Audio Engineering Society

Convention Paper 10033

Presented at the 145th Convention
2018 October 17 – 20, New York, NY, USA

Developing a Method for the Subjective Evaluation of Smartphone Music Playback

Elisabeth McMullin[1], Victoria Suha[1], Yuan Li[1], Will Saba[1], and Pascal Brunet[1]

[1]*Samsung Research America, Audio Lab, Valencia, CA 91355, USA*

Correspondence should be addressed to Elisabeth McMullin (e.mcmullin@samsung.com)

ABSTRACT

To determine the preferred audio characteristics for media playback over smartphones, a series of controlled double-blind listening experiments were run to evaluate the subjective playback quality of six high-end smartphones. Listeners rated products based on their audio quality preference and left comments categorized by attribute. The devices were tested in different orientations in level-matched and maximum-volume scenarios. Positional variation and biases were accounted for using a motorized turntable and audio playback was controlled remotely. Test results were compared to spatially-averaged measurements made using a multitone stimulus and demonstrated that the smoothness of frequency response was a key aspect in smartphone preference. Low-frequency extension, decreased levels of distortion and higher maximum playback levels did not directly correlate with increased preference.

1 Introduction

In the past decade, as smartphone ownership has skyrocketed, mobile phones have become the most ubiquitous audio devices consumers own. As of 2018, 77% of Americans own a smartphone [1] and even though there are many high-quality headphone and speaker options available, at some point virtually all of these consumers listen to music or watch videos using their phone's built-in speakers. Despite the prevalence of their usage as audio devices, there is a scarcity of scientifically rigorous research into what listeners prefer in smartphone audio. This paper proposes an approach to evaluating smartphones both objectively and subjectively in order to find trends in what listeners prefer and why.

Due to the obvious physical limitations of smartphone transducers and enclosures, there are consequently audible issues in the playback of each device. In a typical listening preference test of loudspeakers or headphones listeners are identifying minor impairments of sound quality. In contrast, during the evaluation of mobile phones the question is often not, "Are there audible impairments?" but rather, "Which of the present impairments is the least objectionable?" Furthermore, due to limited frequency reproduction capabilities of these devices, low-frequency extension rarely goes below 400 Hz. Because of this, low-frequency masking effects are lessened and distortions in the midrange and highs will be more audible [2].

While there are many tools available for evaluating speech quality on mobile phones [3][4], researchers ex-

ploring music and media playback have limited tools at their disposal. Algorithmic tools such as PEAQ [5] can be used with recorded signals to evaluate the perceived audio quality, but cannot fully replace a thorough subjective evaluation. Standards such as MUSHRA [6] can also be utilized but are difficult to implement correctly with smartphones since the anchors specified may not be appropriate for the application and the number of devices required would slow down the testing process.

While all double-blind subjective audio tests have their unique challenges, smartphone audio is particularly difficult to test for a number of reasons. Smartphones are watched and listened to by users in a variety of positions, including in the hands of the user, oriented vertically or horizontally, flat on a tabletop, and propped up by a kickstand on a tabletop. For simplicity in these tests, it was decided that handheld positions, in which the user holds the phone without interfering with the speakers would be studied. Another issue is that there are no easy methods for playing back audio remotely through each smartphone. To overcome this challenge, our team loaded each device with remote playback software [7][8] which allowed a test administrator to start audio playback on each phone from a connected tablet. Each phone was connected to its own tablet. A third issue is that smartphone audio relies heavily on compressor and limiter settings and usually has separate tunings for each volume level. In these tests, only two volume scenarios are addressed: maximum volume and level-matched. The actual playback level for the level-matched tests was set at a moderate 64 dB SPL(A) at the seating position. Finally, the typical viewing distances for mobile phone usage are quite varied and range between 19-60 cm [9], which makes it logistically difficult to locate the products within proper listening range of the listeners while still adequately obscuring the turntable and the devices.

2 Experiment Setup

2.1 Phones Tested

Six phones were chosen for these experiments based on their popularity in the American and Chinese markets. Five of the devices have “stereo” playback, meaning both the receiver and main speaker are utilized for media playback. The phones range in screen size from 13.97 cm to 15.75 cm and in price from $649 to $999.

2.2 Multitone Measurements

The phones were measured using the Samsung Audio Measurement System (SAMS). Each phone was loaded with an audio file containing an 8-second multitone stimulus which was looped into a several minute long clip and loaded onto each phone to allow it to play throughout the duration of multiple measurements. A multitone stimulus was selected to allow for easy measurement of frequency response, noise, and distortion as described in [10]. Importantly, it also accurately mimics a musical signal, allowing the phone’s compressors and limiters to engage properly for the measurement. Since most phones are voiced to be as loud as possible, compressors and limiters are used heavily and properly engaging them is vital to getting a realistic representation of what the listener is hearing.

Each phone was mounted in a custom fixture attached to a motorized turntable in an anechoic chamber as shown in Fig. 1. The phones were first measured in their vertical orientation and then in their horizontal orientation. Using a 1/2” free-field microphone located 0.5 m from the phone screen, measurements were made in 10-degree increments horizontally around the phone and vertically at 0 and $\pm$25 degrees.

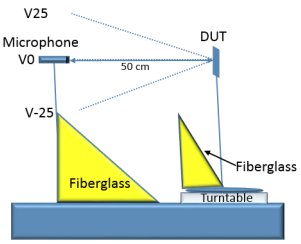
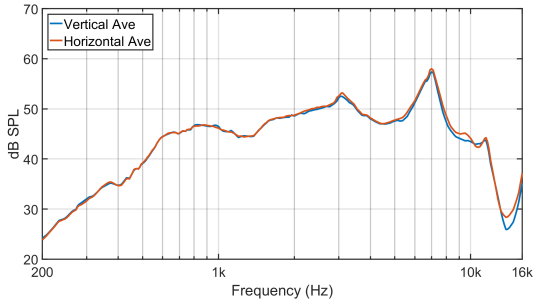


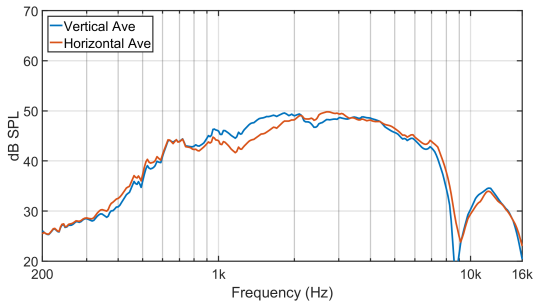
Fig. 1: Diagram of the measurement setup in the anechoic chamber.

The resulting data was weighted vertically by a fixed amount and horizontally to favor the forwardmost regions of the phone using a Gaussian weighting curve. A point-by-point average was then made of the weighted data to calculate a spatially-averaged frequency response. This spatial averaging helped to better characterize the resonance behavior as opposed to acoustic interference and yielded a smoother curve as shown in previous loudspeaker research [11]. The results of these averages in both the vertical and horizontal orientations playing back a stereo multitone file are shown for each phone in Fig. 2.

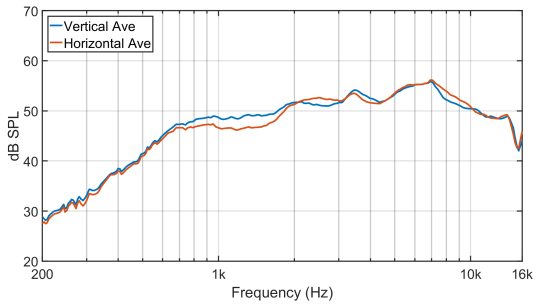
(a) Phone A



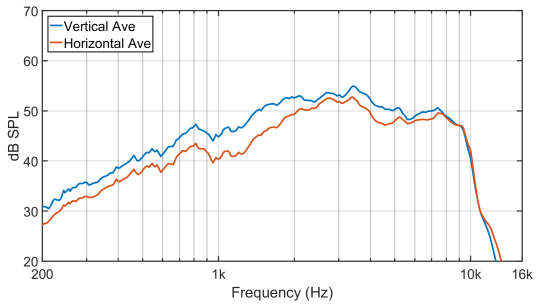
(b) Phone B



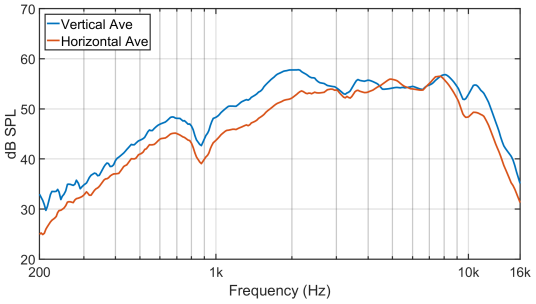
(c) Phone C



(d) Phone D



(e) Phone E



(f) Phone F

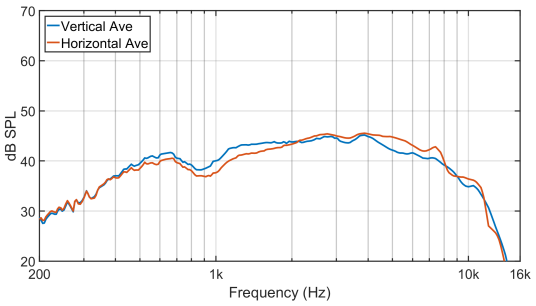


Fig. 2: Spatially-averaged stereo multitone measurements of each phone in the horizontal and vertical orientation at full volume. Note that the level shown in dB cannot be compared to those of a log sweep because the measurement is a power sum of the individual contributions of each sine wave. This means that the levels appear much lower in the graphs than they would for a traditional sine wave sweep.

Noise and nonlinear distortion components were also calculated from the measurements across the frequency playback range. Plots were made of each of the phones on-axis as shown in Fig. 3. Percentages of even, odd, and total distortion were calculated as well as total playback level in unweighted dB SPL.

2.3 Listeners

Eleven employees of Samsung Research America participated as listeners in the study. They ranged in age from 25 to 56 with a median age of 37 (SD=11.2). All listeners were screened for normal audiometric hearing and had participated in previous listening experiments.

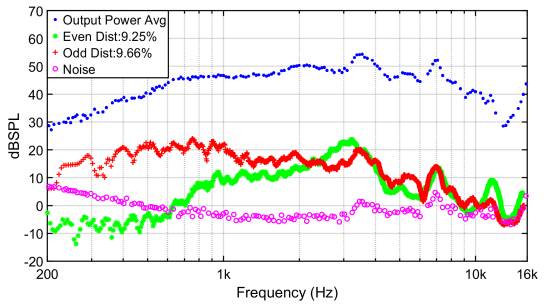


Fig. 3: Example of the on-axis multitone measurement of Phone C including noise, and both even and odd distortion components.

Nine of the listeners were considered trained based on their consistency in past listening tests calculated using their F_l scores [12].

2.4 Test Design

Listening tests were administered using the Samsung Listening Test Software (SLTS), which controlled turntable rotation, randomization of phone playback order, and results storage. The tests took place in Samsung Audio Lab's Small Listening Room [13]. The phones were mounted perpendicular to the turntable on its outer edge using phone clips, which were modified to reduce reflections.

There were four test cases evaluated throughout testing: vertical orientation level-matched (TC1), vertical orientation at maximum volume (TC2), horizontal orientation level-matched (TC3), and horizontal orientation at maximum volume (TC4). Each test case was blocked into three separate sessions in which listeners compared four phones over the course of six trials (three songs, one repeat). This structure allowed for direct comparison of the results for all six of the phones while minimizing listener fatigue by shortening the test sessions to around 25 minutes each. The order in which each listener completed the test sessions was randomized to minimize the impact of learning effects. In total, each listener completed 12 test sessions: three per test case as shown in Table 1.

Table 1: The structure of the sessions for each test case.

Session	Phone			
1	C	B	A	D
2	C	B	E	F
3	A	D	E	F

Listeners were seated 0.5 m away from the phones and were allowed to lean in to listen to the phones closer, since typical viewing/listening distance varies from listener to listener. An acoustically-transparent blackout curtain was located in between the listener and the turntable to prevent sighted biases. A diagram of the setup is shown in Fig. 4.

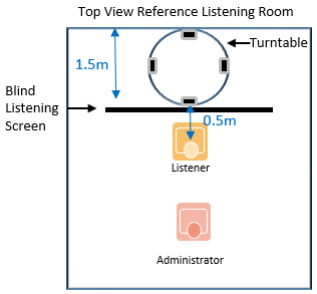


Fig. 4: Top view of the test setup in the small listening room.

Using a Samsung tablet loaded with SLTS, listeners selected from four buttons labeled "A"-"D" to hear the next phone. After selection, the test administrator switched which phone was playing audio based on commands given by the testing software. Each phone was loaded with randomly ordered playlists and the playback was controlled remotely by the administrator using tablets loaded with remote-access software [7][8].

After hearing each phone, listeners left a preference rating for each on a scale from 0 to 100 ("Extreme Dislike"-"Extreme Like"). They could also leave comments by selecting from a predefined list sorted by attribute (Bass, Midrange, Treble, Dynamics, Spatial). Once the task was completed, they pressed a "submit" button to go to the next trial and the phone order and

song selection were randomly reassigned. The interface for the tablet software is shown in Fig. 5.

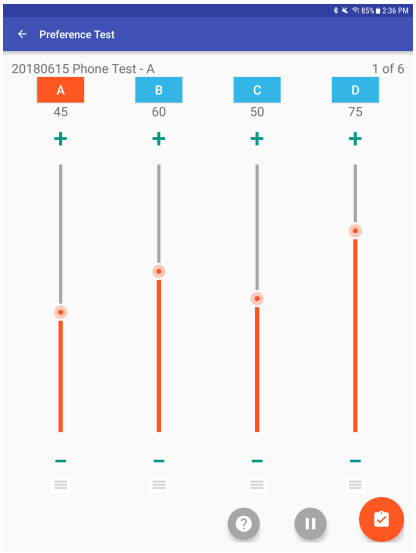


Fig. 5: The tablet interface for the Samsung Listening Test Software (SLTS).

2.5 Programs and Level-Matching

Three songs from different genres were selected as program material as shown in Table 2. The songs were analyzed for frequency (Fig. 6) and dynamic content, and were level-matched to -15 LUFS [14] for the maximum level test cases. For the level-matched sessions, the phones were initially level-matched using pink noise and an SPL meter at the seating position. Each phone was matched to a moderate 64 dB SPL(A) using the volume buttons on the phone. Since the volume buttons on most of the phones made very coarse adjustments, these levels were further refined by slightly adjusting the playback levels of the audio files. After listening to the initial results of this process, a further step was taken of playing back the Steely Dan clip over all the phones, recording each at the seating position and adjusting the files further based on the LUFS level. It should be noted that despite the extensive level-matching process, due to the difference between each phone's DSP processing, some listeners found that certain phones sounded too quiet or too loud on certain tracks after level-matching as is shown later in the results section.

Table 2: Summary of audio Programs used in testing.

Artist / Genre	Song/Album/Label
Mark Ronson and Bruno Mars (MR) / Pop Funk with Male Vocal	Uptown Funk / Uptown Special/ RCA 2014 CD
Sierra Hull (SH) / Bluegrass with Female Vocal	All Because of You/ Daybreak / New Rounder 2011 CD
Steely Dan (SD) / Jazz Rock with Male Vocal	Cousin Dupree / Two Against Nature / Giant 2000 CD

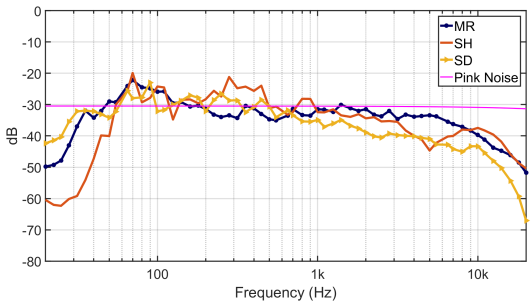


Fig. 6: Average power spectral density of the left and right channels of the audio programs used with 1/6-octave smoothing applied. The curves were weighted for readability with pink noise as the neutral horizontal reference.

3 Results

3.1 Preference Ratings

The results data for each separate test case as well as for all four test cases combined, grouped by level, and by phone orientation was analyzed. As is apparent in Fig. 7, the mean ratings for most of the phones varied noticeably across the test cases. Phone B rated the most consistently across tests while Phone D rated quite low on TC3 compared to the other phones tested. This may have been an interaction with the level-matching of phone orientation as listeners tended to comment that this phone sounded "Too Quiet" in this test case.

When the combined means are observed (Fig. 8), the rank order of the phones from highest rated to lowest

rated is as follows: C (mean= 47.0), F (mean= 45.5), D (mean= 43.6), B (mean= 39.9), E (mean= 35.8), and A (mean = 31.9). Based on the results of a Bonferroni test, Phones C and F statistically tied, Phones F and D tied and all the other means were significant.

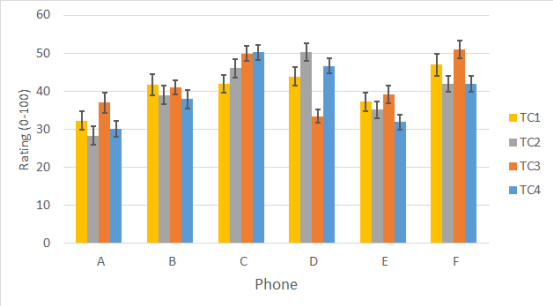


Fig. 7: Means and 95% confidence intervals for all phones in all test cases.

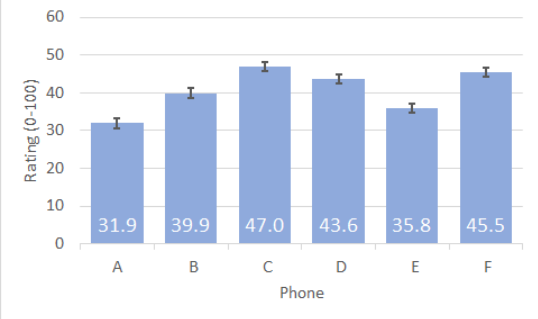


Fig. 8: Means and 95% confidence intervals for all test cases combined.

3.2 Effects and Interactions

A repeated measures ANOVA was conducted on each test case to assess the main simple effects and interactions. The fixed within-subjects factors were phone (6 levels), track (3 levels) and observation (4 levels). All tests were run at a 5% significance level. Each data set was checked for outliers using box plots, normality of the residuals using normal Q-Q plots, and sphericity using a Mauchly's Test. The results of a Mauchly's test proved significant for the track by phone interaction in TC2 and a Greenhouse-Geisser correction was applied to these results.

As shown in Table 3, the phone factor was a simple main effect in all four test cases. Additionally, the interaction between track and phone was significant in TC1 and TC2. In both of these test cases, the phones were vertically oriented. Further analysis of this interaction revealed that it was predominantly isolated to Phone E which rated significantly lower on the Mark Ronson track than on the other tracks tested as shown in Fig. 9.

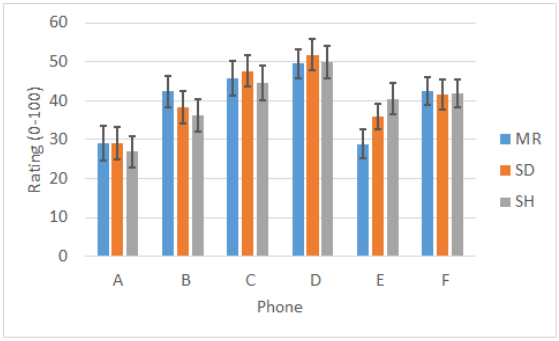


Fig. 9: Means and 95% confidence intervals by track by phone for test case 4. Phone E rated significantly lower on the Mark Ronson Track.

In order to evaluate the effects of playback level on phone playback, two repeated measures ANOVA tests with the data grouped by the orientation were performed after checking for assumptions. The fixed within-subjects factors were volume (2 levels), phone (6 levels), track (3 levels) and observation (4 levels). In the horizontal tests, level was a simple main effect ($F(1,10) = 5.7$, $p < 0.05$). For both the vertical and horizontal orientations, there were interactions by level by phone ($F(5,50) = 3.05$, $p < 0.05$ vertically and $F(5,50) = 11.18$, $p < 0.001$ horizontally).

Additionally, in order to evaluate the effects and interactions of phone orientation, another two repeated measures ANOVA tests with the data grouped by playback level were performed after checking for assumptions. The fixed within-subjects factors were orientation (2 levels), phone (6 levels), track (3 levels) and observation (4 levels). At full volume, orientation was not a significant factor. For the level-matched tests, there was an interaction by phone by orientation ($F(1.699, 16.994) = 5.562$, $p < 0.05$). Note that this result had a Greenhouse-Geisser correction applied as it failed the Mauchly's test.

3.3 Comments

The comments of all six phones across the test cases were compiled to observe trends (Table 4). The comments were sorted by the number of listeners who left

Table 3: Summary of significant effects and interactions by test case.

Test Case	Phone	Phone*Track
1: Vertical, Level-Matched	$F(5,50) = 4.3$, $p < 0.01$	$F(10,100) = 2.36$, $p < 0.05$
2: Vertical, Full Volume	$F(5,50) = 10.4$, $p < 0.001$	$F(3.755,37.549) = 5.5$, $p < 0.01$
3: Horizontal, Level-Matched	$F(5,50) = 9.497$, $p < 0.001$	
4: Horizontal, Full Volume	$F(5,50) = 10.03$, $p < 0.001$	

each comment and number of repetitions. For the top 15 comments, the percentage of listeners that left the comment and overall percentage of that comment for each product was recorded. Correlations were then calculated between this data and the combined means (all four test cases) for each phone. The notable results of these calculations are shown in Tables 5 and 6.

There was a strong negative correlation associated with the phrase "Too Little Bass" in both methods of analysis. The more often this phrase was used, the lower the phone tended to be rated. In contrast, there was a positive correlation between the comment "Clear" in one method of analysis. Surprisingly, the comments "Distorted Bass," "Too Little Treble," and "Needs More Bass Extension" also had positive correlations with the overall means. Since the listeners had many issues to address in commenting on these phones, the authors believe that these issues were considered less objectionable. It is also important to note that the comment "Distorted Midrange" was commonly left for all six of the phones, but did not have any correlation with rating.

The comments were also observed split up into each test case, and it became apparent that the level-matching on Phone D appeared to seem perceptually lower than the other phones in the horizontal level-matched test case. The level-matching method used may not have been as effective for Phone D in this configuration due to its custom DSP processing.

Notably, there was also a very strong negative Pearson correlation ($r = -0.972$, $p < 0.001$) between the overall number of comments left and the overall mean for each product. The more comments that were left, the lower the product was rated. This is likely due to negativity bias, since it has been shown repeatedly that humans are more affected by unpleasant experiences and more likely to comment on them [15].

4 Discussion

4.1 Phone Ratings and Design

Phones B, C, D, and F were all "stereo" phones which used a forward-firing receiver and a downward-firing speaker for music playback. Interestingly, the two lowest rated phones, A and E, had different configurations. Phone E used a forward-firing receiver and forward-firing speaker, while Phone A used a single downward-firing speaker for music playback. Considering this, it is possible that Phone A and E stood out as inherently different when placed in context with the other four devices.

The number of drivers may also have been a factor. Phone A, which only utilized a single downward-firing driver, rated the lowest of all phones tests. Despite its single driver it was not the quietest phone evaluated. That said, it did have stronger resonant peaks than any of the other devices tested, which likely influenced the ratings.

The physical size of the device tested did not appear to affect the overall preference rating. Intuitively it would seem likely that larger phones would have larger back volumes for the transducers, potentially allowing for better low-frequency production. Contrary to this idea, Phones A, B, and C were all similarly sized and rated quite differently in regards to overall preference. In real-world examples, acoustic modules of different mobile phones are usually similarly sized regardless of the actual size of the device.

4.2 Importance of Frequency Response

Aspects of each phone's frequency response were analyzed for comparison with the overall means. The smoothness of each phone's frequency response was evaluated by fitting a 3rd-order polynomial through the curve from 800 Hz to 10 kHz. The coefficient of determination, R^2, of this polynomial was then recorded

Table 4: Top ten comments of all test cases combined listed with (number of listeners who left comment, overall repetitions).

A	B	C	D	E	F
Honky (10, 83)	Bright (11, 95)	Muffled (10, 89)	Too Quiet (11, 41)	Too Little Bass (11, 152)	Too Quiet (11, 53)
Too Little Bass (9, 94)	Needs More Bass Ext (11, 73)	Too Little Treble (10, 43)	Distorted Midrange (10, 39)	Harsh (11, 107)	Distorted Midrange (10, 60)
Distorted Midrange (8, 51)	Harsh (11, 67)	Distorted Midrange (10, 26)	Recessed Midrange (10, 31)	Bright (11, 88)	Clear (10, 42)
Muffled (7, 65)	Distorted Midrange (11, 39)	Hollow (9, 25)	Harsh (10, 29)	Too Much Treble (11, 62)	Distorted Bass (10, 41)
Hollow (7, 53)	Too Much Treble (10, 51)	Honky (9, 15)	Too Little Bass (9, 74)	Distorted Midrange (11, 39)	Bright (9, 23)
Needs More Bass Ext (11, 73)	Aggressive (10, 30)	Needs More Bass Ext (11, 73)	Bright (9, 37)	Needs More Bass Ext (11, 73)	Aggressive (9, 17)
Nasal (6, 51)	Too Little Bass (9, 94)	Clear (8, 26)	Muffled (9, 30)	Aggressive (9, 48)	Recessed Midrange (9, 17)
Right-Heavy (6, 35)	Distorted Treble (9, 33)	Harsh (8, 19)	Honky (9, 25)	Distorted Treble (9, 32)	Distorted Treble (9, 14)
Too Little Treble (4, 31)	Compressed (9, 19)	Aggressive (8, 15)	Aggressive (9, 25)	Clear (9, 18)	Too Little Bass (8, 61)
Recessed Midrange (4, 29)	Clear (8, 21)	Too Little Bass (7, 45)	Distorted Treble (9, 18)	Too Much Midrange (8, 27)	Harsh (8, 24)

Table 5: Most notable correlations between percentage of listeners who left each comment and overall rating.

Comment	Spearman
Distorted Bass	0.880, p = 0.021
Too Little Bass	-0.820, p = 0.046
Too Little Treble	0.754, p = 0.021

Table 6: Most notable correlations between percentage of comment repetitions and overall rating.

Comment	Pearson
Clear	0.896, p = 0.016
Distorted Bass	0.886, p = 0.019
Needs More Bass Extension	0.829, p = 0.042

to represent the goodness-of-fit of the curve. This was done for both the vertical and horizontal orientation and the results were averaged. When a Pearson correlation was calculated between these values and the overall preference means, there was a strong correlation between the two ($r = 0.818$, $p < 0.05$). This indicates that phones with smoother frequency responses were more preferred overall.

Lower frequency extension was also calculated for each phone, based off of the -6 dB down point from the mean level (800 Hz to 7 kHz) starting from 800 Hz, 1 kHz, and 2 kHz. Surprisingly, there were no strong correlations between these results and preference overall or in individual test cases. That said, the phone with the best lower-frequency extension, Phone F, was still one of the most preferred phones.

4.3 Influence of Loudness

While one of the main goals in tuning a phone is to increase the playback level of the device as much as possible, the results of these listening tests actually question whether that has the intended effect of improving listener preference (see Fig. 10 for levels). Correlations between max SPL of each smartphone and its mean rating in each of the full-volume listening tests returned only small to moderate correlations which were not statistically significant. While comments indicated that listeners could identify a phone that sounded quieter, in the full-volume tests they did not strongly penalize Phone F, the quietest phone.

It is important to remember that these tests were run in a quiet listening room rather than in an environment with higher background noise like a busy street or loud coffee shop. It these contexts, loudness performance is likely more important than fidelity.

There were, however, significant differences between the level-matched test cases and maximum-volume test cases. This is unsurprising considering that most mobile phones have different tunings at each volume setting, which could result in the phone sounding very different spectrally at two different volume settings.

4.4 Influence of Phone Orientation

For the maximum-volume test cases, phone orientation was not a significant factor. In contrast, orientation may have affected the preference ratings of phones in the level-matched tests. The phone which saw the most change in overall rating was Phone D. The multitone measures for Phone D demonstrate some broadband

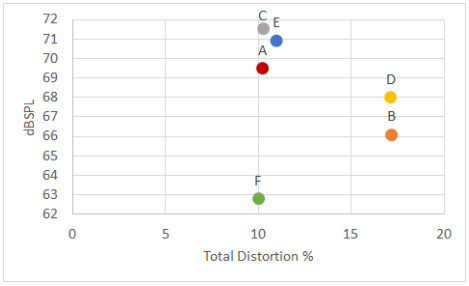


Fig. 10: Max SPL and total distortion percentage for each phone playing back a stereo file.

differences between the vertical and horizontal spatial averages. The horizontal average has less energy below 2 kHz than the vertical average, which may account for some of the preference difference.

4.5 Distortion Correlation

There was no strong correlation between maximum playback level and total distortion levels for each phone, as shown in Fig. 10. Correlations between total distortion and preference rating were computed across all the maximum-volume tests and no significant correlations were found as shown in Fig. 11. The total distortion values measured for Phone B and D were markedly higher than the other phones tested, but it is difficult to say if that directly affected their preference ratings. In looking at the comment data, there was no dramatic difference between the number of comments about distortion values for these phones compared to the other devices tested. In general, distortion comments about the bass and the midrange were common across all the devices tested.

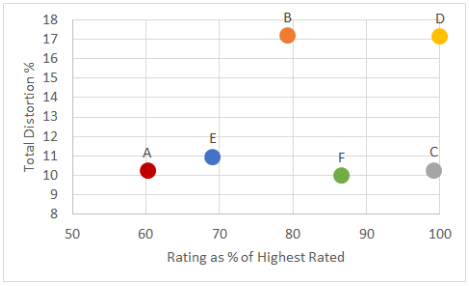


Fig. 11: Total distortion percentage by preference rating as a percentage of highest average rating on the full-volume test cases.

4.6 Influence of Program Material

There was a program interaction between the Mark Ronson track and Phone E in the vertical phone orientation tests. From 200 Hz up, this track is the most neutral of the three used in the test. The multitone measurements of Phone E reveal that it has a low-Q resonance from about 1 kHz to 3 kHz which is not seen in the horizontal orientation. The combination of a very neutral track and a prominent low Q resonance was likely more audible on this track than on the other two tracks, which had more colorations.

5 Future Work

Several follow-up experiments will be run to expand upon this research and address some of its shortcomings. More experiments addressing the listening position of the smartphones will be performed which will include measuring the effects of different handheld listening positions and evaluating the performance difference between tabletop and handheld states. Further research will also be pursued in regard to the effects of nonlinear distortion on the perception of mobile phone audio quality. Additionally, more study of methods to level-match smartphones for subjective evaluation will be pursued.

6 Conclusion

A series of objective and subjective measurements were performed on six high-end smartphones to evaluate their performance and observe trends in what listeners prefer and why. Spatially averaged multitone measurements were made of all the devices in both the vertical and horizontal phone orientations. From the on-axis measurements, distortion components, noise, and overall SPL were also calculated. The phones were subjectively evaluated in a series of double-blind experiments which included four test cases in which the phones were mounted vertically or horizontally playing back at maximum volume or level-matched.

A summary of the findings are as follows:

1. Phones with smoother frequency responses tended to receive higher ratings.

2. Listeners left more comments for the lower-rated phones.

3. Listeners tended to describe phones with higher ratings as "Clear" and phones with lower ratings as having "Too Little Bass."

4. There were no obvious correlations between preference ratings and distortion measurements, but listeners tended to comment that all the phones had distortion in the bass and midrange.

5. There was an interaction between playback level and phone preference ratings.

6. Phone orientation only interacted with phone ratings in the level-matched tests.

7. Device size was not an important predictor of overall audio quality preference.

8. Higher maximum SPL levels did not correlate with higher listener preference ratings in the maximum volume tests.

Acknowledgments

Thanks to Samsung Research America who fully supported this work. Thanks to all the participants of the Audio Lab Listening Team. Special thanks to Glenn Kubota and Felix Kochendoerfer.

References

[1] Pew Research Center, "Mobile Fact Sheet," `http://www.pewinternet.org/fact-sheet/mobile/`, Feb 5, 2018, accessed: May 8, 2018.

[2] Fastl, H. and Zwicker, E., *Psychoacoustics: Facts and Models (Springer Series in Information Sciences)*, Ch, 4. Springer Berlin Heidelberg, 2007.

[3] ITU-T Recommendation P. 862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunications Union*, 2001.

[4] ITU-T Recommendation P. 835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *International Telecommunications Union*, 2003.

[5] ITU-R Recommendation BS. 1387-1, "Method for objective measurements of perceived audio quality," *International Telecommunications Union Radiocommunication Assembly*, 2001.

[6] ITU-R Recommendation BS. 1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunications Union Radiocommunication Assembly*, 2015.

[7] Blue Atlas Technology, "Tango Remote, Music Player with Remote Control," 2018.

[8] FK33, "Phone to Tablet Remote," `https://play.google.com/store/apps/details?id=fk33.remote&hl=en_US`, 2018.

[9] Y. Bababekova, M. Rosenfield , J.E. Hue , R.R. Huang, "Font size and viewing distance of handheld smart phones." *Optometry and Vision Science., 88(7):795-7.*, Jul 2011.

[10] Brunet, P., Decanio, W., Banka, R., and Yuan, S., "Use of Repetitive Multitone Sequences to Estimate Nonlinear Response of a Loudspeaker to Music," *presented at the 143rd AES Convention, Audio Eng. Soc., preprint 9827*, October 2017.

[11] Toole, F., *Sound Reproduction: Loudspeakers in Rooms*, pp. 373-383. Focal Press, 2008.

[12] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation*, John Wiley & Sons, Ltd, 2006.

[13] McMullin, E., Celestinos, A., and Devantier, A., "Environments for Evaluation: The Development of Two New Rooms for Subjective Evaluation," *presented at the 139th AES Convention, Audio Eng. Soc., preprint 9460*, October 2015.

[14] ITU-R Recommendation BS. 1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," *International Telecommunications Union Radiocommunication Assembly*, 2015.

[15] Rozin, P. and Royzman, E. B., "Negativity Bias, Negativity Dominance, and Contagion," *Personality and Social Psychology Review, Vol. 5, No. 4, 296–320*, 2001.