Listening Tests in the Process of Microphone Development

Hendrik Paukert[1], Jonathan Ziegler[1,2]

[1]*Hochschule der Medien Stuttgart, Germany, Email: paukert@hdm-stuttgart.de*

[2]*Eberhard Karls Universität Tübingen, Germany Email: zieglerj@hdm-stuttgart.de*

Abstract

Preliminary listening tests play a key role in the development of novel types of digitally enhanced microphone arrays. The assessment of different types of noise and signal degradation can lead to a better understanding of which factors need the most attention in future development of signal processing algorithms. Along with the results, the principles of the listening test will be discussed, as well as the creation of suitable sound files.

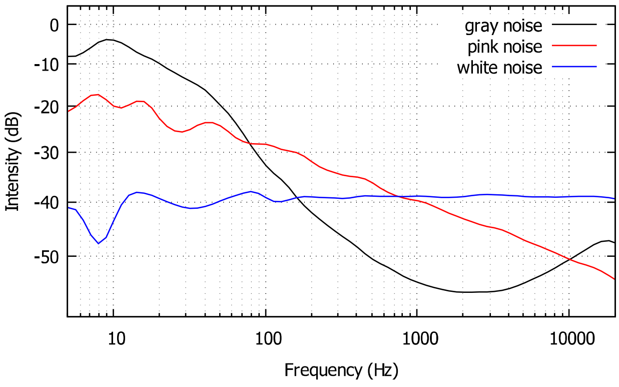
1. Preparation

The entire listening test was programmed using Max/MSP, an object oriented programming environment created by Cycling'74 [1]. This was chosen for its powerful and straightforward audio manipulation capabilities and the ability to quickly design a GUI for the test subjects. The test was comprised of pair comparisons [2], rankings [3] and active evaluations. To match the DSP algorithms for which the listening test was devised, the selected noise sources are jitter, compression artifacts and various colors of random noise. The creation of degraded speech signals took place using band pass filters, gates, limiters and audio clipping. All test subjects experienced the test on the same laptop with Beyerdynamic DT770 headphones [4] and data was acquired automatically. The set of test subjects consisted of audio professionals between 28 and 55 years of age. Special thanks go to Johanna Zehendner and Jo Jung for the generous contribution of speech samples [5-7].

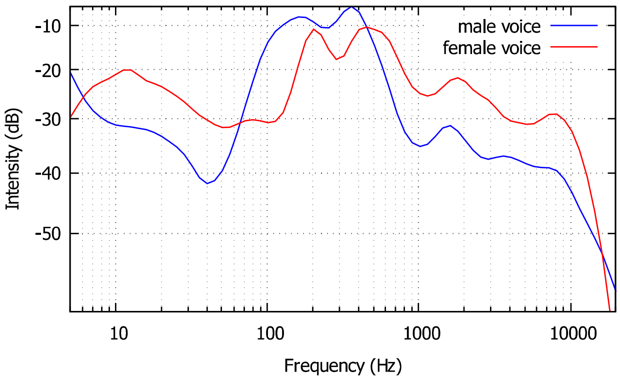
2. Synthesis of noise samples

White and pink noise were created using the noise generators integrated in Max/MSP. In addition, a type of noise was introduced to the test, which matches the spectral sensitivity of human hearing and thus should be more tolerable to an average listener. This was achieved by modeling white noise with appropriate equalization. As seen in Image 1, gray noise has a strong attenuation around 2000 Hz, which matches the heightened sensitivity of human hearing at this frequency [8].

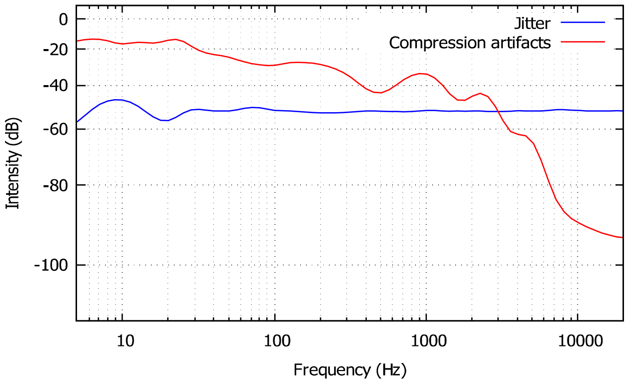
Jitter was captured by recording the signal of a damaged Toslink optical ADAT cable connecting a digital console to a recording interface. Compression artifacts were created using a specifically designed Max/MSP patch. Peak and RMS levels of the signals were analyzed using Audacity [9].



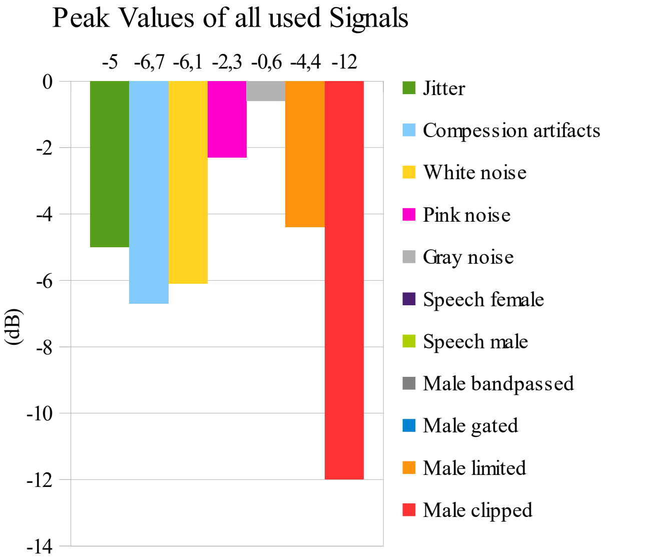
Img. 1: Overlay of white, pink and gray noise. White noise shows an even energy distribution over the entire audible spectrum, whereas pink and gray noise have specific spectral attributes.



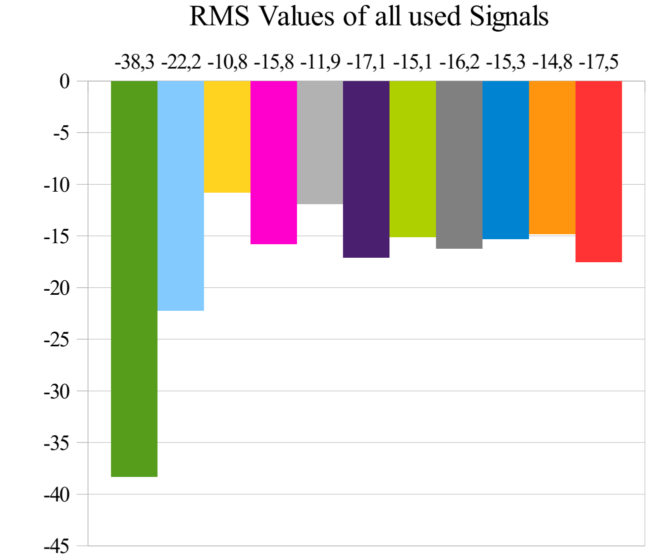
Img. 2: Spectral energy distribution of the speech samples used in tests 3-6.



Img. 3: Spectral analysis of compression artifacts and jitter. The compression artifacts contain little energy above 5000 Hz.



Img. 4: Peak values of all signals used in the listening tests. Before evaluation of the results individual gain matching was applied. ‘Speech female’, ‘Speech male’, ‘Male band passed’ and ‘Male gated’ have 0dB peak level.



Img. 5: RMS values of used signals. While jitter and white noise have similar peak levels, the difference in mean energy is significant.

3. Execution of listening tests

The program in use consists of 6 separate experiments and returns a total of 52 parameters for each test subject. Before the test is started, the subject is reminded that the listening test is devised for a speech-specific microphone and that therefore a focus should be placed on sound quality in regard to such signals.

Test 1: “Identical Disturbance Level”

Initially the test subject is asked to start a calibration signal to set the listening volume to a comfortable level. This process ensures that every listener is evaluating the signals within his or her own listening comfort zone.
Due to a variety of sonic differences in the noise samples, a direct comparison is not possible. A recording of white noise will be perceived to be louder than, for example, jitter at the same peak level. Therefore, every subject is asked to set the noise samples to a subjectively identical level. These gain values are consequently incorporated into the pair comparison tests.

Test 2: “Pair Comparison”

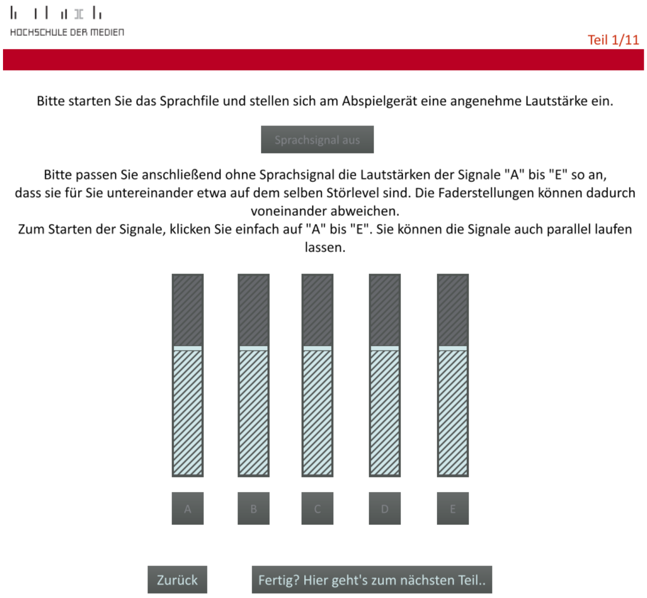
All noise signals are evaluated in pairs and the signal with a higher disturbance potential is chosen. Due to the fact that the test subject previously matched all signals to a subjectively identical level, the decision is based more on spectral and temporal energy distribution than on a general difference in volume between the signals.

Test 3: “Perception Threshold”

The signals are now examined based on an individual threshold of perception. In addition to the noise samples, a vocal sample is played back. Both male and female speech are used in order to examine a difference in signal masking. The user interface is similar to the one used in test 1.

Test 4: “Disturbance Threshold”

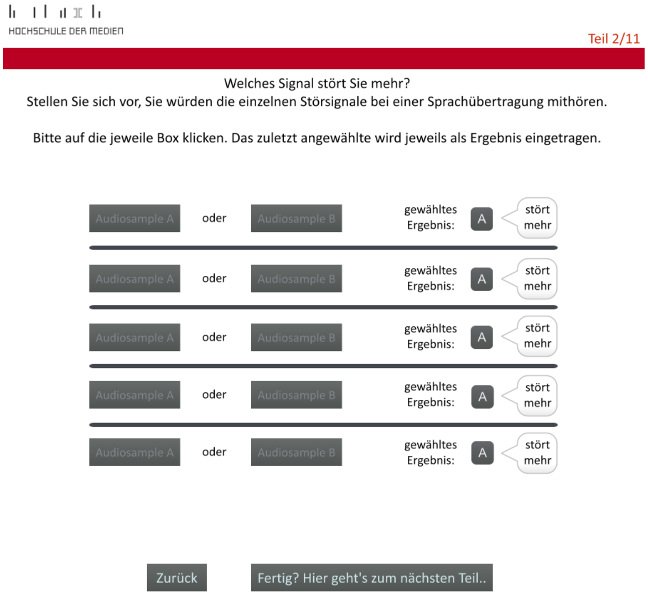
The disturbance threshold for each signal is determined using the same methodology as in the previous test.



Bitte starten Sie das Sprachfile und stellen sich am Abspielgerät eine angenehme Lautstärke ein.

Bitte passen Sie anschließend ohne Sprachsignal die Lautstärken der Signale "A" bis "E" so an, dass sie für Sie untereinander etwa auf dem selben Störlevel sind. Die Faderstellungen können dadurch voneinander abweichen.
Zum Starten der Signale, klicken Sie einfach auf "A" bis "E". Sie können die Signale auch parallel laufen lassen.

Img. 6: "Identical Disturbance Level." The test subject is asked to set the noise signals to identical levels using the provided faders and on/off buttons.



Welches Signal stört Sie mehr?
Stellen Sie sich vor, Sie würden die einzelnen Störsignale bei einer Sprachübertragung mithören.

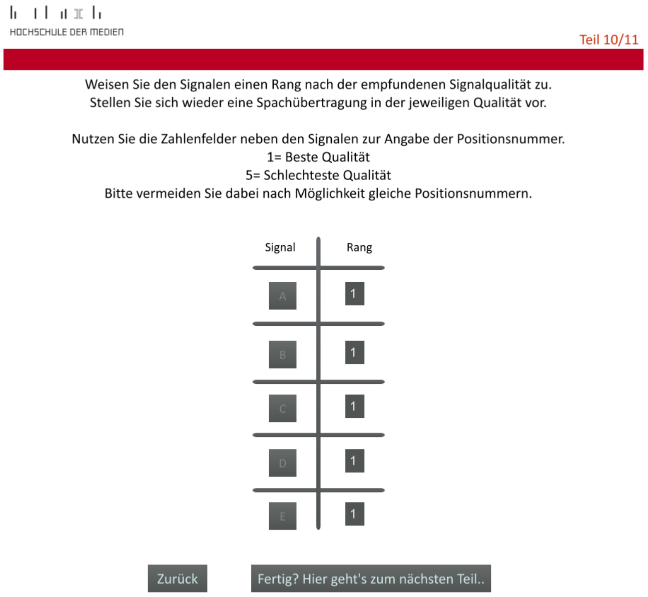
Bitte auf die jeweile Box klicken. Das zuletzt angewählte wird jeweils als Ergebnis eingetragen.

gewähltes Ergebnis: A
gewähltes Ergebnis: A
gewähltes Ergebnis: A
gewähltes Ergebnis: A
gewähltes Ergebnis: A

Img. 7: "Pair Comparison." The test subject is asked to determine the more bothersome signal within a pair.

Test 5: "Ranking"

The male speech sample is now played back in a variety of modifications, created with a gate, a band pass filter, a limiter and a clipper. All signals were modified to a similar degree. This insures that the character of a modification is perceived rather than its intensity. The subject is asked to rank the sound samples according to perceived signal quality. This test contains a blind reference.



Weisen Sie den Signalen einen Rang nach der empfundenen Signalqualität zu.
Stellen Sie sich wieder eine Spachübertragung in der jeweiligen Qualität vor.

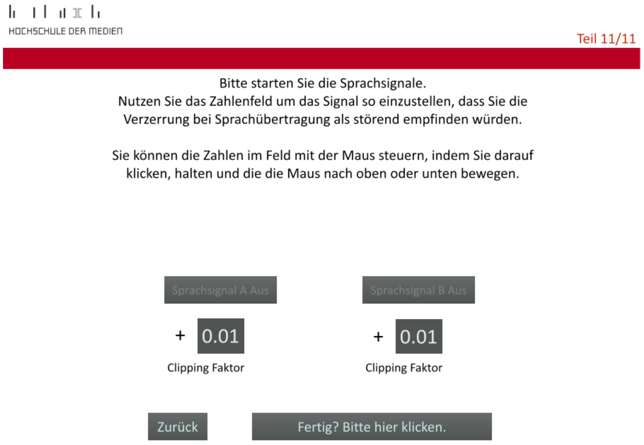
Nutzen Sie die Zahlenfelder neben den Signalen zur Angabe der Positionsnummer.
1= Beste Qualität
5= Schlechteste Qualität
Bitte vermeiden Sie dabei nach Möglichkeit gleiche Positionsnummern.

Signal	Rang
A	1
B	1
C	1
D	1
E	1

Img. 8: "Ranking." The subject compares the modified speech signals by clicking the tiles A-E and assigning a numerical rank to each sample.

Test 6: "Clipping"

In this test the amplitude of the audio samples of male and female speech are clipped. The subject is asked to set a tolerable level of clipping using a number box, as shown in Image 9. To prevent habits of audio professionals from influencing their decisions, no level meters are supplied. The factor is converted to decibels for the evaluation process and can be compared to the fixed clipping from test 5.



Bitte starten Sie die Sprachsignale.
Nutzen Sie das Zahlenfeld um das Signal so einzustellen, dass Sie die Verzerrung bei Sprachübertragung als störend empfinden würden.

Sie können die Zahlen im Feld mit der Maus steuern, indem Sie darauf klicken, halten und die die Maus nach oben oder unten bewegen.

+ 0.01 Clipping Faktor
+ 0.01 Clipping Faktor

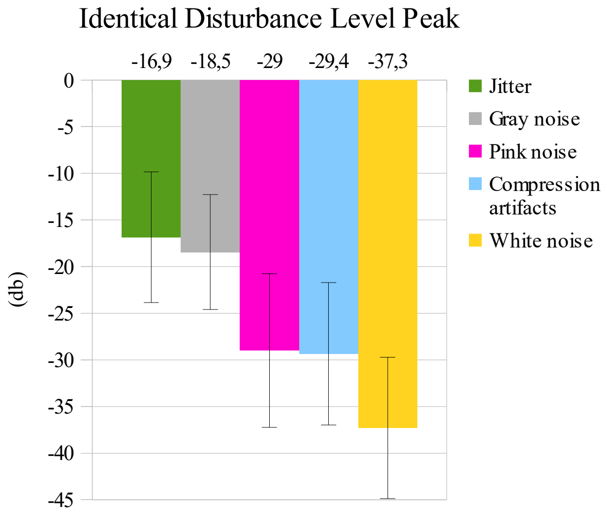
Img. 9: "Clipping." Test subjects set a tolerable level of clipping in male and female voice samples using a number box.

4. Results

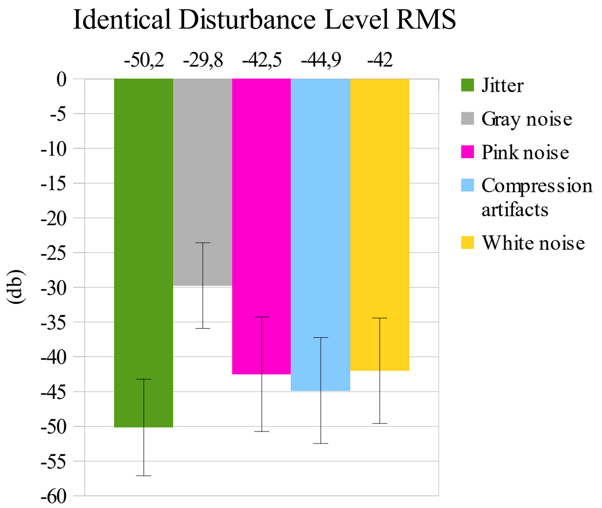
All results are gain-corrected using the peak level offsets shown in Image 4. By compensating for differences in peak and RMS levels, a uniform evaluation of all tests is achieved. The depicted results are gathered using averages of all test subjects.

Test 1: "Identical Disturbance Level"

Image 10 makes it quite clear that gray noise and jitter have a very high disturbance threshold. Thus, much higher peak levels can be tolerated than, for example, with white noise. The test subjects set pink noise and compression artifacts to similar peak levels. This could very likely be due to the spectral similarity of the two signals. RMS values for the chosen levels show more similarity. This can be observed in Image 11. The only exception is gray noise, where much higher RMS values are chosen due to reduced signal energy in the frequency bands most sensitive in human perception. The average standard deviation is 7.31 dB.



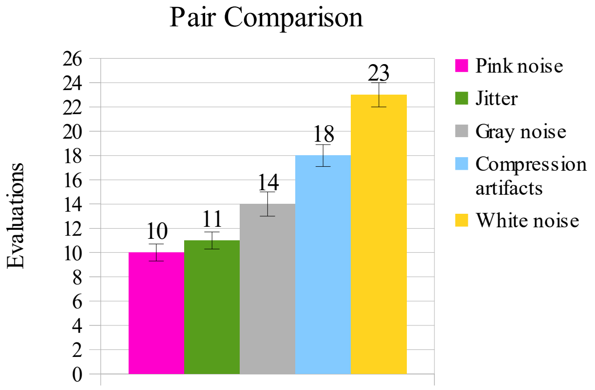
Img. 10: Results of “Identical Disturbance Level Peak.” Jitter is set to the highest, white noise is set to the lowest peak level with a difference of approximately 20 dB.



Img. 11: Results of “Identical Disturbance Level RMS.” Compared to peak levels in Image 5, RMS levels show much higher correlation.

Test 2: “Pair Comparison”

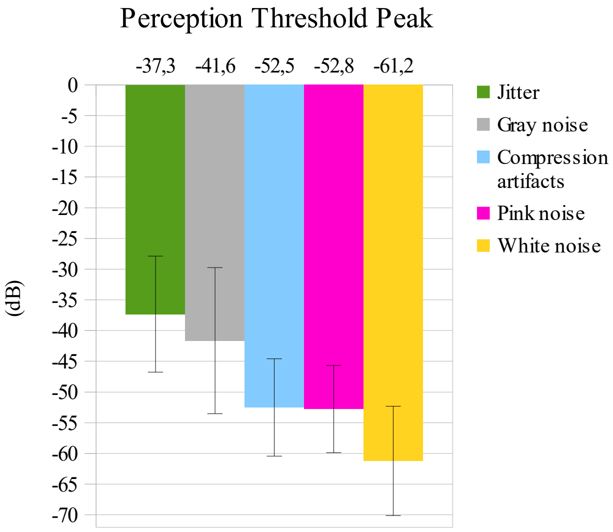
Pair comparison tests are run using the resulting disturbance levels from test 1, thus a consistent level of audible noise is achieved. Each noise sample is compared to every other noise sample, resulting in 10 choices per test subject. During evaluation the number of losing pair decisions is calculated per sample, determining the most disturbing source. As seen in Image 12, white noise clearly leads the list, followed by compression artifacts and gray noise. Pink noise and jitter were perceived to be the least bothersome.



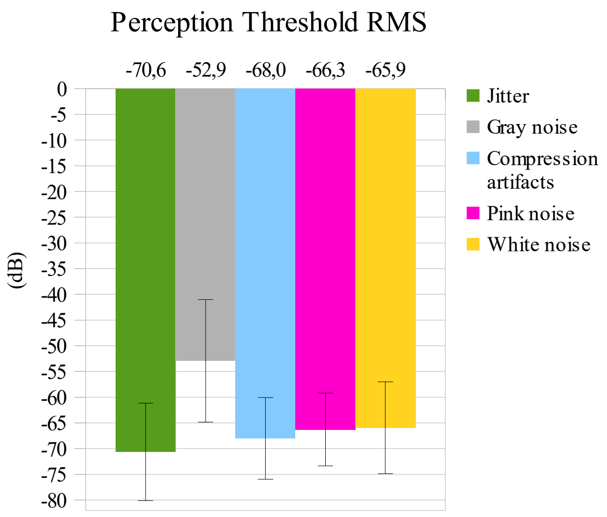
Img. 12: Results of “Pair Comparison.” The list of most disturbing noise sources is clearly led by white noise. Pink noise and jitter were perceived as the least disturbing.

Test 3: “Perception Threshold” - Part I

The perception thresholds of jitter and gray noise are highest, followed by compression artifacts and pink noise. White noise, on the other hand, can already be heard at very low signal levels. The average standard deviation is 9.1 dB. Pink noise shows the lowest, and gray noise the highest standard deviation of the tested signals. This could be due to hearing capability of the test subjects. Gray noise has the highest spectral energy in low and high frequency ranges. The high frequency sensitivity of human hearing is decreased with age and over-exposition to high sound pressure levels. This can cause a higher fluctuation in perceived noise levels.



Img. 13: Results of “Perception Threshold Peak.” Jitter can be added at the highest peak level without disturbance, while white noise can be detected at very low levels.



Img. 14: Results of "Perception Threshold RMS." A convergence of measured values can be detected. The highest standard deviation is found within gray noise.

Test 3: "Perception Threshold in Female Speech" - Part II

In contrast to the previous test, the order of the noise samples is changed: compression artifacts trade places with pink noise. The average standard deviation is 8.3 dB. As shown in Image 20, the perception threshold of the noise samples in female speech is on average slightly below the results in male speech. This can be attributed to a 2 dB higher RMS level of the male speech sample.

Comparative Analysis

When comparing perception thresholds of pure noise samples and noise in added speech, it becomes apparent that especially pink and gray noise can be increased in volume. Jitter and compression artifacts are masked least.

Perception threshold in female speech	Difference due to masking (dB Peak)	SD
Jitter	9.9	8.6
Compression artifacts	9.9	8.7
White noise	11.0	6.7
Pink noise	12.8	6.6
Gray noise	12.4	10.9
Average	*11.2*	*8.3*

Tab. 1: Results of "Perception Threshold in Female Speech." Pink noise profits most from masking effects and additionally shows the smallest standard deviation.

Test 3: "Perception Threshold in Male Speech" - Part III

The order of the samples is analogous to part 2 of this test and standard deviation is 7.8 dB.

Comparative Analysis

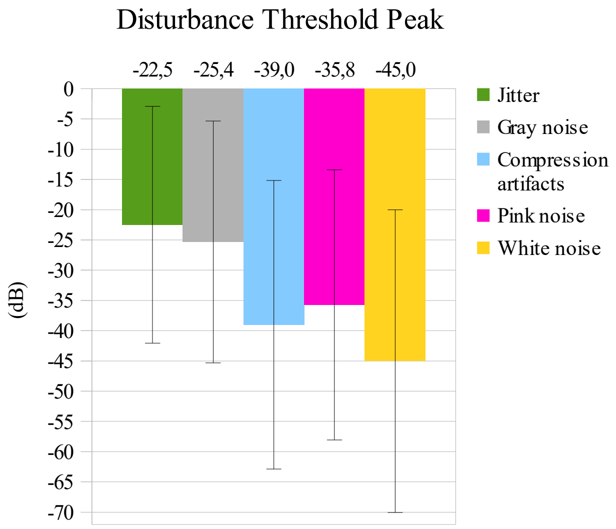
Masking effects are least apparent with jitter. As with female speech, pink and gray noise show the strongest masking characteristics. Additionally, pink noise shows an increase in masking of 1.5 dB compared to female speech.

Perception threshold in male speech	Difference due to masking (dB Peak)	SD
Jitter	9.2	9.0
Compression artifacts	9.8	8.5
White noise	11.8	7.0
Pink noise	14.3	5.1
Gray noise	12.7	9.4
Average	*11.6*	*7.8*

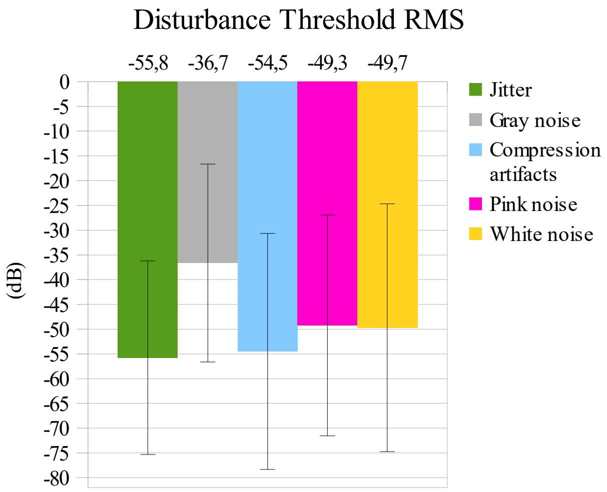
Tab. 2: Results of "Perception Threshold in Male Speech." As with female speech, pink noise shows both the most effective masking and lowest standard deviation.

Test 4: "Disturbance Threshold" - Part I

Compared to the tests concerning perception thresholds, pink noise shows a higher disturbance threshold and thus is slightly less bothersome. The highest levels are set for jitter, indicating the lowest relative disturbance among the compared signals. Compression artifacts and white noise show the lowest tolerance level. On average, the disturbance threshold is 15.6 dB above the perception threshold. The standard deviation of 22.2 dB on average is 13.1 dB higher than for the perception threshold. This could be due to a missing reference, unclear definitions of disturbance or diverging sensitivity for noise among test subjects.



Img. 15: Results of "Disturbance Threshold Peak." A significant difference is the greatly increased standard deviation.



Img. 16: Results of „Disturbance Threshold RMS." As with peak results, a significant fluctuation within the group of test subjects is registered.

Test 4: "Disturbance Threshold in Female Speech" - Part II

The average standard deviation is reduced through the introduction of a speech signal (female voice) from 22.2 dB to 15.6 dB. This level is still 6.5 dB above those of the tests concerning perception thresholds. The order of signals remains unchanged.

Comparative Analysis

Pink noise is masked most in the hearing tests with speech signals while the disturbance threshold changes most for jitter. Also, no direct correlation between the disturbance threshold and the perception threshold can be detected.

Disturbance threshold in female speech	Difference due to masking (dB Peak)	SD
Jitter	6.5	14.1
Compression artifacts	5.2	16.2
White noise	3.4	16.7
Pink noise	4.3	16.0
Gray noise	4.8	15.1
Average	*4.8*	*15.6*

Tab. 3: Results of "Disturbance Threshold in Female Speech." Concerning disturbance thresholds, jitter profits most from masking effects. The most notable difference to the tests concerning perception thresholds are the much higher standard deviations.

Test 4: "Disturbance Threshold in Male Speech" - Part III

The levels set by the test subjects are up to 2.2 dB higher than with female speech, thus confirming tendencies of the perception threshold tests. This is due to the 2 dB higher RMS value of the male speech sample. The order stays unchanged and standard deviation is at 15.5 dB.

Comparative Analysis

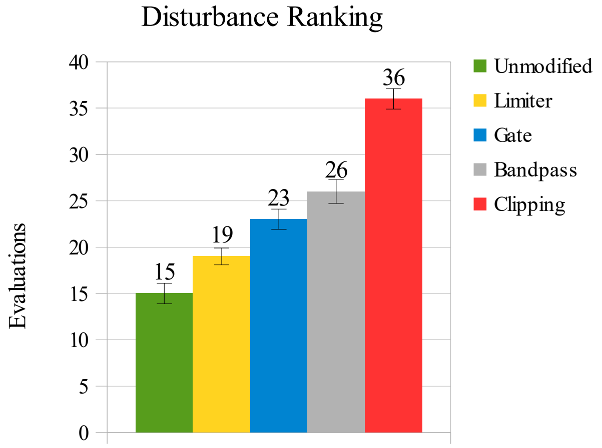
On average, the difference in masking is increased by 1.4 dB. In comparison to the perception threshold, masking occurs less, especially for gray noise. Jitter and compression artifacts profit most from masking effects.

Disturbance threshold in male speech	Difference due to masking (dB Peak)	SD
Jitter	7.0	14.1
Compression artifacts	7.3	18.3
White noise	5.5	16.4
Pink noise	6.5	13.8
Gray noise	5.0	14.9
Average	*6.2*	*15.5*

Tab. 4: Results of "Disturbance Threshold in Male Speech." Compression artifacts and jitter profit most from masking effects with male speech.

Test 5: "Ranking"

The unmodified sound sample is ranked highest with the signal treated with a limiter coming in second. The worst marks are given to the signals treated with clipping and band pass filters.



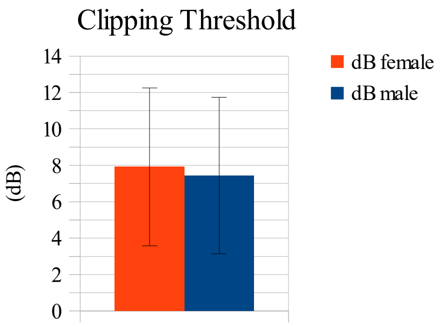
Img. 17: Results of "Disturbance Ranking." The original sample is ranked highest and the signal with clipping lowest.

Test 6: "Clipping Threshold"

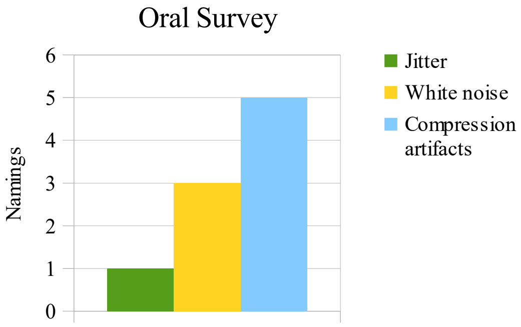
Clipping thresholds set by the test subjects are nearly identical between male and female speech with a difference of 0.5 dB. The higher RMS of the male speech sample could result in more noticeable clipping effects and would explain the lower threshold. Standard deviation is 4.3 dB for male speech and female speech. This indicates a very individual perception of disturbance.

Concluding Oral Survey

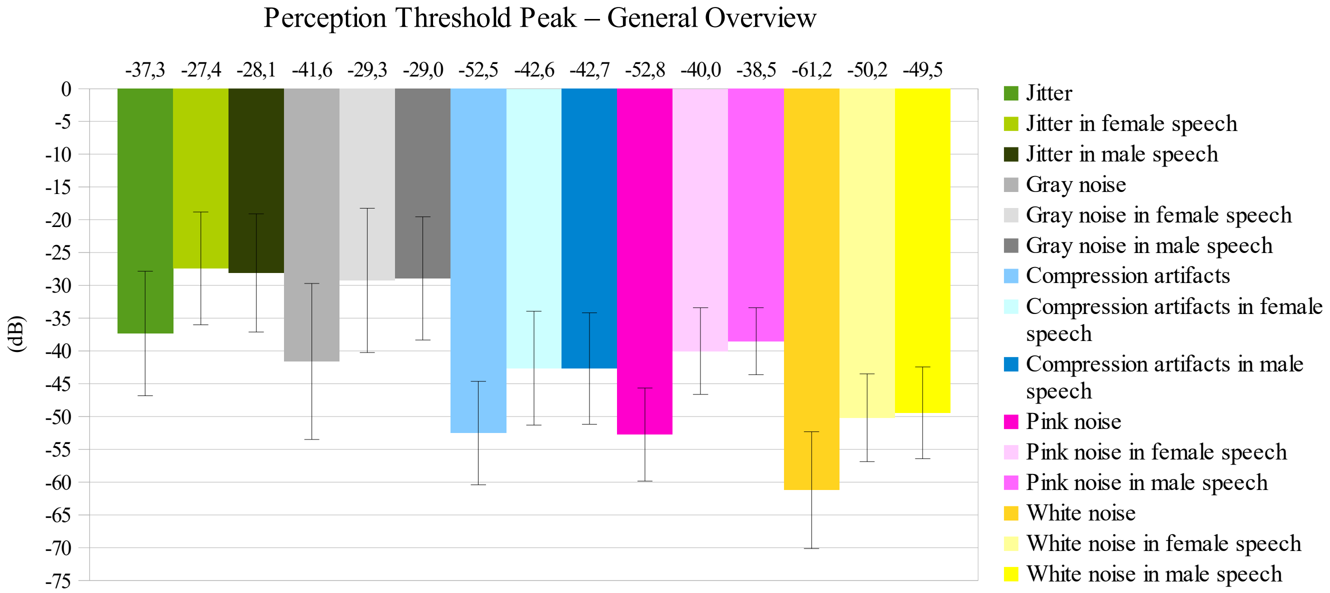
After completion of the test, the subjects were asked to name the least pleasant signal. The results of the survey can be seen in Image 19.



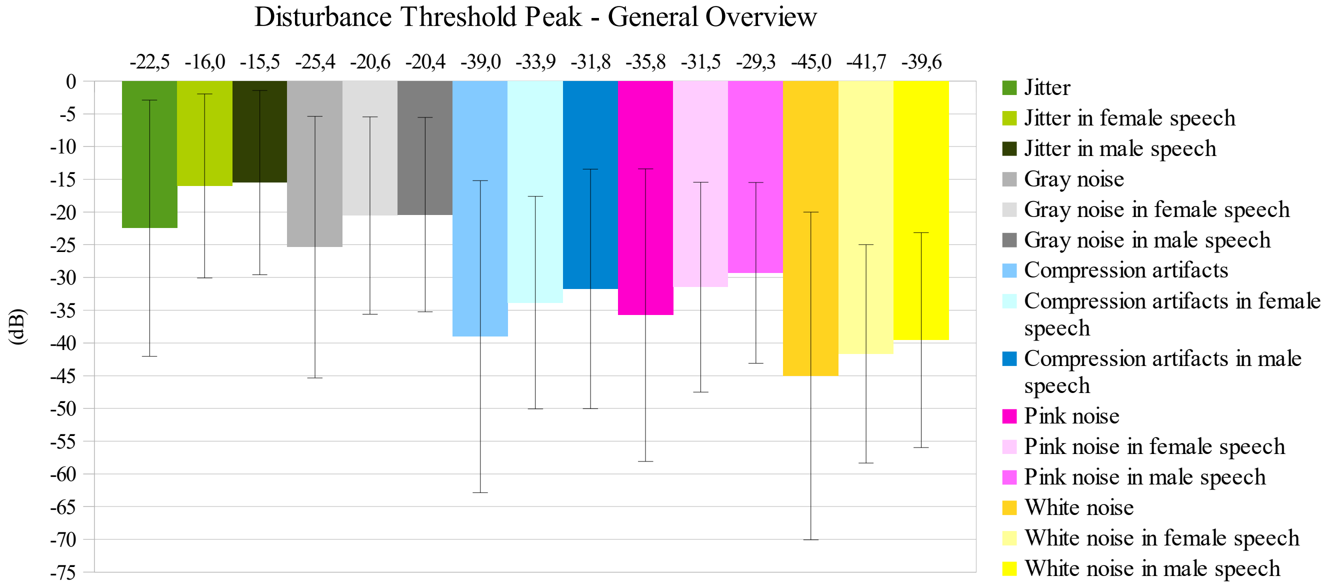
Img. 18: Results of “Clipping Threshold.” Differences between male and female speech are below statistical relevance.



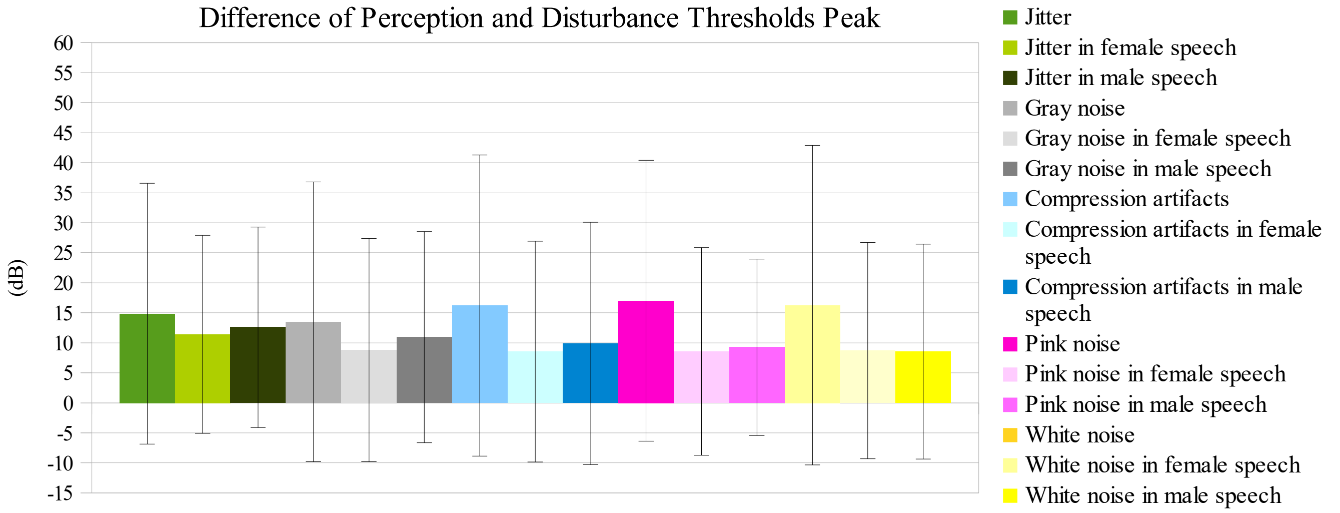
Img. 19: Results of “Oral Survey.” Compression artifacts were perceived as the most disturbing. Of the five options, gray and pink noise were not mentioned at all.



Img. 20: Results of “Perception Threshold Peak.” In general, white noise has the lowest and jitter the highest perception threshold. Pink noise profits most from masking effects by speech signals and has the lowest standard deviation.



Img. 21: Results of “Disturbance Threshold Peak.” Jitter provides the least potential for disturbance, while white noise has the lowest disturbance threshold. The high standard deviation shows a wide variety of sensitivity towards noise in the test subjects.



Img. 22: Results of "Difference of perception and disturbance thresholds Peak." Gray noise shows the smallest level difference between perception and disturbance thresholds, thus being felt as disturbing relatively quickly after its perception. Pink noise has a threshold interval of 17dB and therefore is tolerated at levels well above the perception threshold. The pure noise samples uniformly show larger differences between perception and disturbance. The combined standard deviations of perception and disturbance thresholds create a larger spread.

5. Conclusion

In regard to peak levels, jitter provides the highest tolerability of all tested noise samples and can be present at the highest signal to noise ratio without being considered disturbing. White noise is detected and perceived as disturbing at much lower levels. Subjective determination of disturbance thresholds results in a significant spread of the results. The same occurs in perception thresholds of signals with a high percentage of spectral energy at the upper and lower end of human hearing.

Heavily clipped audio samples were considered to be of the poorest signal quality, while the unmodified signals were ranked highest. Additionally, modifications which increase intelligibility, such as noise gates and limiters did not significantly reduce the perceived signal quality.

6. Outlook

The described listening tests compose a foundation for further, more complex examinations within a larger project. The results will be used for prioritization within the development of DSP algorithms and for the creation of more detailed testing environments. If needed, various combinations of modified signals and noise sources can be surveyed within similar listening tests. Additional focus can be placed on the analysis of age distribution among test subjects.

More detailed evaluation of near-production prototypes will take place following the suggestions of ITU-R BS.1116[10] and ITU-R BS.1534[11]. In addition, listening tests with hidden reference and anchor (ABC/HR or MUSHRA) will be used.

7. References

[1] CYCLING '74, Max 7 perpetual licence, URL: https://www.cycling74.com

[2] HEAD Acoustics Application Note, Page 2, URL:https://www.headacoustics.de/de/nvh_application_notes_jury_evaluation.htm

[3] HEAD Acoustics Application Note, Page 1, URL: https://www.headacoustics.de/de/nvh_application_notes_jury_evaluation.htm

[4] Beyerdynamik GmbH & Co. KG, URL: http://www.beyerdynamic.de/shop/dt-770-pro.html

[5] Johanna Zehender, Radioplay „Menschlich ist", URL: https://www.johannazehendner.allyou.net

[6] Jo Jung, Radioplay „Menschlich ist", URL: http://www.jo-jung.eu

[7] A. Hummel, V. Kuptsov, M. Köhler, S. Kreuzer, H. Paukert: Radioplay „Menschlich ist", recorded at University of applied Science Stuttgart, URL: https://www.hdm-stuttgart.de/ URL: https://www.facebook.com/menschlichist/

[8] Michael Dickreiter, Volker Dittel, Wolfgang Hoeg, Martin Wöhr: Handbuch der Tonstudiotechnik, Band 1, 7. Auflage (2008), Saur Verlag, ISBN 979-3-598-11765-7, Seite 100, Abb.3/4

[9] Audacity, open source audio software, URL: http://www.audacityteam.org/

[10] International Telecommunication Union, URL: http://www.itu.int/rec/R-REC-BS.1116

[11] International Telecommunication Union, URL: http://www.itu.int/rec/R-REC-BS.1534