# 400 voices in a jiffy: a verification of the Cocktail experiment platform

*Christina Tånnander[1,2], Per Fallgren[2], Joakim Gustafson[2], Jens Edlund[2]*
*[1] Swedish Agency for Accessible Media (MTM)*
*[2] KTH Speech, Music and Hearing*
christina.tannander@mtm.se, perfall@kth.se, jocke@speech.kth.se,
edlund@speech.kth.se

## Abstract

In this paper we present a baseline experiment intended to verify the Cocktail experiment platform, an interactive platform in which large amounts of sound snippets are mixed into a dynamic buzz that varies as listeners move around in a two-dimensional space. An artificial task was created for validation purposes and given to 18 respondents: to localize clusters of voices of the same gender in an otherwise mixed-gender soundscape made up of 400 voices. The results show that respondents agree on two areas in the two-dimensional space, and that these areas coincide with the areas where there were only female or male voices, respectively. We also present a method to empirically test whether results are likely to be random and find some evidence of a cognitive bias: female respondents tend to perform this task with higher precision than male respondents, although they perceive it as more difficult.

## Introduction

The Cocktail experiment platform makes it possible to listen to and make statements about a multitude of sounds, for example voices, simultaneously. One of its intended uses is annotation and evaluation of speech and voices – a time-consuming task as it often involves many respondents and many voice samples. In this paper, we report the results of a study intended to verify Cocktail as an experiment platform and demonstrate the analysis we use to interpret and validate experimental results on Cocktail. The study is an extension of a small sanity test performed earlier. The purpose is two-fold: (1) we want to verify the general, positive sanity test results; and (2) we have revised the instructions to avoid a possible artefact affecting the results of the initial sanity test. We used a pool of 18 respondents who listened to an artificially constructed soundscape built from 400 voices and gave them a task with a known ground truth. The results verify that the system works as intended and that the revised instructions remove the suspected artefact.

## Background

### The Cocktail experiment platform

The Cocktail experiment platform combines a touch screen for control with a soundscape creation method (massively multi-component audio environments; Edlund, Gustafson & Beskow, 2010 that allows us to change the soundscape dynamically with low latency. The method is inspired by studies of the cocktail party effect, which states that that important signals carry through noise with more ease than others (Moray, 1959). The soundscapes used in Cocktail are reminiscent of the buzz heard at cocktail parties.

In Cocktail, we control the general composition of the buzz by distributing sounds over a two-dimensional space in which respondents can move around while the buzz changes accordingly. The tool is intended to make the human voice evaluation process more efficient by facilitating multi-sound listening, and also to move listener focus from individual voice characteristics to more general ones. The respondent moves around in the two-dimensional space on a touchscreen (or with a mouse), while hearing a soundscape that continuously updates to reflect the sounds placed in different positions in this space. This simulates the effect of moving around in a room or a cocktail party location while listening to the buzz of the different guests. The platform is designed with voice buzz in mind but can obviously be used with all sorts of sounds.

### Cocktail experiments

We have previously used Cocktail for in a task where respondents were asked to point to the location where they found the buzz of voices most pleasant (Tånnander, Fallgren, Edlund & Gustafson, submitted, as a first attempt towards finding out what voice parameters affect voice likability. That experiment was accompanied by a sanity test to ensure that the method works for tasks where we have the ‘correct’ answer: finding voice clusters of the same gender in an otherwise mixed gender soundscape. Eight respondents took part in the experiment, four females and four males. We got significant results of the male voice cluster, but not for the female, which we ascribed to the small number of respondents. Surprisingly, we found that four out of four female respondents found the female voice cluster first, and three out of four male respondents found the male voice cluster first. The work presented here is an extended version of this sanity test and is intended to validate Cocktail and the analysis we use to interpret results.

## Method

### Cocktail settings

Before an experiment, Cocktail is loaded with sound files, which are organized in the two-dimensional room according to some principle. When the listener points to a location, a circular marker is moved to that position, and short segments from the sounds in the area covered by the marker (the uptake area) are selected at random and played repeatedly, resulting in a voice buzz. Depending on configuration, such as the length of the segments or the size of the uptake area, we can make it easier or harder to distinguish separate voices or words.

Cocktail can be configured in several ways, of which four are described here:

*Uptake area:* The radius of an inner circle, where the sounds are played louder, and an outer circle, where they are played softer. We used the areas illustrated in Figure 1, where the radius of the inner (black) circle is half of the radius of the outer (grey) area.

*Update interval.* The audio snippets can be scheduled in batches to conserve some processing cycles at the expense of some latency. How often a new set of sampled audio snippets are scheduled is governed by this setting, and here, 1000 ms was used.

*Launch interval:* The interval between firing of each individual audio snippet. Together with segment duration, this controls how many simultaneous sounds a listener hear at any given point. 20 ms was used.

*Segment duration:* The length of the sound snippets in milliseconds. 500 ms was used. With a sound snippet length of 500 ms and a launch interval of 20 ms, 25 voices are heard simultaneously at any given time.

### Sound snippets (stimuli)

The audio used consisted of 400 voices of both genders distributed evenly in the two-dimensional space, except for two areas of 5x5 voices of the same gender, as illustrated in figure 1.

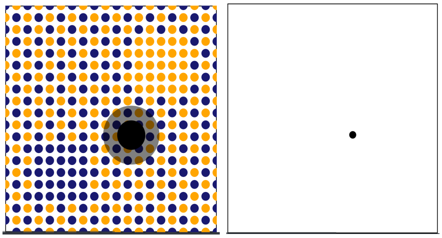

Figure 1. Left pane: Clusters of female (yellow) and male voices (blue) in development view, with the uptake area represented by two concentric circles. Right pane: Respondents' blind view of the same configuration.

The sound files are part of a speech corpus based on more than 12,000 information files, originating in the file sets of Swedish talking books (Tånnander, 2018). The files contain information about the talking book, such as information about the copyright law and header levels. The files are mp3 encoded and sampled at 22,050Hz. Under the Cocktail settings used, many voices are played simultaneously, while respondents can perhaps distinguish the occasional word. The overall effect is however a murmured voice buzz without much semantic information.

## Respondents

18 adult respondents, 9 females and 9 males, took part in the experiment. The average age was 51 years. Two of the respondents reported that they had a diagnosed hearing impairment.

## Control and systematic variation

In order to investigate the previous result, where a tendency to find one's own gender first was noted, we controlled for gender. The orientation of the two-dimensional space was varied systematically throughout the experiment.

## Procedure

The respondents were first asked to take a couple of minutes to get acquainted with the Cocktail soundscape. They were then asked to find clusters with voices of the same gender (in any order). The coordinates of the selected location were logged, as well as the locations the respondents had stopped by at before making their decision. Finally, the respondents were asked about their opinion of how easy they thought it was to find the single-gender locations.

## Analysis

The uptake area is registered for each respondent's choice. We take the loudness of the inner circle (louder playback) and the outer circle (softer playback) into account, so that 1 is added to the inner circle an 0.2 to the part of the outer circle that does not intersect with the inner circle. For each respondent input, these areas are summed in a manner similar to kernel density estimates (KDE), and the result can be plotted in a coordinate system or described statistically in the same manner as KDEs.

If the selections were done randomly, the distribution should be evenly plotted in the coordinate system, given that the number of respondents is large enough. But if the respondents select the areas based on a specific question, higher values will be seen at certain locations. We estimate significance in this system empirically. We find a single highest value of the experiment under investigation using simulated, random responses as follows: (1) for each response in the real experiment, generate a simulated response at random coordinates in the same coordinate system, (2) calculate the sum of their uptake areas, (3) sample the results at the same interval as you will sample the original experiment, and (4) find the highest value in the set of samples. We then repeat this for some large number of iterations and store each max value. Finally, sort the max values and find the Nth percentile – for example 95th for an estimate of the threshold at which the chance of a max value occurring by coincidence is less than 5% or the 99th for a less than 1% estimate. Here, we use the 99.5th percentile for an estimation of significance at the 0.005 level.

### Comparison to gold standard

This experiment is intended to validate the platform and method, and we use controlled stimuli where the "correct" answers are known beforehand. We deem a selection made by a respondent to be accurate when the centre of the selected location is within the actual voice cluster of the gender reported to be localized, and as false when the centre does not overlap with the gender cluster.

## Results

### Agreement between respondents

The results show that as a group, the respondents agreed on two areas. Figure 2 shows two response clusters, one in the upper right corner (coinciding with the female voices) and in the lower left corner (coinciding with the male voices). The green-coloured areas represent values that are less than 0.5% likely to be the result of random choice.

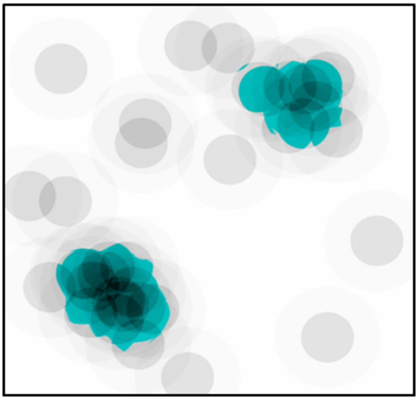

Figure 2. The areas all respondents identified as consisting of only female or male voices. Values above the significance estimate at 0.005 are coloured.

Figure 3 shows the results after splitting them by gender. The results of the female respondents are shown to the left, and male respondents to the right.

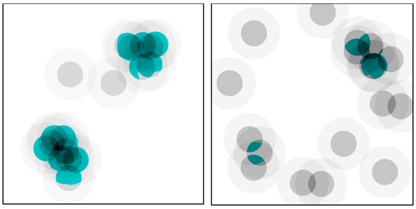

Figure 3. The results by female (left pane) and male (right pane) respondents. Values above the significance estimate at 0.005 are coloured.

### Agreement with gold standard

The two areas spotted correspond well with the actual locations of the 5x5 voices of a single gender. As individuals, there was some variation in the respondents' ability to find the places with female and male voices in the otherwise gender mixed sound environment.

Table 1 shows which gender the respondents reported they found first in the soundscape. Note that all reported locations are presented in the table, not only the correctly spotted voice clusters.

Table 1. The gender clusters that the respondents reported having found first.

| Respondent's gender | Reported first | |
|---|---|---|
| | Female | Male |
| Female | 3 | 6 |
| Male | 5 | 4 |
| Sum | 8 | 10 |

In table 2, we can see the accurately spotted voice clusters by female and male respondents.

Table 2. The columns show the number of respondents that correctly spotted (F) the female voice cluster; (M) the male voice cluster; (1) at least one voice cluster; and (2) both voice clusters.

| Resp's gender | Correctly spotted | | | |
|---|---|---|---|---|
| | F | M | 1 | 2 |
| F | 7 | 8 | 9 | 6 |
| M | 3 | 5 | 6 | 2 |
| Sum | 10 | 13 | 15 | 8 |

### Task difficulty

On the task difficulty scale from 1 to 5, where 1 is very easy and 5 very difficult, the average was 3.83. The average for the female respondents was 4.00 and for males 3.67. Some of the respondents commented that they thought it was harder to find the cluster of female voices.

## Discussion

We note that people are able to perform this straightforward but not trivial task in Cocktail. Three respondents failed to accurately pinpoint any single-gender area. Two of these reported that they had a hearing loss.

The tendency towards finding one's own gender first, as seen in a previous experiment, did not hold. Instead, only three of the nine female respondents (33%) reported that they had found the female voice cluster first, and four of the nine male respondents (45%) reported that they had found the male voice cluster first. These numbers are well within random choice.

We found a possible cognitive bias in the results. Although the female respondents generally thought the task was more difficult than the male respondents, they were better at localizing the single-gender areas. As shown in table 2 and visualized in figure 3, women pointed out areas that corresponded to the intended single-gender area 15 of 18 times, resulting in an error rate of 16.67%. The corresponding number for men is 8 of 18 times, which gives an error rate of 55.5% (whereof 67.5% for female voices and 45% for male voices). Even if we eliminate the two respondents with a hearing loss, the men’s error rate is 33%, twice the result of the females’ error rate. We have not crunched these numbers for significance.

## Conclusions

In summary, the system allows us to rapidly listen through large numbers of voices while retaining the ability to make accurate statements of what they hear – our respondents were able to spot the single-gender areas in the voice buzz of 400 voices in a matter of minutes. The previously observed phenomenon that respondents found their own gender cluster first did not stand up to scrutiny, which we think is for the best.

## Acknowledgements

This work is funded in part by Vinnova (2018-02427), the Swedish Governmental Agency for Innovation Systems and 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).

## References

Edlund, J., Gustafson, J., & Beskow, J. (2010). Cocktail - a demonstration of massively multi-component audio environments for illustration and analysis. In *The Third Swedish Language Technology Conference (SLTC 2010)* (pp. 23–24). Linköping.

Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60.

Tånnander, C. (2018). Speech Synthesis and evaluation at MTM. In *Proceedings of Fonetik* (pp. 75–80). Gothenburg: Gothenburg University.

Tånnander, C., Fallgren, P., Edlund, J., & Gustafsson, J. (submitted). Spot the pleasant people! Navigating the cocktail party buzz. In *Interspeech 2019*. Graz, Austria.