Perception of Nonlinear Distortion in Music Signals Reproduced by Microspeakers

Pei Yu, Yong Shen*, Ziyun Liu
Institute of Acoustics, Key Laboratory of Modern Acoustics of Ministry of Education, Nanjing University, Nanjing, 210093, China

A lot of work has focused on compensating nonlinear distortions of the microspeaker under large excitation, yet it is unclear at which level the effect of nonlinear distortion is imperceptible or not annoying. In this study virtual listening tests were performed to evaluate the deterioration of sound quality in music signals with different levels of nonlinear distortion by a microspeaker. Binaural recordings of the music played by the microspeaker were made at different voltage levels, adjusted to the same RMS power, and afterwards reproduced by a low distortion reference headphone. The "double-blind triple-stimulus with hidden reference" method was used in the listening tests. Listeners were required to not only discern the distortion, but also rank the severity of the distortion. The effects of several factors including subject, program material and distortion level on the perceived distortion are discussed.

Keywords: Nonlinear distortion, microspeaker, virtual listening test.

1 Introduction

With the widespread use of microspeakers in smart phones, laptops and other portable devices in recent years, people are used to listening to music reproduced by these tiny drivers. Due to the limited size and to pursue high output power, microspeakers usually work under large signals[1, 2]. In this case, nonlinear distortions rise considerably and deteriorate sound quality[3]. Many efforts have been put into reducing the distortions[4-6], which arouse interest in studying how people perceive the deterioration of sound quality caused by nonlinear distortions in microspeakers.

Research of the audibility of distortion in loudspeakers has a long history[7-15], but few of them are focused on microspeakers. In early times, most research[7-9, 11-13] focused on determining the threshold for just detectable distortion. Cabot[7] examined several factors affecting the distortion audibility and pointed out that on account of these factors, it is impossible to establish a single number which represents the minimum audible distortion. Klippel[14] proposed a speaker auralization method based on parameter identification to assess the linear and distortion components in real time. Recently, Temme et al.[15] studied the distortion audibility in headphones, it turned out that it was difficult to hear audible distortion in the headphones and nonlinear distortion in headphones was not a significant factor affecting the sound quality.

In this study, based on binaural recording and reproduction, a virtual listening test was performed to assess the perception of different levels of distortion in microspeaker. Compared with traditional live listening test, the virtual listening test is more effective and efficient. It can better control nuisance parameters in the test, and help create an equal listening condition for every test subject[16-18]. Discussions of several factors affecting the perception of distortion are also presented.

2 Objective Measurements of the Microspeaker

A typical microspeaker of size 1.1×1.5 cm^2 was used in this research. The working range of this microspeaker is 100-20 kHz according to the specification.

To have a general impression of the nonlinear performance of the microspeaker, we measured the frequency response and harmonic distortion in an anechoic chamber. Figure 1 shows the on-axis frequency responses of the microspeaker measured at a distance of 10 cm. The measuring level varied from 0.4 V to 1.2 V in 0.2 V increments. Clearly there are huge roll-offs in the low frequency range. The frequency response at 300 Hz is about 10 dB below the sound pressure level averaged over a bandwidth of one octave in the region of maximum sensitivity. Nonlinear amplitude compression also appears as the measuring level increases in equal increment while the uplift of the output amplitude shrinks.

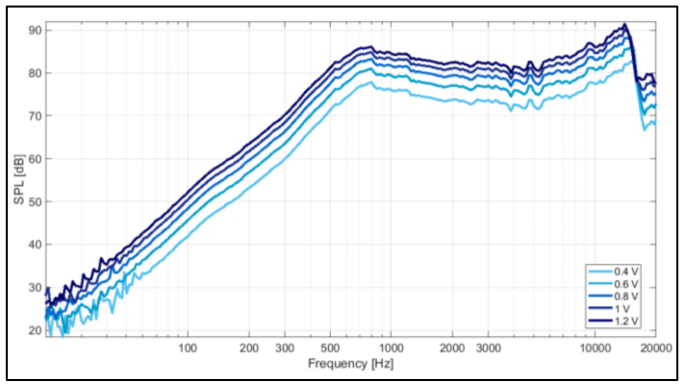

Figure 1. On-axis frequency response curves at different levels.

Figure 2 depicts the total harmonic distortion (THD) measured at the same levels, ranging from 0.4 V to 1.2 V. Distortions are obvious in the low frequency range. Even for a small stimulus level of 0.4 V, the THD can be up to 16% around 300 Hz. For higher stimulus level like 1.2 V, the peak value of THD increases to an amazingly 57% at 200 Hz. Considering that most music signals contain extended low frequency content, especially in the range of 200-300 Hz, distortions in this range shall have huge effects on the perceived sound quality.

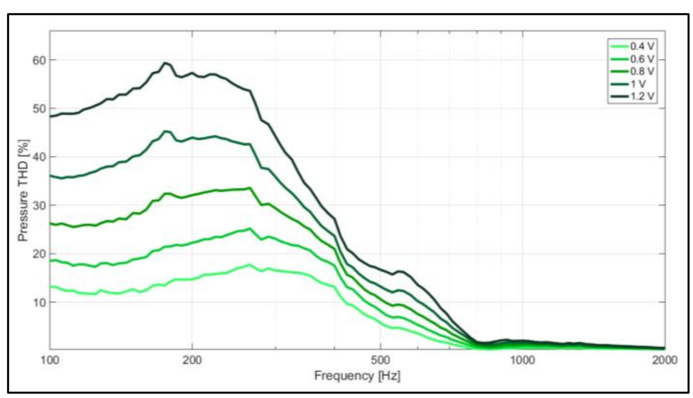

Figure 2. Total harmonic distortion curves.

3 Virtual Listening Tests

This part describes the setup and procedures of the listening experiments.

3.1 Program Selection

Table 1 lists the four program selections used in the listening test. The first two programs were reliable test signals for identifying the spectral distortions according to a previous study[19]. Another two programs were chosen from the "2016 GRAMMY Nominees" considering that most consumers prefer pop music when listening through their mobile phones. Each program was clipped to 28-38 s and applied with gains normalized to the same average RMS power.

Table 1. Program materials used in tests.

Code	Artist	Track
TC	Tracy Chapman	"Fast car"
JT	James Taylor	"That's why I'm here"
TS	Taylor Swift	"Blank space"
SH	Sam Hunt	"Take your time"

As Figure 1 shows, the microspeaker we used has poor low frequency performance. To avoid the effect of possible harmonic components of low frequency content in the program spectrum, all the four programs were preprocessed by high-pass filter at 100 Hz.

3.2 Binaural Recording of the Microspeaker Playback

Binaural recordings were used in this virtual listening test since a previous study[17] showed that binaural recordings could produce the most accurate and reliable preference ratings compared with mono/stereo recordings.

In this study, binaural recordings were made by the Head and Torso Simulator (HATS, Brüel & Kjær Type 4128C) in the anechoic chamber.

The microspeaker was mounted at a height equal to the ear of the HATS with a horizontal distance around 1 m. The RMS voltage of program playing was set to 0.1 V to 1.5 V in 0.1 V increments in 15 separated recording sessions, and output signals from the HATS were recorded by a Brüel & Kjær PULSE Type 3120A. All the recordings were then adjusted to equal RMS power. In this way, the distinguishing characteristic between recordings was limited to the amount of nonlinear distortion besides the program material.

3.3 Headphone Equalization and Playback

A Beyerdynamic DT 990 Pro circumaural headphone was chosen as the reference headphone due to its relatively smooth frequency response, wide bandwidth, and low distortion. To accurately reproduce the recordings, the transfer function of the ear canal of HATS and the frequency response of the reference headphone should be compensated[16, 17]. The solid curves in Figure 3 depict the measured frequency responses of the left and right channels of the headphone based on an average of several resets of the headphone on the HATS. The dashed curves are the equalization function applied to flatten the frequency responses[17].

The playback level was calibrated to 72 dBA using simulated program signal as IEC 60268-1: 1985. The simulated program signal was scaled to the same average RMS power as the 4 program materials.

3.4 Listening Panel

The listening panel was composed of 9 listeners, including 6 males and 3 females, with ages ranging from 22 to 32. All of the test subjects reported no hearing damage.

The listeners were categorized into two groups: Experienced listeners and naïve listeners. All the experienced listeners had reached skill level 8 or higher in Harman's "How to Listen" listener training software[20].

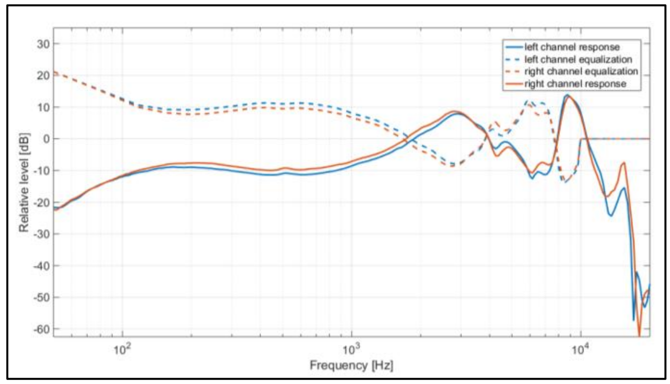

Figure 3. Frequency responses of the reference headphone and equalization.

3.5 Test Procedure

A pilot listening test was done by two experienced listeners to evaluate the distortion levels of the 15 separated recording sessions. Recordings under low voltages suffered from the unavoidable background noises, which may distract listeners from assessing the distortion. As a result, the recordings recorded at 0.4 V were judged as reference signals with ignorable background noise and the least perceivable distortion. Others recorded at 0.6 V, 0.8 V, 1 V and 1.2 V were chosen as test signals for they showed distinguishing distortions. The two listeners didn't participate in the subsequent listening tests.

The "double-blind triple-stimulus with hidden reference" method was used in the listening tests refer to ITU-R BS.1116-1. Custom software was written to administer the test and provide a graphic user interface (GUI) for the listener. The GUI is shown in Figure 4. In a trial, the reference signal is always available as stimulus "Reference", while the test signal is randomly assigned to "A" or "B" and the other is also the reference signal (known as the hidden reference). The subject may freely switch between the three sound files to compare and then identify which one is distorted between "A" and "B". Rating of the severity of the distortion on a continuous five-grade impairment scale should also be given. The grading scale is shown in Table 2.

A whole listening test for one subject was divided into two successive phases: the training phase and the evaluation phase. The GUI was used in both phases, but with different audio test files. The training phase consisted in explaining to the subjects how to use the GUI and what they are expected to do. Examples of the four program files were used in this session to help subjects get used to the music. In the evaluation phase, there were 16 trials (4 programs and 4 distortion levels) for each subject. The whole listening duration ranged from 18 mins to 32 mins for different subjects.

Table 2. Grading scale on distortion.

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

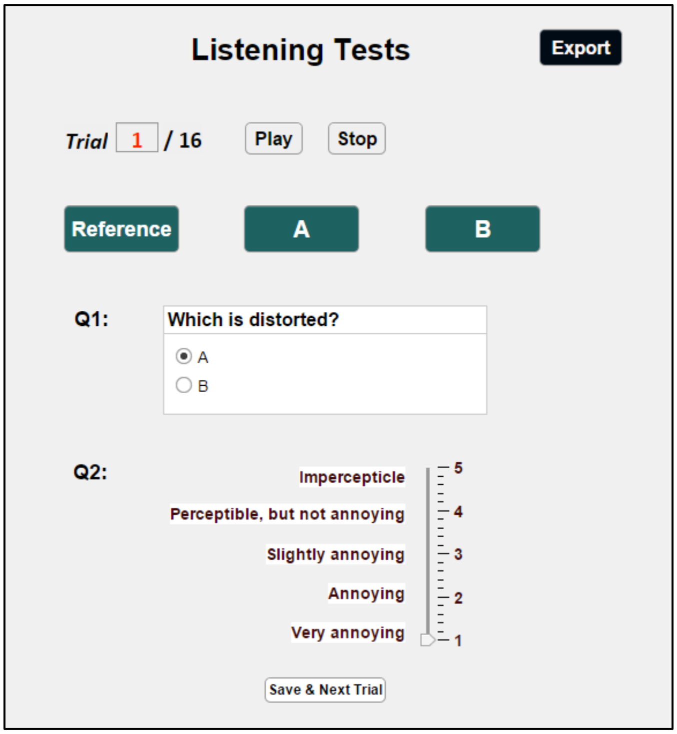

Figure 4. GUI used in the virtual listening tests.

4 Results

4.1 Evaluation of Listener Expertise

T-tests were performed to evaluate the reliability of the listening panel. For each trial, the double-blind triple-stimulus hidden-reference method provided two grades. One was default 5 points for the signal judged as hidden reference, and the other was the rating given by the subject for the other signal judged as distorted. Subtracting the grade for the actual hidden reference from the grade for the actual test signal, the data thus obtained were subjected to a one-sided t-test, to assess the likelihood that the mean of the distribution for each subject is zero. If the subject had difficulty identifying the distorted signal, there would be both positive and negative difference grades in the data, which could tend to balance each other out on the average, resulting a zero mean of the distribution and the null hypothesis would be accepted. On the contrast, if the null hypothesis was rejected for a given subject, we may conclude that the data for that subject originates from a distribution with a mean greater than zero in a negative direction. In this case, the subject shall be considered to be able to correctly distinguish the distorted signal from the hidden reference at a given level of confidence.
Table 3 illustrates the statistics of the listening panel. The percentage of the subject correctly differentiated the hidden reference from the distorted signal was also calculated, shown in Table 3 as Accuracy.

Table 3. Statistics of the listening panel.

Subject	Experience	Accuracy	*t*	*P* Value
1	Experienced	75%	-2.477	0.0256
2	Experienced	75%	-2.645	0.0184
3	Naïve	87.50%	-3.971	0.0012
4	Experienced	50%	-0.734	0.4743
5	Naïve	75%	-2.126	0.0505
6	Experienced	100%	-10.90	<0.001
7	Experienced	81.25%	-3.573	0.0028
8	Experienced	93.75%	-5.604	<0.001
9	Naïve	43.75%	-0.324	0.7502

Note: Significance level 0.05.

It could be observed that Subject 4, 5 and 9 showed poor performance in the t-tests with *p* values higher than the significance level, indicating they were unreliable in assessing the severity of distortion. The data of these three listeners would be excluded from further analysis.

4.2 ANOVA Tests

Analysis of Variance (ANOVA)[21, 22] was performed to evaluate the effect of different factors on listeners' perception of distortion. Table 4 shows the results of three single factor ANOVA tests. The dependent variable was rating for the severity of distortion, and the fixed independent variable was subject, program and level separately.

Table 4. The results of single factor ANOVA.

Factor	Sum of Squares	df	Mean Square	*F*	Sig
Subject	26.652	5	5.330	5.568	<0.001
Program	20.380	3	6.793	6.762	<0.001
Level	16.684	3	5.561	5.323	<0.001

All of the three factors were significant. There was also a significant interaction influence between subject and program (F=1.622, p=0.089), meaning that for different listeners the perception of nonlinear distortion in certain programs may differ a lot.
The results of Least Significant Difference (LSD) comparison of each program pair and level pair are shown in Table 5 and Table 6.

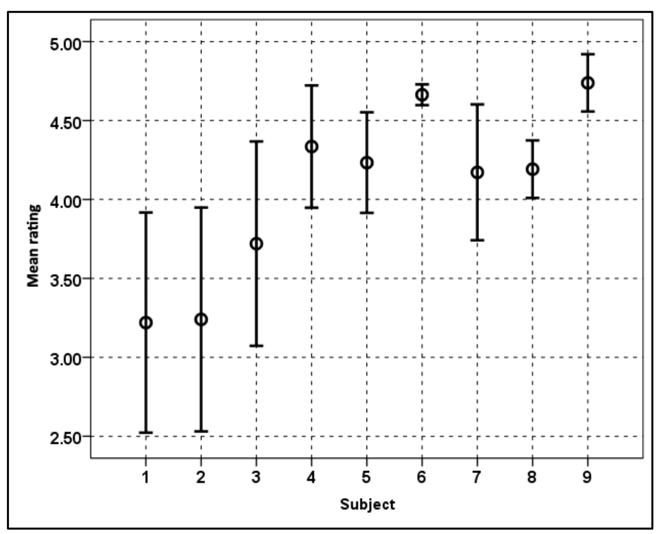

Figure 5. Mean ratings of different subjects.

Table 5. LSD comparison of programs.

(I) program	(J) program	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
TC	JT	-0.103	0.2207	0.641	-0.542	0.336
	TS	-1.177*	0.2207	0.000	-1.616	-0.738
	SH	-0.436	0.2207	0.052	-0.874	0.003
JT	TS	-1.073*	0.2207	0.000	-1.512	-0.634
	SH	-0.332	0.2207	0.136	-0.771	0.107
TS	SH	0.741*	0.2207	0.001	0.302	1.180

** The mean difference is significant at the 0.05 level.*

Table 6. LSD comparison of levels.

(I) level	(J) level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
0.6 V	0.8 V	0.438	0.2207	0.051	-0.001	0.877
	1.0 V	0.663*	0.2207	0.003	0.224	1.102
	1.2 V	1.157*	0.2207	0.000	0.7178	1.596
0.8 V	1.0 V	0.226	0.2207	0.310	-0.213	0.664
	1.2 V	0.719	0.2207	0.002	0.280	1.158
1.0 V	1.2 V	0.493*	0.2207	0.028	0.0545	0.932

** The mean difference is significant at the 0.05 level.*

The program TS showed significant difference from other three programs as illustrated in Table 5, while the other program pairs were not statistically differentiated. Table 6 presents that the recordings with highest distortion level, which was recorded at 1.2 V, were remarkably differentiated from any other level. There was also significant difference between 0.6 V level and 1.0 V level.

5 Discussion

5.1 The Effect of Subject

A former study[23] pointed out that notwithstanding several differences in their performance when conducting listening tests, experienced listeners had very similar preference with the naïve listeners. Similar discovery can be found in this study. Comparing the accuracy and *t*-test results of Subjects 1, 2 and 5 in Table 3, we may find that they had the same accuracy in the listening tests, but only Subject 5 was considered unreliable. This difference may relate to the subjects' different experience in participating in listening tests. Figure 5 plots the mean ratings of all subjects. As can be seen, Subject 5 tended to give higher ratings for distortion than Subjects 1 and 2, which means his difference data would be closer to zero, thus leading to poor *t*-test result. Compared with experienced Subject 1 and 2, Subject 5 was untrained and had little experience in judging the sound quality of loudspeakers. This agrees with the former study[23] that the trained listeners are more critical and also the most reliable listeners.

Interaction between subject and program also showed to be significant, which can be seen in Figure 6 where the subjects' ratings for different programs are depicted. Subject 6 rated all programs in a rather condensed range, although he correctly discriminated the distorted signal in all trials.

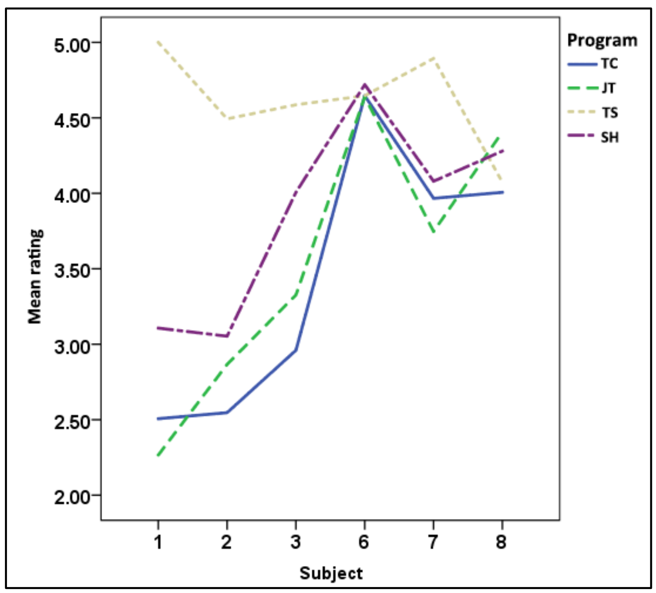

Figure 6. Mean ratings for different programs.

5.2 *The Effect of Program*

As Figure 6 illustrates, the audibility of distortion depends on the program materials. The listening panel graded program TS considerably higher than other programs. It can be illustrated as most of them had difficulty in distinguishing the distorted version of program TS from the hidden reference. Besides, listeners seemed to be more easily perceiving distortions in program TC and JT than in program SH and TS. The general order of listeners rating programs is TC<JT<SH<TS.

In order to determine whether is the spectral content of the program account for the difference, one-third octave spectrum analysis were made to the program recordings at 0.4 V, which were used as reference signals in the formal listening tests. The one-third octave spectral curves of the four recorded programs with headphone equalization are presented In Figure 7. A general similarity in their shapes can be seen. There are huge roll-offs below 500 Hz for all programs due to the limited low frequency reproduction of the microspeaker. Compared with other three programs, program TS has noticeably less content between 200 Hz and 800 Hz, and more high frequency content above 3000 Hz. As Figure 2 shows, most nonlinear distortions appear in low frequency due to the large displacement of voice coil. It can be safely assumed that the relative lack of low frequency content cause program TS to be the least to appear distortion.

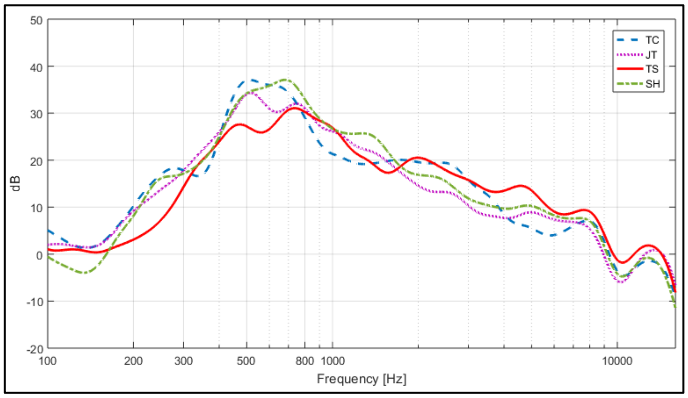

Figure 7. One-third octave spectra of the recorded programs after equalization.

Besides the spectral influence, there may be another factor concerning the temporal aspect of programs. Figure 8 compares the time waveform of the four original programs. Apparently, program TS and SH have much smaller dynamic range compared to program TC and JT. The dynamic range is measured as a base-10 logarithmic value of the difference between the maximum and minimum RMS amplitude. For program TC and JT, the dynamic ranges are separately 23.6 dB and 18.6 dB, while for tack TS the value is only 12.7 dB and for tack SH the value is even smaller, around 10.6 dB. Probably the small dynamic range stemming from highly compressed mixing process makes the program TS and SH sound distorted at the beginning, which also explains why these two programs are less susceptible to added distortions.

Combining both spectral and temporal factors, we believe that the program materials suitable for revealing nonlinear distortions have two characteristics: enough low frequency content and relatively large dynamic range. The first characteristic agrees with a former study[14] that for generating substantial nonlinear distortions, the audio signal should contain low frequency components of sufficient amplitude. The second characteristic means that the program should not be compressed too highly in the mixing process. The program TC and JT used in this study are typical examples.

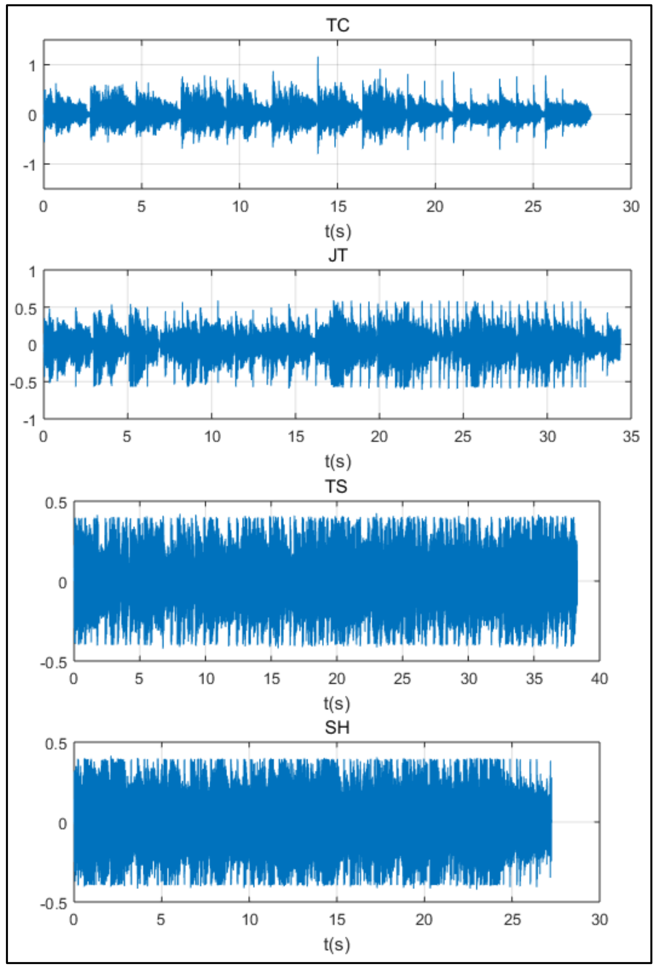

Figure 8. Time waveform of the four programs.

5.3 The Effect of Level

Figure 9 plots the relationship between programs and recording levels. An obvious downward trend of ratings could be seen in all programs except program TS. Generally, listeners regarded nonlinear distortion in programs recorded at lower levels as not annoying. As the recording level increased, they felt more annoyed and gave lower ratings. In this virtual listening test, recordings at 0.4 V were regarded as the reference signal with least perceptible distortions. By comparison, listeners considered nonlinear distortion in recordings at 0.6 V as nearly imperceptible and the average grade decreased to slightly annoying when the recording level increased to 1.0 V.

The THD metric is commonly used to assess the performance of nonlinear compensation method[6]. Combining the THD measurement results in Figure 2 with the subjective evaluation ratings in Figure 9, we may get a guidance for distortion compensation in some extent. For example, Figure 2 demonstrates that the absolute decrease of the THD peak value is about 23% when the stimulus level decreases from 1.0 V to 0.6 V. Hence the reduction value 23% may be used as a reference when evaluating the effect of compensating nonlinear distortion in such a microspeaker.

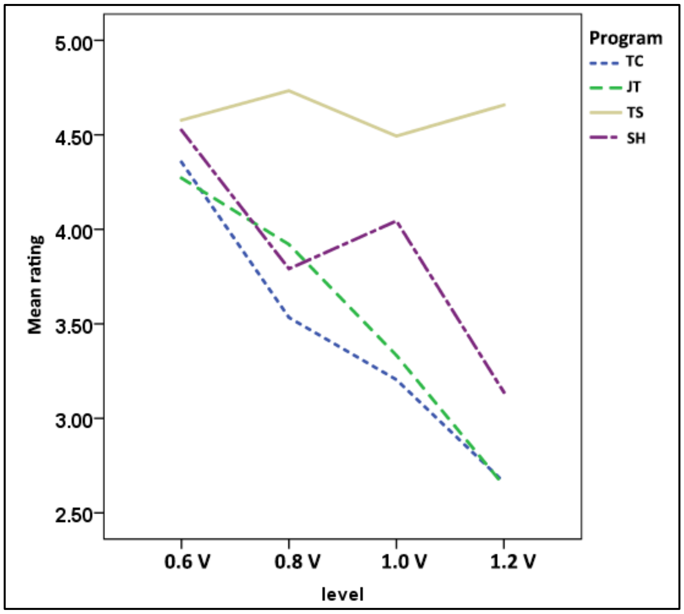

Figure 9: Mean ratings for different recording levels as well as different programs.

6 Conclusion

In this study, the question of how people perceive the deterioration of sound quality in music signals reproduced by microspeakers is studied through virtual listening tests. Based on the experimental evidence, the major conclusions are summarized as follows:

1) Listeners are able to perceive the growing impairment in sound quality of the music signals as the distortion level increased.

2) Program materials have huge effects on the audibility of distortion in microspeakers. Listeners are more difficult to perceive the distortion in programs with less low frequency content.

3) Dynamic ranges of programs also affect the perception of distortion. The influence of distortion level is limited on the programs with smaller dynamic ranges.

Reference

1. Klippel, W., “Large signal performance of tweeters, micro speakers and horn drivers,” 118th Audio Engineering Society Convention, 2005.
2. Klippel, W., “Modeling the large signal behavior of micro-speakers,” 133rd Audio engineering society convention, 2012.
3. Klippel, W., “Tutorial: Loudspeaker nonlinearities-Causes, parameters, symptoms,” Journal of the Audio Engineering Society, Vol. 54, pp. 907-939, 2006.
4. Klippel, W., “The mirror filter-A new basis for reducing nonlinear distortion and equalizing response in woofer systems,” Journal of the Audio Engineering Society, Vol. 40, pp. 675-691, 1992.
5. Klippel, W., “Direct feedback linearization of nonlinear loudspeaker systems,” Journal of the Audio Engineering Society, Vol. 46, pp. 499-507, 1998.
6. Hu, Y. X., Wang, M., Lu, J., Qiu, X. J., “Compensating the distortion of micro-speakers in a closed box with consideration of nonlinear mechanical resistance,” Journal of the Acoustical Society of America, Vol. 141, pp. 1144-1149, 2017.
7. Cabot, R. C., “Perception of nonlinear distortion,” Audio Engineering Society Conference: 2nd International Conference: The Art and Technology of Recording, 1984.
8. Fryer, P. A., “Intermodulation distortion listening tests,” 50th Audio Engineering Society Convention, 1975.
9. Boer, M. A., Nijmeijer. A. G. J., Schurer, H., Druyvesteyn, W. F., Slump, C. H. and Hermann, O. E., “Audibility of nonlinear distortion in loudspeakers,” 104th Audio Engineering Society Convention, 1998.
10. Voishvillo, A., “Assessment of nonlinearity in transducers and sound systems-From THD to Perceptual Models,” 121st Audio Engineering Society Convention, 2006.
11. Bryan, M. E. and Parbrook, H.D., “Just audible thresholds for harmonic distortion,” Acta Acustica united with Acustica, Vol. 10, No. 5, pp. 87-91, 1960.
12. Petri-larmi, M., Otala, M., and Lammasniemi, J., “Psychoacoustic detection threshold of transient intermodulation distortion,” Journal of the Audio Engineering Society, Vol. 28, pp. 98-105, 1980.
13. Brinicombe, D., “Just detectable distortion,” Wireless World, Vol. 87, pp. 56-57, 1981.
14. Klippel, W., “Speaker auralization-Subjective evaluation of nonlinear distortion,” 110th Audio Engineering Society Convention, 2001.
15. Temme, S., Olive, S. E., Tatarunis, S., Welti, T. and McMullin, E., “The correlation between distortion audibility and listener preference in headphones,” 137th Audio Engineering Society Convention, 2014.
16. Hiekkanen, T., Makivirta, A., Karjalainen, M., “Virtualized listening tests for loudspeakers,” Journal of the Audio Engineering Society, Vol. 57, pp. 442-442, 2009.
17. Sun, S. Y., Shen, Y., Liu, Z. Y., Feng. X. L., “The effects of recording and playback methods in virtual listening tests,” Journal of the Audio Engineering Society, Vol. 63, pp. 570-582, July 2015.
18. Olive, S. E., Welti, T., McMullin, E., “A virtual headphone listening test methodology,” Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones, 2013.
19. Olive, S. E., “A method for training listeners and selecting program material for listening tests,” 97th Audio Engineering Society Convention, 1994.
20. Olive, S. E., “A multiple regression model for predicting loudspeaker preference using objective measurements: Part I-Listening test results,” 116th Audio Engineering Society Convention, 2004.
21 Keppel, G., Wickens, T. D., Design and Analysis: A Researcher’s Handbook, 4th ed., Pearson Prentice Hall, 2004.
22 Bech, S., Zacharov, N., Perceptual Audio Evaluation-Theory, method and application, John Wiley & Sons, Ltd., 2006.
23 Olive, S. E., “Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study,” Journal of the Audio Engineering Society, Vol. 51, pp. 806-825, 2003.

The author can be reached at: yshen@nju.edu.cn.