INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Spot the pleasant people! Navigating the cocktail party buzz

Christina Tånnander[1,2], *Per Fallgren*[1], *Jens Edlund*[1], *Joakim Gusafsson*[1]

[1]Speech, Music and Hearing, KTH, Sweden
[2]Swedish Agency for Accessible Media, Sweden
christina.tannander@mtm.se, perfall@kth.se, edlund@speech.kth.se, jocke@speech.kth.se

Abstract

We present an experimental platform for making voice likability assessments that are decoupled from individual voices, and instead capture voice characteristics over groups of speakers. We employ methods that we have previously used for other purposes to create the Cocktail platform, where respondents navigate in a voice buzz made up of about 400 voices on a touch screen. They then choose the location where they find the voice buzz most pleasant. Since there is no image or message on the screen, the platform can be used by visually impaired people, who often need to rely on spoken text, on the same premises as seeing people. In this paper, we describe the platform and its motivation along with our analysis method. We conclude by presenting two experiments in which we verify that the platform behaves as expected: one simple sanity test, and one experiment with voices grouped according to their mean pitch variance.

Index Terms: voice likability, voice preferences, evaluation

1. Introduction

The cocktail party effect is the observation that important signals carry through noise with more ease than others [1], and most of us feel that we can pick up the general atmosphere in a crowded room quickly by simply listening to the murmur of voices. In the present work, we take steps towards harnessing these skills for voice likability and voice preference assessment.

These assessments are carried out for various reasons, from the matching of products with voices that may increase sales, through selecting suitable voices for entertainment, through exploring what makes voice attractive in mating situations, to a main motivation for the work presented here: finding which voices are best suited for reading texts aloud in learning situations.

Today, most voice likability assessments are tests in which the respondents are instructed to rate individual voices on a Likert scale. If many voices are to be assessed, this process is time-consuming and cumbersome, with a risk of increasingly bored and careless respondents. Perhaps more importantly, the results of this procedure are difficult to interpret. A respondent rating an individual voice might respond to any combination of an almost infinite set of traits of greatly varied nature: enunciation, pronunciation, intonation, speaking rate, accent, speaking style, voice quality, gender, pitch and loudness, to mention but a few. While Likert scales may tell us which of a few voices is preferable in some context, they are less likely to tell us why.

As a step towards a better understanding of the makings of a likable or a preferred voice, we are looking for a method in which individual speakers and voices become less prominent and give way to quantifiable similarities in groups of voices. Our goal is to be able to answer questions like "Are voices sharing trait T preferable to those that do not, generally speaking?", without having to risk exhausting respondents with a seemingly endless string of voice samples for assessment.

Our previous experiences with artificial cocktail party-style buzz led us to speculate that this could be a meaningful approach to the problem. A technique to create such buzz in such a manner that its composition and characteristics can change within moments was already known to us. We also have the means to control these changes intuitively, in a manner reminiscent of navigating through a space or in a room [2],[3]. Finally, we have positive experiences of asking respondents to listen to such buzz and report their impressions. That, however, is a far cry from a reliable and validated framework for assessment of voices. Here, we focus on the following questions: (1) What is a sensible way to analyse the reactions we elicit? and (2) What is a suitable way to organise voices in the listener space in order to get useful responses? We hope to be able to investigate which voice traits affect voice likability by letting respondents listen to voice buzz created from voices with one or more voice features in common.

To tackle (1), we propose a method of analysis and we test it with a manufactured mini-study to exemplify its effects. To tackle (2), we follow this up with a study in which we organise a 2-dimensional listening space according to the pitch characteristics of voices, with voices sorted according to their average pitch on one axis, and according to the standard deviation of their pitch on the other. As a test of this first attempt, we present a study where this organisation is used to find preferences for pitch characteristics.

2. Background and related work

2.1. Voice likability

Voice likability has been described as "the sound of her/his voice and manner of speaking" [4], or "a speaker social characteristic that can determine the listener's attitudes and decisions towards speakers and their message." [5]. Both definitions make the point that voice likability does not only concern the speakers' inherent voice characteristics, but the manner in which speakers intentionally or unintentionally communicate their message.

http://dx.doi.org/10.21437/Interspeech.2019-1553

Voice likability assessment has been used for a number of purposes in varying areas. This ranges from marketing/sales (e.g. [6], [7]) and propaganda/politics ([8]) to entertainment (e.g. fiction audiobooks or digital assistants). A number of studies has investigated attraction, love, flirt and mating.

Voice likability is often associated with voice attributes, or voice traits. Objective or objectively measurable attributes include pitch, speech rate, intensity, centre of gravity, skewness and other spectral parameters (see e.g. [9]), while subjective attributes, such as trustworthiness, clarity, charisma, attractiveness and engagement, are generally accessible only with the help of human input.

2.2. Voice likability assessment

Voice likability studies typically use a Likert scale from 1 to 5, 7 or 9 (e.g. [5]–[10]), but there are also examples of binary (non-likable or likable) or trinary likability (non-likable, neutral or likable) likability assessment [11]. In these studies, respondents judge voices sequentially, one at a time, which makes evaluations of many voice samples tedious and time-consuming. In other studies, different voices are pitted against each other and their likability compared, sometimes after separating female and male voices (e.g. [5]). [5] also shows that comparisons between likability scores obtained in laboratory settings correlate well with scores from crowdsourcing. There is also an increasing amount of work in which spectral parameters are used to predict voice likability automatically (e.g. [9], [12]).

General findings include that likability ratings increase with higher articulation rate, lower spectral centre of gravity, lower f_O (only male speakers), higher spectral standard deviation and skewness (only female speakers), and with lower 3rd central moment (only female speakers) [9]. [13] sees "a significant increase in voice attractiveness ratings as the risk of conception increased across the menstrual cycle in naturally cycling women", but "no effect for women using hormonal contraceptives".

2.3. Artificial cocktail party buzz

We have previously used a technique called massively multi-component audio environments (MMAE) for purposes ranging from perception experiments to building dynamic soundscapes to audio and speech corpus browsing. In short, MMAE builds a dynamic audio environment by firing a multitude of very short (typically 200-1000ms) sound snippets at a high rate (typically every 50-100ms), so that a number of sounds are played simultaneously at any given time. Our original tool for this was called Cocktail, as a homage to the cocktail party effect, since the result of building MMAEs from speech is very similar to authentic cocktail party buzz [2].

More recently, we have combined MMAE with temporally disassembled audio (TDA), that is longer audio sequences that are segmented into small snippets, which are then reorganized along some other dimension(s) than the temporal. If organised along two dimensions, the sound snippets can be visualized intuitively on a screen, which can in turn be used to control the MMAEs, such that they are built from sound snippets in a specific area of the screen, and change dynamically as another area is targeted using a mouse or touch screen [3]. The inspiration for this mode of control came from Manny Tan's and Kyle McDonald's Google AI experiment *Bird Sounds*[1].

3. Method

3.1. Speech corpus

The Swedish Agency for Accessible Media (MTM) produces over 2 200 talking books per year [14]. While most fiction is produced with human voices, about 50% of university text books are made with speech synthesis. The talking book production takes place in accordance with an exception in the Swedish copyright law permitting MTM to produce talking books for people who cannot read for one reason or another, without permission from the publisher. At the beginning of each book, there is a spoken copyright disclaimer referring to this judicial exception. It also contains some meta information about the book in question, such as the number of pages and the number of header levels. Physically, this information is stored separately in single sound files, and it is these sound files that constitute our speech corpus. The messages in these files are very similar, though not identical, across the corpus. The files are mp3 encoded and sampled at 22 050 kHz.

In total, MTM administer a collection of more than 124 000 talking books. After filtering out everything but Swedish books read by human narrators produced by MTM after 2011 and removing files that showed signs of being the wrong type of file, we have a corpus of 12 083 readings.

3.2. Cocktail experiment platform

A new platform based on TDA and MMAE was implemented. The platform runs on any standard compliant web browser, with pages that can be served by any conventional web server.

The goal of the platform is to encourage a sense of moving around in a room filled with people talking to each other in the respondents. It is designed to be used with a touch screen, and to be as useful to visually impaired people (who are big consumers of spoken text, such as talking books, news, or web pages) on the same premises as seeing people. In experiment mode, the interface shows nothing but a white square (Figure 1, right pane).

The left pane of Figure 1 shows the platform in demonstration mode, and we see an example of how audio files can be organised over the surface of the interface. The two black concentric circles show an example of the uptake area – this is the area from which the sounds that create the MMAE are taken, as follows. (1) each of the audio files under the circle is segmented into short snippets of duration D, (2) at a rate of R, a random snippet is selected from this set and immediately replayed, (3) sound snippet loudness is adjusted according to some algorithm, which is visualized in the demonstration mode (in Figure 1, we use an inner circle at full loudness and an outer at 20%), which allows us to create a sense of foregrounded and backgrounded speakers, and (4) as the respondent touches different areas of the screen or drags around it, the set of sound snippets from which the random selection is made is updated to reflect this. When the respondent finds what is sought after (this depends on the task), they simply tell the experiment leader that they are done and leave the screen as is, and the system registers the final position.

[1] https://experiments.withgoogle.com/bird-sounds

3.3. Analysis

We aim for an analysis that models what we know: what a respondent is able to hear when they make their choice. For each respondent, we register the uptake area where they make their choice, and we take into account the loudness of sounds taken from the different part regions of the uptake area. In Figure 1, if the two concentric circles denoting the uptake area of a respondent at the point where they make their choice, we add 1 to the area of the inner circle, and 0.2 to the area of the outer circle that is not in the inner circle. We sum all respondents results in a manner reminiscent of kernel density estimates. We then sample the coordinate system at some interval, again as one would a kernel density estimate, and either plot the result or describe them statistically. With random selections, we would expect an even distribution given a sufficient number of respondents. If respondents are indeed selecting one or more specific areas based on the soundscape and the question they have been asked, we will see areas with higher values and areas with lower values.

In order to get smooth results that could perchance be modelled mathematically, we would need large numbers of respondents. Here we are interested in the ability to make rapid initial studies, with only a few handfuls of respondents. To still be able to say something about reliability, we implement an empirical method to provide a significance estimate. We first find a single highest value of the experiment under investigation with 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, amongst the samples. Repeat this for some large number of iterations and store each max value. Finally, sort the max values and find the N^{th} percentile – 95^{th} for an estimate of the threshold at which the chance of a max value occurring by coincidence is less than 5% or the 99^{th} for a less than 1% estimate.

3.4. Experiments

We conducted two experiments with different purposes. The first is a *sanity test for the platform, including our significance estimates*, and the second an initial attempt to see whether our method allows us to (a) *say something about likability in general* and (b) *say something about how speech attributes affect likability*.

3.5. Experiment 1: Where are the men?

As a first evaluation of the Cocktail platform, we gave respondents a task intended to be relatively difficult, yet very easy to evaluate. A small group of respondents were asked to find two places in the "room": one with the most women, and one with the most men.

3.5.1. Stimuli

We used a subset of 200 female and 200 male voice recordings from the corpus. These were organised in an equidistant 20x20 grid, such that every other recording was female and every other was male, with the exception of two 5x5 areas that were uniformly female and male, respectively. An uptake area covering roughly nine recordings was used. The effect is that on most places, a respondent will hear a balanced mix of female and male voices, and in two areas, covering one sixteenth part of the total area each, will voices from one gender only be heard. The respondents do not know this, however, and if they stand a good chance of making their choice without having even passed through one of the single-gender areas. The layout is shown in Figure 1. We used a snippet length of 500ms and a firing rate of 50ms, meaning that in effect, 10 voices played simultaneously at all times.

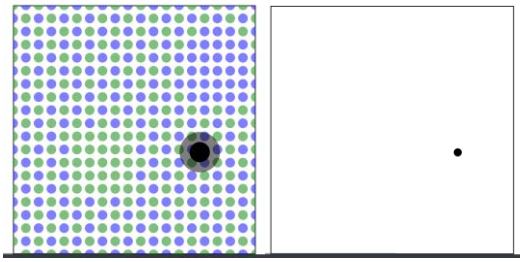

Figure 1 *The left pane shows the demonstration view, with 200 female voices (blue) and 200 male voices (green) evenly distributed, with the exception of two 5x5 clusters. The uptake area is visualized as two concentric black circles. The right pane shows the respondents' blind view of the same configuration.*

3.5.2. Respondents and procedure

Eight respondents, four women and four men, were asked to find the clusters of female and male voices by moving around in the room, displayed as a white space surrounded by walls (the right pane in Figure 1) with the respondent's position marked as a movable dot. The experiment leader noted the coordinates of the two spots selected by each respondent by triggering a function in the platform.

3.5.3. Analysis

The eight selections of particularly female places and the corresponding eight selections of male spots were analysed in the manner described in 3.3, with 10 000 iterations of eight random selections to acquire a confidence estimate at the 0.005 level (the 50^{th} highest max value created by random selections).

3.6. Experiment 2: Where are the pleasant guests?

The second experiment was designed to investigate whether the Cocktail platform can in fact shed light on voice likability issues. A group of respondents were asked to find the spot which they found most pleasant in what they were told was an emulation of a cocktail party. We primarily wanted to know if respondents would make similar choices or respond randomly. If there was some agreement, we wanted to know if this could be associated with objectively observable voice attributes

3.6.1. Stimuli

The second experiment was divided into two Cocktail tests, one with female and one with male voices. A subset of 400 readings was used in each test, plotted along the X axis according to their mean pitch and along the Y axis according to their pitch standard deviation. f_0 extraction was done with getF0 and converted to pitch estimates on a semitone scale. Outliers were excluded to avoid too large empty areas in the corners. Figure 2 show the resulting layouts.

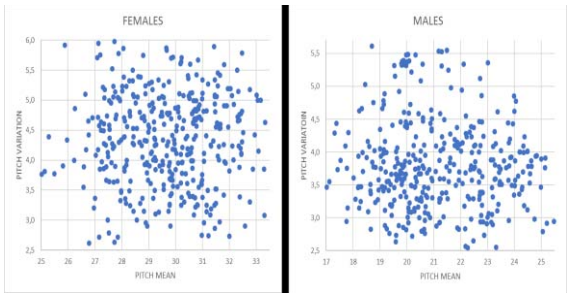

Figure 2 *Plot of 400 female (left pane) and 400 male (right pane) voices as they were organized in the tests.*

3.6.2. Respondents and procedure

24 respondents took part in the experiment, 12 women and 12 men. There was no other pre-selection. Respondents with hearing impairments (2 persons) and other mother tongue than Swedish (4 persons) we allowed to participate.

The respondents were asked about gender, age, hearing level and knowledge of the Swedish language. They were then asked to move around the two-dimensional room ("the Cocktail party") using the touch screen, to get acquainted with the sound environment, and then to choose the location where they found the voices most pleasant. The coordinates of this location were registered. This was done twice for each respondent, once with male voices and one with female voices. Finally, the respondents were asked if they found the Cocktail test fun to conduct (Likert scale 1-5 where 1 is "very boring" and 5 "very fun"), and if they found it hard to find the place in the room where they like the voice the best (scale 1-5 where 1 is "very easy" and 5 "very hard").

3.6.3. Analysis

For each of the two tests, the 24 selections of particularly pleasant places were analysed in the manner described in 3.3, with 10 000 iterations of 24 random selections to acquire a confidence estimate at the 0.005 level (the 50th highest max value created by random selections).

4. Results

In experiment 1, finding the female and male voice clusters among the 400 voices in the first test took on average about 2,5 minutes altogether.

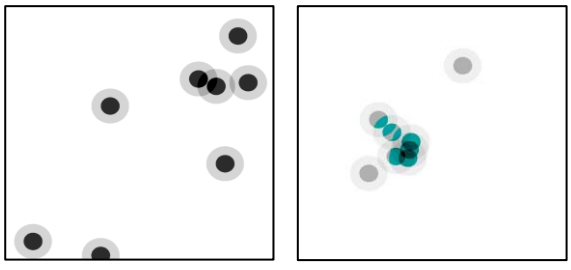

Figure 3 *Plots describing analysed responses of finding the women (left pane) and the men (right pane). Values above significance estimate at 0.005 are coloured.*

Figure 3 shows the results after analysis.

In experiment 2, the test took 3,5 minutes per voice on average. The respondents thought that is was fun to do the tests (3,8 on a Likert scale 1-5), and that it was rather hard to find most pleasant sound (3,75 on a scale where 1=very easy, 5=very difficult).

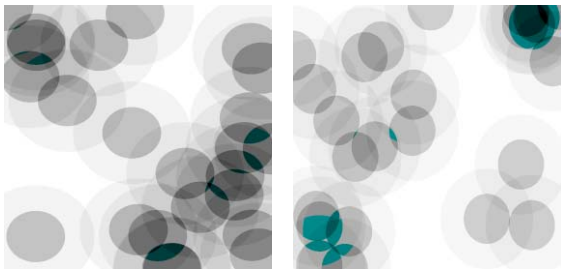

Figure 4 *The analysed results for the female voices (left pane) and the male voices (right pane). Values above the significance estimate at 0.005 are coloured.*

Figure 4 show the distribution of choices for the two tests, after analysis.

5. Discussion

In experiment 1 (Figure 3), the upper right cluster of four responses in the left pane coincides with the female only area, but too few of the respondents found it to make it different from the random iterations. In the right pane, we see that most respondents found the all-male area, and the cluster yields values that are unlikely to occur by coincidence. This is encouraging and lends support to the empirical significance estimates.

In experiment 2 (Figure 4), we note that both tests give results that are above our significance estimates. The clearest area is the upper right part of the male voices (right pane). Unfortunately, we know from several spontaneous comments from respondents that that area was chosen because of its relative quietness. If we disregard that, there is a clear cluster in the lower left corner, corresponding to low pitch and low pitch variation. In the female voices (left pane) there seems to be two clusters as well, one corresponding to low pitch and high pitch variation, and one to high pitch and low pitch variation. The latter is seemingly stronger.

We are encouraged by these results and note that the system works well in practical terms, and respondents are happy to use it. Our initial tests yielded positive results, and although we would not go as far as to say that the second test clearly shows for example a preference for low mean pitch in male voices, it does show that respondents seem to agree to a sufficient extent to make this type of testing interesting. We also note that our respondents in effect listened through hundreds of voices in mere minutes. The next step is to run a series of larger tests with voices are organised along other attributes.

6. 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). The framework will be made publicly available under an open source license through the national research infrastructure Nationella Språkbanken (Swedish Research Council 2017-00626) once it is robust and ready for release.

7. References

[1] N. Moray, "Attention in Dichotic Listening: Affective Cues and the Influence of Instructions," *Q. J. Exp. Psychol.*, vol. 11, no. 1, pp. 56–60, 1959.

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

[3] P. Fallgren, Z. Malisz, and J. Edlund, "Bringing order to chaos: a non-sequential approach for browsing large sets of found audio data," in *Proc. of the 12th International Conference on Language Resources (LREC2018)*, 2018.

[4] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "'Would You Buy A Car From Me?' - On the Likability of Telephone Voices," in *Interspeech*, 2011, pp. 1557–1560.

[5] L. F. Gallardo, R. Zequeira Jiménez, and S. Möller, "Perceptual Ratings of Voice Likability Collected through In-Lab Listening Tests vs. Mobile-Based Crowdsourcing," in *Proceedings of Interspeech*, 2017, pp. 2233–2237.

[6] A. Chattopadhyay, D. W. Dahl, R. J. B. Ritchie, and K. N. Shahin, "Hearing Voices: The Impact of Announcer Speech Characteristics on Consumer Response to Broadcast Advertising," *J. Consum. Psychol.*, vol. 13, no. 3, pp. 198–204, Jan. 2003.

[7] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and W. J. Barry, "Modelling personality features by changing prosody in synthetic speech," in *Speech Prosody*, 2006.

[8] E. Strangert and J. Gustafson, "What makes a good speaker? Subject ratings, acoustic measurements and perceptual evaluations," in *Proceedings of Interspeech*, 2008, pp. 1688–1691.

[9] B. Weiss and Burkhardt, "Voice Attributes Affecting Likability Perception," in *Interspeech*, 2010, pp. 1934–1937.

[10] B. Schuller *et al.*, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proceedings of Interspeech*, 2012.

[11] S. Hantke, E. Marchi, and B. Schuller, "Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification," in *LREC*, 2016, pp. 2156–2161.

[12] F. Eyben, F. Weninger, E. Marchi, and B. Schuller, "Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation," in *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[13] R. Nathan Pipitone and G. G. Gallup, "Women's voice attractiveness varies across the menstrual cycle," *Evol. Hum. Behav.*, vol. 29, no. 4, pp. 268–274, 2008.

[14] MTM, "MTM Årsredovisning 2018," Stockholm, Sweden., 2019.