

# Emotion Detection in Commercial Applications:

Multimodal remote research combining face, respiration, voice and sentiment analysis

## Authors and Affiliations

Divya Seernani<sup>1</sup>, Anna Derington<sup>2</sup>, Jessica Justinussen<sup>1</sup>

1. iMotions A/S

2. audEERING®

# Table of Content

<b>Introduction</b>	<b>4</b>
<b>About the Study</b>	<b>5</b>
Participants	5
Study Design	5
Baseline Condition	5
Watching Ads	5
Participant Reflections	6
<b>Metrics and Methods</b>	<b>7</b>
Watching Ads	7
Participant Reflections	7
<b>Guidelines for Study Design</b>	<b>8</b>
Collect quality baseline data	8
Why do I need a baseline?	8
Baseline Best Practice	9
Baseline Duration and Timing	9
Collect Voice and Respiration data separately	10
Include Technical Checks	10
Head Check with Web Camera	10
Microphone and Audio Check	10
<b>Guidelines for Data Collection</b>	<b>11</b>
Obtaining Participant Consent	11
Participant Recruitment and Exclusion Criteria	11

<b>Data Analysis</b>	<b>12</b>
Signal Processing	12
Voice-Valence	12
FEA-Adaptive Valence	12
Respiration-specific exclusions	12
Results: How to get insights from your data	13
Baseline Analysis	13
Watching Ads	13
Participant Reflections	14
Why is this useful?	15
<b>Guidelines for Data Analysis</b>	<b>19</b>
Statistics Strategies	19
Mixed models to study effects	19
Controlling for individual differences on a segment level	20
Data Visualization: Circumplexes	20
Comparing Circumplexes	20
Circumplex Results: Experience Questions (Like and Dislike)	22
Circumplex Results: Intention Questions (Playing and Buying)	22
<b>Conclusions and Takeaways</b>	<b>23</b>

# Introduction

The purpose of this report is to give researchers best practice guidelines for online research using voice analysis in combination with other metrics such as facial expression analysis and respiration.

In this report, you will find:

Guidelines for:

- **Study design, data collection, and data analysis.**
- **How to conduct multimodal online research** for commercial applications such as advertisement, package, shelf, and product testing with iMotions' remote data collection.
- **Give examples of evaluating emotions using voice analysis, sentiment analysis, webcam respiration, and facial expression analysis** from a single remote data collection study. This example is used to explore the guidelines.

# About the Study



## Participants

Sixty participants were recruited through Prolific. Participants were 18-35 year olds (mean age 28 +/- 4) from the US, UK, and Ireland. Participants were sent a link to participate in this study from their own computers, using their own web camera and microphone.

## Study Design

This study began with some technical checks to ensure that the sensors (web camera and microphone) were properly positioned and functioning.

## Baseline Condition

First, there was a baseline condition. In this study, participants were asked to read the following text aloud in a neutral tone while their face and voice were recorded.

*The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.*

## Watching Ads

Then, participants watched advertisements for the SIM re-release and Grand Theft Auto VI. The order of the ads was randomized. While participants were watching the ad, respiration metrics were recorded.



# Participant Reflections

After seeing each ad, participants were asked to respond to four questions aloud.

- What did you like about the ad?
- What did you dislike about the ad?
- Would you play the game?
- Would you buy the game?

The first two questions gauged their experience with the ads and the last two questions gauged their intentions. While answering these questions, participants' faces and voice were recorded. We used the baseline face and voice data for normalizing data from the participant reflections.

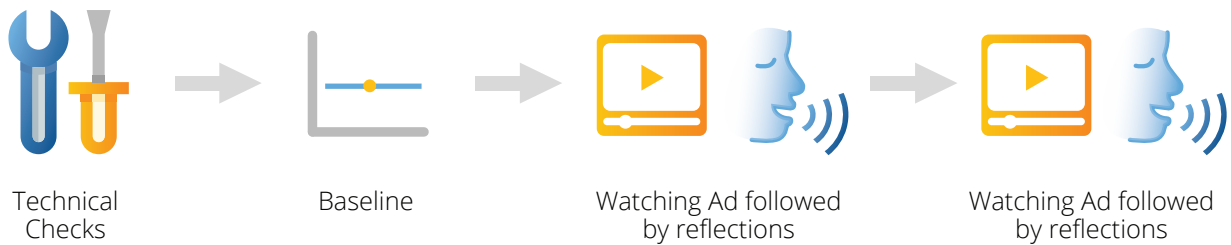


Figure 2. Flowchart of Study Stimuli. Note that this study had a randomized design so some participants saw the Grand Theft Auto Ad first and others saw the SIMs ad first.

# Metrics and Methods

# 2

## Watching Ads

**Respiration rate** provides insight into the intensity of participants' emotional responses. In this study, respiration rate was extracted from webcam video using the iMotions Webcam Respiration module. For each participant, the module analyses subtle movements associated with inhalation and exhalation. Signal processing in iMotions is then used to compute respiration rate and cycle duration. These metrics were used to assess how arousing the two ads were for the sample.

## Participant Reflections

**Sentiment analysis** (from Speech-to-text) was measured during voice activation during questions following advertisements. In this technical report, sentiment analysis specifically refers to the words being said while people are speaking, regardless of how they are being said.

- The duration of positive and negative sentiment can be taken per participant per stimulus.
- To normalize this, take the duration of the positive or negative sentiments and report them as a percentage of the total duration of the prompt.

**FEA-Adaptive Valence** (Facial Expression Analysis, FEA): We used Adaptive valence rather than valence, as it is better suited for analyzing facial expressions while talking. This metric accounts for muscle movements during speaking and adjusts the facial expression valence scores accordingly. Percentage of positive and negative adaptive valence was considered per participant per stimulus, then the baseline was subtracted from this to normalize the data.

**Voice-Valence:** We considered aggregate-level data analysis, correcting for baseline. However, aggregating across voice segments loses the variability needed for capturing emotional reaction per speech segment.

In a second, more promising analysis, mixed models were used. For this analysis, the baseline valence value per participant was subtracted from each individual speech segment of that participant. The speech segments were used in the mixed model with participants as a control variable, to compare per question across stimuli. The mixed models proved to be a better fit to leverage voice analysis and are discussed in detail in the Data Analysis section.

# Guidelines for Study Design

# 3

In this section, we will discuss:

- **How to collect quality baseline data and why this is important to your analysis**
- **Why you should collect voice and respiration data separately**
- **Which technical checks are important to include**

## Collect quality baseline data

Baseline data is useful in situations where we expect a lot of individual variation. It allows us to normalize individuals' responses during the test condition (answering questions about the ads) to their own baseline (reading the text).

This means **rather than looking at how frequently participants showed positive valence, we look at changes in frequency of positive valence** (between the baseline condition and when we asked participants about the ads).

## Why do I need a baseline?

In our experience with remote data collection for voice, face, and respiration, this is exceptionally important. In the case of voice analysis, as is the focus here, if a person is speaking loudly and quickly when we ask them about their buying intentions, it is useful to know if they usually speak loudly and quickly. **People have unique voices and vary in how expressive and dynamic their voices are. Understanding each individual's neutral tone helps us better understand their expressive tones and account for individual variability.**

**This is also true for facial expressions; there can be considerable individual variation.** If we have two participants smiling throughout the SIMs ad, **it is helpful to know if they were also smiling in a fairly neutral scenario.** If we are interested in the facial expressions when people are watching the video, we can add a neutral video (passive stimulus) to act as a baseline stimulus. If we are more interested in people replying to the questions (active stimuli), we can use the same baseline condition we used for voice as a baseline for FEA-adaptive valence. A short discussion on why we used adaptive valence rather than valence for FEA is included in the metrics section above.

## Baseline Best Practice

### *Baseline Stimuli*

In this study, the text used for the voice baseline is **Aesop's fable, 'The North Wind and the Sun'**. This text covers many of the different sounds (phonemes) in the English language and is commonly used as a baseline for voice analysis. It has been translated into other languages and works well to capture phonemes in other languages, but not all. If your study is not in English, it is important to determine what a good baseline text is for the language your participants will be speaking.

*Note: This baseline text is not a prosody or vocal emotion control. Asking the participants to say the text in a neutral tone is the control for vocal expression and emotion.*

### Baseline Duration and Timing

For voice analysis, it is ideal to have a baseline recording of **at least 30 seconds speaking**. When building the study, researchers can design the study so that participants can choose when to advance to the next part of the study and/or different parts of the study have a designated duration.

Consider the stimulus settings so that participants are speaking for 30 seconds. Often, the stimulus will start and participants will take time to read instructions or reflect, so they will not start speaking immediately. **In this study, participants took between 2-6 seconds to start talking** when asked to answer one of the four questions after viewing the ads. For baseline measurement, it took participants longer to start speaking (10-15 seconds). This is likely because the instructions were longer and the baseline occurred earlier in the study (before the participants were familiar with how the study would progress). This could be changed by presenting the instructions in a separate slide before the baseline.

Considering study design and aligning it with research needs is important. **We recommend either keeping the instructions short and anticipating a 2-6 second delay before participants begin to speak OR expect longer delays if you have longer instructions on the same stimulus.**

### *Optional Baseline for Watching Ads*

This study did not include a baseline for respiration or facial expression while viewing a passive stimulus, but you can include one. Present a neutral stimulus (emotionally neutral screensavers or a focus crosshairs image) for about 30 seconds.

# Collect Voice and Respiration data separately

**Webcam respiration measurements are affected by talking**, so it is best to separate measuring respiration from measuring voice-related metrics. For example, in this study, we measured respiration while participants were watching the ads and measured voice when they were answering questions aloud afterwards.

## Include Technical Checks

### Head Check with Web Camera

For facial expression analysis and webcam-respiration, **it is important to have a head check**. This looks at the position of the head both to see that the participant is upright, that they have an optimal distance from the camera and the face is well lit for webcam based metrics to be calculated.

### Microphone and Audio Check

**For voice analysis, it is important to have a microphone check**. Because this study presented videos, we included an audio check to ensure that participants could hear the advertisements.

# Guidelines for Data Collection

## Obtaining Participant Consent

iMotions has a built-in consent form for using data from the webcam and microphone in multiple languages. However, we recommend telling people about the webcam and microphone during the recruitment process. Otherwise, around 50% may drop out when asked to consent to being recorded.

## Participant Recruitment and Exclusion Criteria

Participants who did not comply with the instructions were excluded from this study. We recommend overrecruiting by 20-30% to be sure that you have sufficient data for your study.

In this study, 14 participants were excluded for not complying with the online instructions, not reading the text, and/or not replying to the questions prompted. In 2 cases, participants were excluded because speech was detected in the background (e.g. voice data was recorded from a virtual meeting running parallel to the study and the participant did not answer questions).

**The data included in this report are from the remaining 46 participants.**

# Data Analysis

# 5

## Signal Processing

In iMotions, R Notebooks are used to process data and signals from sensors to produce metrics. This section describes the R Notebooks used to produce some of the metrics used in this study.

### Voice-Valence

In iMotions, voice analysis is available for two different emotional models: the Emotion Categorical Model and the Emotion Dimension model. In this study, we used the emotion dimension model. The emotional dimensional model describes speech segments in terms of activation, valence, and dominance. In this study, we focused on valence. Negative valence includes emotions such as anger, sadness, fear, grief, and boredom, while positive valence includes relaxation, contentment, happiness, and excitement.

Later, we will discuss a data visualization called circumplexes, which use two of the emotion dimensions, valence and activation, to visualize approach/avoidance and arousal.

### FEA-Adaptive Valence

The R Notebook for FEA allows for thresholding the signal, and classifying moments where the signal, which is a probability graph, crossed the threshold set or not. This step is carried out so we do not take all the variance shown in the signal, but only take the data considered meaningful by the researchers. The threshold set for this study was 50 for all expressions and emotions. Therefore, when we calculate metrics like percentage of time the emotion was shown, this refers to the percentage of time the threshold was crossed.

### Respiration-specific exclusions

The R Notebook for Respiration, there is a quality index for respiration that filters out low-quality data (we used a 0.5 quality index threshold). Because of this quality index filter, any cycles with poor quality will be excluded. If many cycles are excluded, the respiration rate may be lower than expected for a spontaneous breathing condition. For this reason, respiration rates that are 5-7 cycles per minute were excluded. They are a result of low-quality data (not slow respiration).

When making respiration-specific exclusions, you must decide whether to exclude only the low-quality recordings or exclude all data from that participant. This decision is based on your study design and which comparisons you plan to make.

## **Respiration Exclusion Example:**

*In this study, 16 recordings (out of 92) were excluded. We had no “within-subjects” comparisons planned (we were not comparing Participant A’s reaction during the SIMS ad to their reaction in the GTA ad). We did not need similar sample sizes for the GTA condition and the SIMS condition, because we were not comparing these conditions. Thus, we chose to only exclude the low-quality recordings rather than exclude the participant from all of the data.*

## **Results: How to get insights from your data**

Here we will go through the data analysis stimulus by stimulus. We start with the baseline where we will get an overview of the variance in our participants and whether there are any outliers to be aware of. Then, we look at participants during the ad, where we measured respiration. Finally, we look at facial expression, voice, and sentiment metrics when participants were asked questions about their experience (What they liked and disliked) and intentions (whether they would like to play and/or buy).

### **Baseline Analysis**

Here we want to look to have an idea of how much variance there is in the baseline and get an indication of whether or not there were any outliers. In this study, at Baseline, Voice Valence ranged from -0.28 to 0.23 but showed a mean value of -0.01 and a median of 0.03. In this study, at Baseline, the percentage of positive FEA adaptive valence showed a range from 0.00 to 7.49 had a mean of 0.40 and a median of 0.00; and the percentage of negative FEA adaptive valence showed a range from 0.00 to -7.98 and had a mean of -0.25 and a median of 0.00.

These values for individuals will be used to normalize the data for the stimulus where we ask them about their likes, dislikes, purchasing intention and desire to play.

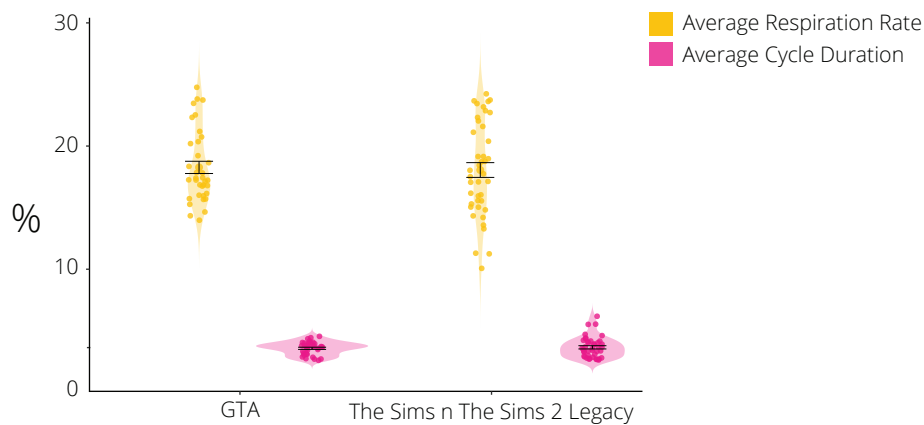
### **Watching Ads**

#### **Respiration Rate**

Respiration rates were compared directly per participant per stimuli. In this study, we used a t-test to compare respiration rates during the two different advertisements.

In this study, comparing the average respiration rate across the ads, there was no significant difference in respiration rate between the two advertisements. In this case, we can see that both ads had similar effects on arousal levels. If we wanted to better understand the different emotions the ads evoked, we could examine facial expressions during the ad from the same video data the respiration data was extracted from.

A different research question could have been to understand how different demographics react to the ads. If you include a survey, you could segment the data based on age, gender, etc and evaluate if respiration rates changed across these groups.



**Figure 3.** Respiration metrics while watching ads. Figure shows both the average respiration rate and the average respiratory cycle duration while participants watched ads. Summary statistics indicate means with standard error of the mean.

Another analysis idea would be to look at respiration rate/cycle duration during certain scenes, or comparing respiration rate at the beginning of an ad compared to the end of an ad. Be mindful that respiration rate needs some time to be calculated and shouldn't be measured over a few seconds. On average, we take up to 5 seconds to finish one breath cycle. Shorter comparisons can skew the results based on when a cycle started or ended.

## Participant Reflections

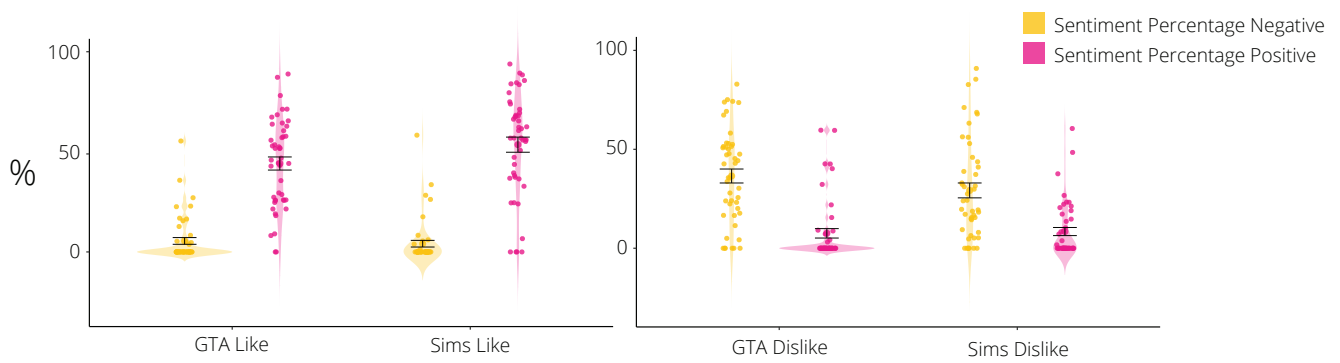
Participants were asked about their experience and intentions while we measured facial expressions, voice, and sentiment during each question.

### Experience Questions: Like and Dislike

When using surveys for online research, researchers often ask whether people liked or disliked the ad by presenting survey questions with scales. Sometimes surveys include a free-text field to encourage participants to be specific about what they did and did not like. There may even be more targeted questions to see if people liked and disliked certain aspects of the ad (the actors, the music, certain scenes, a punchline, etc).

What does data from faces, voices, and words provide that is different from a survey?

Measuring emotional expression gives insights into what people feel when they recall the ad in question. For example, if people are talking about what they liked about an ad, we generally expect more positive valence during their responses. For dislike, we might expect more negative valence. In this study, sentiment analysis for the Like/Dislike questions shows what we would expect for both ads (Figure 4).



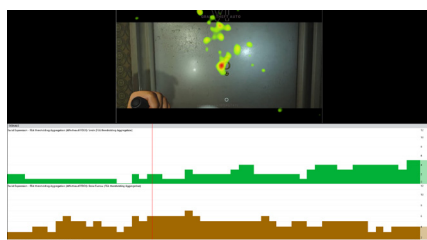
**Figure 4.** Sentiment Analysis from Voice data. The left figure shows sentiment analysis when participants were asked about what they liked about the ads, while the right figure shows the analysis of when the same participants were asked about what they disliked. Summary statistics indicate means with standard error of the mean.

## Why is this useful?

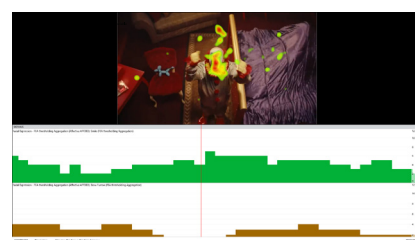
1. Without reading through all of the responses, we see that participants' expressed sentiments align with what the questions were asking.
2. In an A/B test like this, we could compare which ad elicited more positive or negative valence. If we collect reactions during the ad, we could compare this to their reflections after the ad.
3. If people strongly disliked an ad, they may show negative valence, when asked to reflect on what they liked about the ad. You can also have the opposite situation; if people really like the ad, they may not express much negative valence when they talk about what they disliked.
4. We can further use FEA to check whether participants had divided reactions to certain scenes. This could be a division in engagement and/or valence.

## Scene Analysis Example

In this screenshot from iMotions, we can see smile (green) and brow furrow (brown) across two different scenes, one from the SIMS ad and one from the GTA ad. In both cases, we see 5-second bins and how many participants crossed the threshold set for the FEA in each bin. For more information about FEA thresholding, see the section on FEA Signal Processing.



The GTA scene shows the opposite trend of increased brow furrow and reduced smiling as they are about to break into a room (with possible illegal activity).



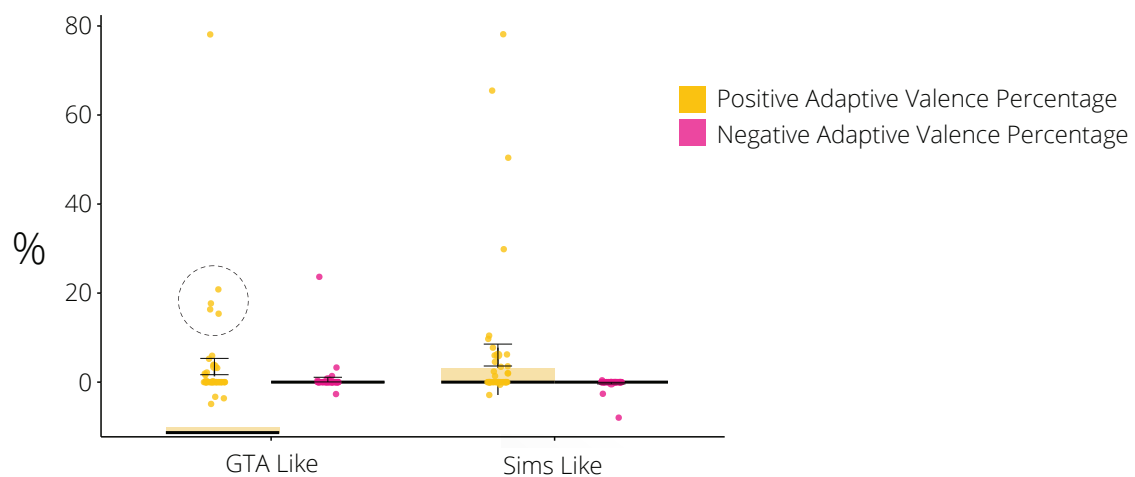
The scene from SIMS shows an increased smiling response and decreased brow furrow when the sex scene is playing.

What are the advantages of having multiple metrics for positive and negative valence for the same questions?

We do not necessarily see the same pattern in facial expression or voice analysis. When different modalities do not show the same trend, there may be an opportunity for more complex considerations.

## Cluster Analysis Example

If we look at the facial expression analysis data for GTA-Like, we see a small cluster of participants who seem to have more positive facial expressions when talking about what they liked about the GTA ad.



**Figure 6.** Facial expression analysis with adaptive valence. This figure shows positive and negative adaptive valence when respondents were discussing what they liked about the ads. Summary statistics indicate means with standard error of the mean.

One interpretation could be that these individuals feel exceptionally positive about the ad compared to the rest of the group. We could segment the data from these individuals to figure out what they may have in common.

For example, we could look at their respiration rate or facial expressions while watching the ad to see if these individuals were particularly expressive during the ad or which scenes elicited strong expressions. This could tell us which scenes contribute to their positive valence during reflection.

With demographic information, we could see if this cluster represents a specific segment of potential customers (Are they similar age, gender, location, or vocation?). This is useful for knowing which audience this ad may be most effective for. If we had information about participants' experience playing GTA, we could see if this cluster represents people who have played GTA before.





# Guidelines for Data Analysis

## Statistics Strategies

### Targeted t-tests based on descriptive statistics or hypotheses

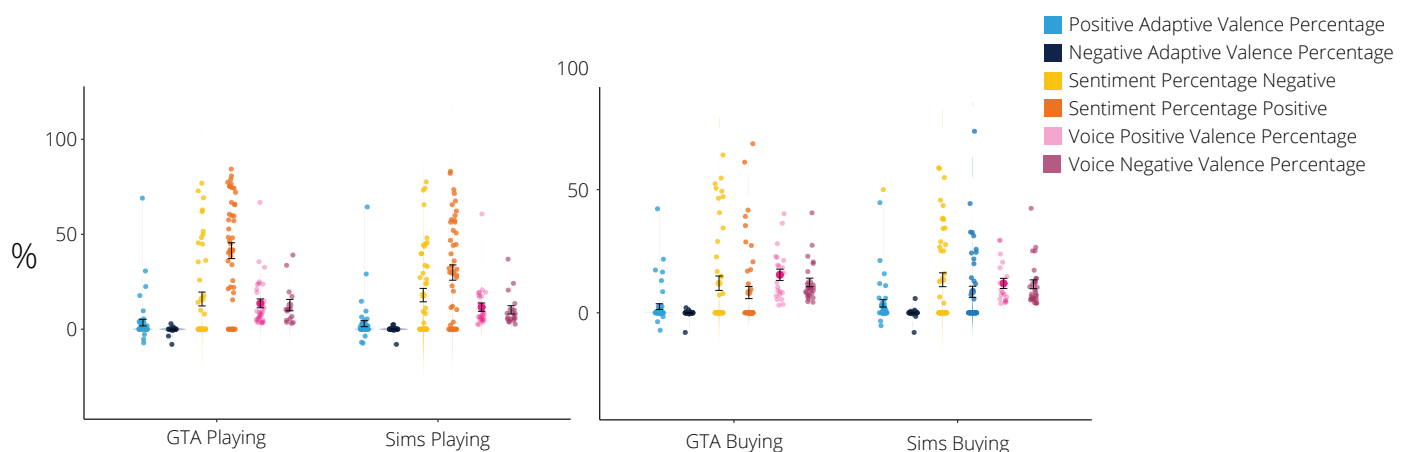
If you have a specific hypothesis that includes a specific comparison, you could perform a t-test.

For example, looking at the “Playing” question and considering the panel on sentiment % and focusing on positive sentiment, as shown above ( $t = 1.9802$ ,  $df = 89.889$ ,  $p\text{-value} = 0.05$ ). This is an example of how you might approach A/B testing.

The advantage of asking more specific questions is that you are likely to find more specific answers. The disadvantage is that you do have to be decisive about what you are looking for.

### Mixed models to study effects

Since there are multiple independent and dependent variables when analysing multimodal biometric data, one may be tempted to run ANOVAs across multiple IVs and DVs. These are considerable interactions and comparisons. Running large models like these can seem like fishing for significance.



**Figure 11.** Example of combining metrics from voice and face analysis. These figures illustrate the complexities of large models.

For more explorative research, we still recommend having a rationale driven mixed models. One approach could be to create a model per research question. In the present study we created a mixed effects model (this could be a MANOVA or ANCOVA depending on your research questions and variables) per question asked. This reduced the model to one main effect that we wanted to study (e.g. Do people react differently when asked about playing SIMS vs GTA).

## Controlling for individual differences on a segment level

Voice can have a lot of variance and aggregating like above may lose some of the variability from between speech segments. Imagine talking for 30 seconds where the voice only peaks in emotion for some relevant segments and stays neutral the rest of the time. To account for this, the present study baseline corrected per speech segment and added the individual level differences as a second controlled factor in the mixed model. This proved promising in the present study using valence and activation from voice analysis.

In this case, multiple MANOVAs provided insights into how the ads differ on voice analysis alone:

- **For activation** - there was a significant difference ( $p < 0.0001$ ) between the ads in all four questions, experience and intent, with GTA showing consistently higher activation.
- **For valence** - there was a significant difference ( $p < 0.0001$ ) when asked about intent to play and intent to buy, with GTA again showing consistently higher valence.

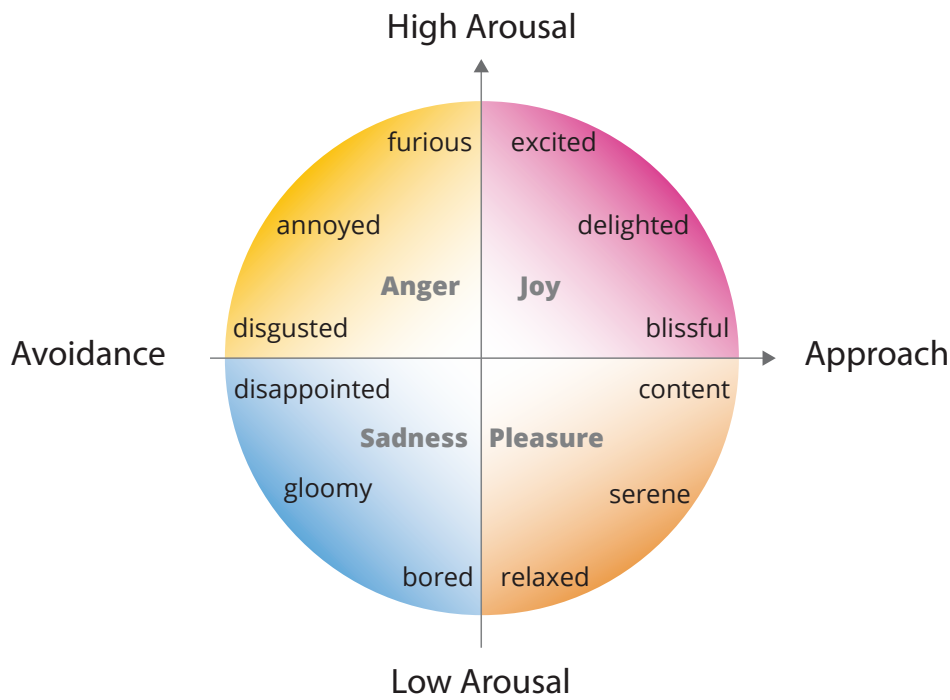
## Data Visualization: Circumplexes

With circumplexes, you can compare approach and avoidance behavior for both ads. Circumplexes can be used as graphical representations of emotions. In consumer research, this is typically shown as approach-avoidance (or valence) on the x-axis and arousal (or activation) on the y-axis, creating quadrants. In the figure below are examples of emotions that fit into the respective quadrants.

With the data set from this study, there are two ways to create circumplexes. For both, you use voice activation to represent arousal (y-axis). Then you can decide whether to use valence from facial expressions, voice, or create a composite score of both. Below, circumplex 1 uses a composite score of valence from face and voice and circumplex 2 uses voice alone.

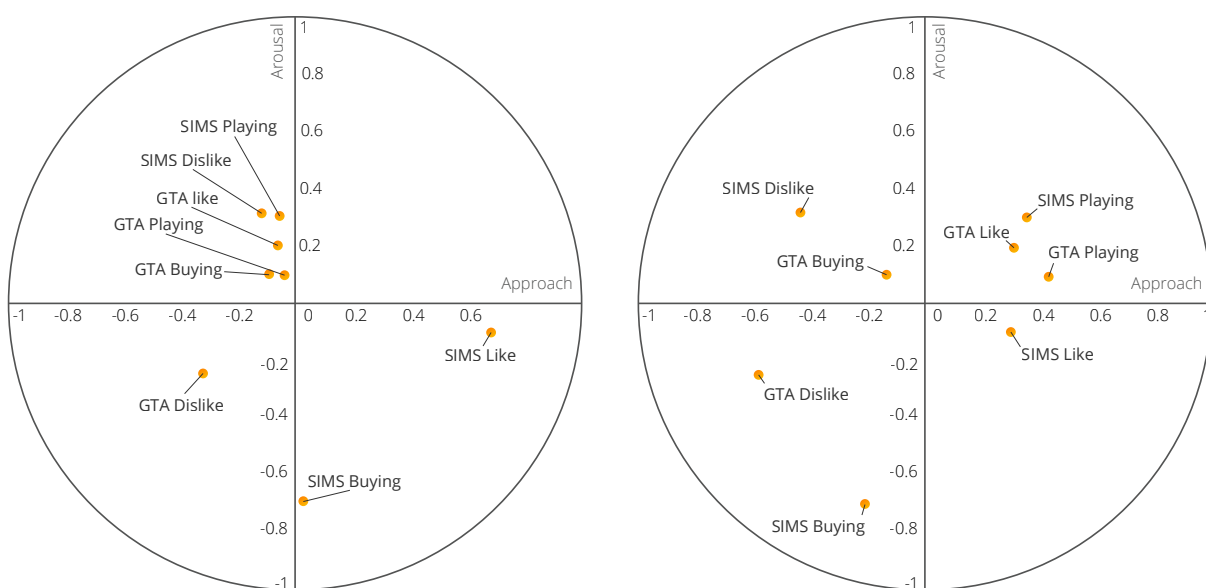
## Comparing Circumplexes

As you can see in the figure above, these circumplexes give quite different information. Keep in mind that these two circumplexes have the same y-axis. When we compare these circumplexes, we are only comparing the difference in the x-axis.



**Figure 12:** Circumplex diagram showing how different emotions align amongst axes of arousal compared to approach/avoidance behavior.

For FEA-Adaptive Valence, we did not see a lot of variability, suggesting that facial expressions were fairly neutral, and participants' facial expressions did not vary in valence as much as their voices did while answering the questions. This is reflected in the FEA and Voice circumplex which pulls the conditions to the center, whereas the Voice circumplex maintains a lot of the variability in expression. For this reason, we continue deriving insights from the Voice-Only circumplex.



**Figure 13.** Both circumplexes use voice data for arousal (y-axis) but differ in which data was used for approach (x-axis). The left circumplex uses facial expression adaptive valence and the right circumplex uses valence from voice.

## **Circumplex Results: Experience Questions (Like and Dislike)**

Looking at the experience questions (like and dislike), the Voice-Only circumplex shows that responses for Like are higher in approach and for Dislike are lower in approach. This is what we expect. When people express what they like about something, we expect to measure more approach behavior. When they express what they dislike, we expect more avoidance.

When questions presented in the survey are focused on discussing something strictly positive or strictly negative (as in the case of the Like and Dislike questions), verifying that the emotional expression aligns with that focus gives us confidence in which circumplex best suits our data.

## **Circumplex Results: Intention Questions (Playing and Buying)**

For both ads, the question regarding playing falls in the quadrant we associate with excitement in the Voice Only-circumplex. For buying, the responses are more neutral. Based on this discrepancy, we would further investigate why our participants would like to play a game, but not buy it. This can be done by adjusting pricing in a follow-up study, or having a second study just to estimate what income groups to target with the release of an ad. We could also investigate if the increased arousal when talking about buying GTA could be converted to purchase behavior faster than a game like SIMS.

# Conclusions and Takeaways

7

This report is an example of best practice guidelines for online research using voice analysis in combination with facial expression analysis and respiration.

We covered guidelines for the study design phase, data collection, signal processing and interpreting results. This report also covered examples of data visualizations (Word Clouds and Circumplexes) and examples of potential analyses and future studies, given the findings in this dataset.

This study focused on ad testing, but these same considerations are relevant for package and shelf testing, UX research, or Entertainment content testing. **These guidelines are applicable for investigating emotional expression in a screen-based setting, whether that be remote or in-lab.**

## References

1. AudEERING GmbH (2024) What the voice reveals [whitepaper]
2. Bishay, M., Preston, K., Strafuss, M., Page, G., Turcot, J., & Mavadati, M. (2023, January). Affdex 2.0: A real-time facial expression analysis toolkit. In 2023 IEEE 17th international conference on automatic face and gesture recognition (FG) (pp. 1-8). IEEE.
3. Derington, A., Wierstorf, H., Özkil, A., Eyben, F., Burkhardt, F., & Schuller, B. W. (2025). Testing Correctness, Fairness, and Robustness of Speech Emotion Recognition Models. *IEEE Transactions on Affective Computing*, 16(3), 1929-1941. <https://doi.org/10.1109/TAFFC.2025.3547218>
4. iMotions (10), iMotions A/S, Copenhagen, Denmark, (2024)
5. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.