eliminate ill effects due to the random selection of snippets within the cocktail.

For the **HearMost** experiment, **ContrastProportion** (the amount of snippets randomly chosen from **ContrastSound** as a proportion of all snippets, i.e. **ContrastSound**+**BackgroundSound**) was varied from 0 - 1.0 with a step size of 0.1, adding another 11 combinations. The total number of cocktails in **HearMost** was 2 376 (=9x6x4x11). For the **HearSome** experiment, the amount of **ContrastSound** was varied starting at 0.0 then 0.005 and increased exponentially up to 0.64, making for 9 settings for **ContrastProportion**. **ContrastProportion**

[2] **FadeDuration** was set to 0.25 x **SnippetDuration** for all cocktails

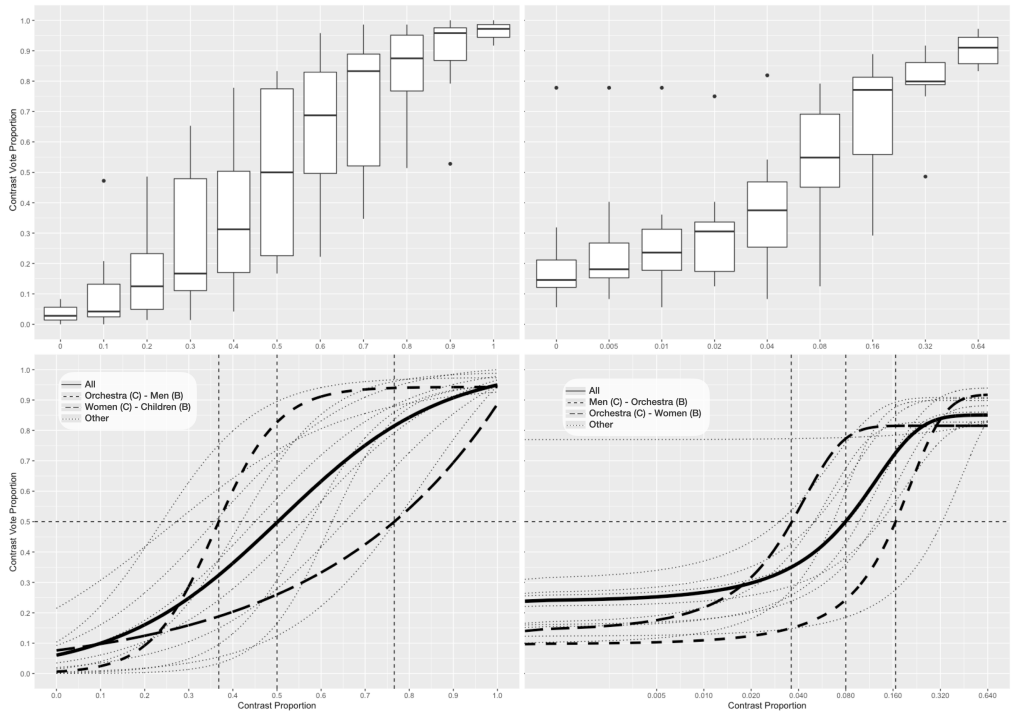

Figure 2: *The left column represents* **HearMost** *and the right* **HearSome**. *The upper row shows boxplots for the distribution of* **ContrastVoteProportion** *for each* **ContrastSound-BackgroundSound** *combination on the y-axis and* **ContrastProportion** *on the x-axis. The lower row shows the fitted sigmoids for each* **ContrastSound-BackgroundSound** *(dotted), and the fitted sigmoid for the entire data (solid). Two* **ContrastSound-BackgroundSound** *combinations are highlighted in each pane in the lower row:* **Orchestra** *contrast against* **Men** *background (dashed) and* **Children** *against* **Women** *(longdash) in the left pane;* **Men** *contrast against* **Orchestra** *background (dashed) and* **Orchestra** *against* **Women** *(longdash) in the right. Guidelines at Y=0.5 and vertical lines where the highlighted sigmoids intersect with Y=0.5 are added for clarity.*

was assymetric in this setup (**ContrastSound** = **Women** vs **BackgroundSound** = **Men** was not the inversion of **ContrastSound** = **Men** vs **BackgroundSound** = **Women**), so all pairs had to be generated, making for 12 **ContrastSound-BackgroundSound** configurations, and a total number of cocktails of 3 888 (=9x9x4x12).

3.5. Crowdsourcing participants and experiment platform

The crowdsourcing platform Mechanical Turk was used. For both experiments, age, gender and information on hearing impairments was collected, and the workers were encouraged to use headphones. Every cocktail was labeled by two unique workers.

3.6. Analysis

In this analysis, all of the 9 cocktail parameters configurations were combined, leaving 9 parameter settings x 4 variants = 36 variants of each cocktail representing a specific combination of (**ContrastSound**, **BackgroundSound** and **ContrastProportion**). The proportion of crowd workers who picked **ContrastSound** as the most prominent in the **HearMost** experiment, or heard **ContrastSound** at all in **HearSome** was calculated for each such combination. We call this proportion **ContrastVoteProportion**.

In **HearMost**, the response data was doubled and flipped for every **ContrastSound/BackgroundSound** pair, utilising the symmetric distribution of the **ContrastSound/BackgroundSound** ratio. In the coming analyses, this creates a symmetric distribution of the data around the vertical 0.5 axis, as for example the responses for **ContrastSound** = **Women**, BackgroundSound = **Men**, and **ContrastProportion** = 0.2 is the same as **ContrastSound** = **Men**, **BackgroundSound** = **Women**, and **ContrastProportion** = 0.8.

The continued analysis focused on the relationship between **ContrastProportion** (the proportion of the contract sound in the cocktail) and **ContrastVoteProportion** (the proportion of workers voting for the contrast in the same cocktail). Both experiments are perception experiments where the amount of the perceived **ContrastSound** is controlled. Thus when placing **ContrastProportion** along the x axis and **ContrastVoteProportion** along the y-axis, we expected an S-shaped distribution and a good fit for a sigmoid. Box plots and fitted sigmoides over the responses were used to test this expectation. Further, taking inspiration from the JND paradigm, the point where the sigmoid crosses 0.5 (majority vote threshold, **MVT**, representing 50% of the crowd workers) was calculated, as were the residual standard errors (**RSE**).

4. Results

4 752 responses were gathered for **HearMost** and 7 776 for **HearSome**. Figure 2 shows results for all cocktail pairs. For **HearMost**, the leftmost outlier in the boxplot represents **ContrastSound** = **Children** and **BackgroundSound** = **Women**, and the rightmost is the same data point flipped. For **HearSome**, the first five outliers from the left represent **Women** as a contrast against **Children**, and the rightmost

		Men	Women	Children	Orchestra
Men	RSE	-	0.112	0.21	0.14
	MVT	-	0.575	0.359	0.368
Women	RSE	0.113	-	0.22	0.128
	MVT	0.419	-	0.272	0.239
Children	RSE	0.209	0.214	-	0.14
	MVT	0.654	0.766	-	0.45
Orchestra	RSE	0.143	0.126	0.14	-
	MVT	0.625	0.769	0.557	-

		Men	Women	Chi ldren	Orchestra
Men	RSE	-	0.183	0.145	0.166
	MVT	-	0.082	0.066	0.05
Women	RSE	0.173	-	0.167	0.208
	MVT	0.051	-	0.03	0.036
Children	RSE	0.169	0.154	-	0.166
	MVT	0.097	*N/A*	-	0.134
Orchestra	RSE	0.168	0.163	0.173	-
	MVT	0.166	0.329	0.125	-

Table 1: *Residual standard error (**RSE**) and the majority vote threshold (**MVT**) for each contrast-background combination in* HearMost *(left pane) and* HearSome *(right pane). Rows represent* BackgroundSound *and columns* ContrastSound

is ContrastSound = Women and BackgroundSound = Orchestra with ContrastProportion = 0.32.

Table 1 summarises the fitted sigmoid models for all pairs in both experiments. Overall **RSE** and **MVT** for the whole of HearMost were 0.23 and 0.502 respectively, and for the whole of HearSome they were 0.233 and 0.08.

5. Discussion

The distributions in the upper panes of Figure 2 are reassuring, with a continuous increase in ContrastVoteProportion as ContrastProportion increases. We observe the expected S-shapes. The left pane is symmetrical around x=0.5, which is a necessary effect of the doubling and flipping of the data (but note that the data used to model individual ContrastSound-BackgroundSound pairs in the bottom row is not doubled in either experiment).

For HearMost, we note great variation in ContrastVoteProportion when ContrastProportion is close to 0.5. This is predictable if our audio cocktails are indeed composed, perceptually, in the way we intended them to be. If the blend is close to 50/50, we would expect it to be difficult to answer the question "which do you hear the most of". Closer to the edges, the interquartile range is much smaller. For HearSome, we note that our intuition to increase the proportion of contrast sound exponetially from a small starting value in order to achieve higher granularity at low values has paid off, as the distributions increase evenly over the exponetial steps.

Moving on to the fitted sigmoid curves in the lower row, we see that the dotted and dashed lines representing specific ContrastSound-BackgroundSound pairs vary considerably. Our results show that **RSE** is larger for the combined model (0.233) than for any of the individual models, suggesting that the contrast-background combination indeed matters for the distribution.

Turning to the outliers, we see that almost 50 percent of workers have reported that they mostly heard children in a cocktail blend with 90 percent women, a trend that was verified when we inspected the underlying data. From our point of view, this is an indication that our method is successful for investigating whether a particular task is suitable for audio cocktail browsing - the method clearly point out that it is useful for distinguishing between certain pairs, and less useful or even useless for others. This is exactly the type of information we need before investing huge resources into large scale audio browsing of sound archives. We propose that the fitted sigmoid curves can be used to quantify the usefulness and characteristics of particular tasks. For example, the dashed line in the left pane represents the response to an increase of Orchestra against a background of Men. The line rises sharply, and has its main increase clearly to the left of most other lines, indicating, predictably, that orchestral sound dominate over those of male voices. The flatter inclination of long-dashed line shows that Children against Women is a less straightforward decision to make, which is also clearly seen in the high corresponding **RSE** in Table 1. Similarly, the long-dashed line in the right pane shows that Orchestra against a background of Women is dominant even if it is barely present (**MVT** = 0.036) whereas the dashed line shows that it is harder to note the presence of Men against a background of Orchestra (**MVT** = 0.166).

6. Conclusion and future work

Approximations on what listeners perceive as prominent or not given different sound blends are interesting on their own. More importantly however they provide valuable insight for using the audio cocktail as a sound browsing tool. Previous findings on using audio cocktails have been promising, but there has until now been a lack of empirical data on how listeners perceive the sound of a cocktail and its contents. Substantial work has been put in the study to be able to properly reason about human perception of speech blends - the included categories were carefully selected, meticulous curation steps were added to ensure data quality, and more than twelve and a half thousand listens laid ground for the analysis.

We are using crowd workers in these perception experiments, and the reliability of our results is affected not only by the experiment itself, but by the reliability of crowd workers. In the results of this study we find high reliability among workers, providing additional ground towards the usefulness of using crowdsourcing for perception experiments. Moreoever, we have collected sufficient data to evaluate how many crowd workers are needed to reliably judge a task, although this analysis is outside the scope of the present paper. Further, we need a method to minimise the number of crowd workers needed to both decide if a particular decision is suitable for our method and to know when we have enough data. Repeated labelling, as described by [17], is a starting point for achieving this. Finally, the data will allow us to evaluate the effect of cocktail parameters on the reliability of crowd workers' responses.

We propose that the method we have investigated can be used (a) to evaluate the suitability of audio cocktails for a particular task, and (b) can be used to actually perform the task (i.e. to find out whether e.g. a speech type, or music, is present in a large audio collection).

7. Acknowledgements

This work was funded in part 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] J. Edlund, J. Gustafson, and J. Beskow, "Cocktail–a demonstration of massively multi-component audio environments for illustration and analysis," in *Proc. of SLTC 2010*, Linköping, Sweden, 2010, pp. 23–24.

[2] ISOF. (2021) Institutet för språk och folkminnen. [Online]. Available: isof.se/om-oss/kontakt/sprakradet/in-english.html

[3] A. Clifton, A. Pappu, S. Reddy, Y. Yu, J. Karlgren, B. Carterette, and R. Jones, "The spotify podcast dataset," 2020.

[4] K. Biblioteket. (2021) Kungliga biblioteket. [Online]. Available: https://www.kb.se/in-english/

[5] 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.

[6] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[7] C. Tånnander, P. Fallgren, J. Edlund, and J. Gustafson, "Spot the pleasant people! navigating the cocktail party buzz," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 4220–4224.

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

[9] P. Fallgren Z. Malisz, and J. Edlund, "How to annotate 100 hours in 45 minutes," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 341–345.

[10] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision making*, vol. 5, no. 5, pp. 411–419, 2010.

[11] T. M. Byun, P. F. Halpin, and D. Szeredi, "Online crowdsourcing for efficient rating of speech: A validation study," *Journal of communication disorders*, vol. 53, pp. 70–83, 2015.

[12] M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons, 2013.

[13] A. M. Sescleifer, C. A. Francoisse, and A. Y. Lin, "Systematic review: Online crowdsourcing to assess perceptual speech outcomes," *journal of surgical research*, vol. 232, pp. 351–364, 2018.

[14] B. Naderi and S. Möller, "Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[15] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, 2011.

[16] 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.

[17] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, "Repeated labeling using multiple noisy labelers," *Data Mining and Knowledge Discovery*, pp. 402–441, 2014.