Development and Evaluation of a New CI Pitch Perception Test - the Glide Tone Test

I. M. Venema

Student number: 5185513
Study: Biomedical Engineering
Track: Medical Devices
Supervisor: dr. ir. R. C. Hendriks (Technical University Delft)
dr. ir. J. J. Briaire (Leiden University Medical Center)

30th October 2022

Abstract

Current speech-encoding strategies in cochlear implants (CI's) do not provide their users with sufficient spectral resolution for complex listening tasks and pitch perception. This paper presents a new test that has been developed to assess pitch perception: the glide tone test (GTT). A total of 25 normal hearing (NH) participants were tested to find their just noticeable difference (JND) between glide tones with a fixed range but a different exponent. The stimuli used mimic the sound spectrum of a violin and have 20 harmonics added to the fundamental frequency (F0). This paper also investigated possible learning effects the test might possess, evaluated user-friendliness and optimised the procedure. The paper presents the JND scores obtained by the NH subjects, as well as the new testing method itself. Suggestions for further experiments with the GTT are made.

Contents

1 Introduction

1.1 General introduction

Since 1980, cochlear implants (CI) became in use to help the profound to severely deaf restore their hearing. Now it has been considered one of the most successful sensory prostheses (Mudry & Mills, 2013). Recent estimates show that approximately 800,000 people world wide have received such an implant (Boisvert, Reis, Au, Cowan, & Dowell, 2020). The primary goal of the implant is to restore speech understanding and enable danger detection. With current speech-encoding strategies, the average CI user can recognize about 95% of the words of a sentence using only the implant (for non-tonal languages) (Shannon, Fu, & Galvin 3rd, 2004). However, complex listening tasks such as the perception of speech-in-noise, perception of music and tonal language understanding are still problematic for most CI users. All these require a higher degree of spectral resolution than is currently provided by the implant (Smith, Delgutte, & Oxenham, 2002).

A higher spectral resolution is necessary for listeners to distinguish between frequency patterns of sound (also known as the spectral envelope, or, in simpler terms 'pitch') (Winn & O'Brien, 2022). The broader effect of limited pitch perception has been extensively researched and spans skills most of normal hearing (NH) people take for granted. CI users have difficulty recognizing vocal emotion (Jiam, Caldwell, Deroche, Chatterjee, & Limb, 2017), listening to and enjoying music (Looi, Rutledge, & Prvan, 2019), and detecting the shifts of pitch that are part of tonal languages such as Mandarin (Wei, Cao, & Zeng, 2004). Lack of these skills has a negative effect on social integration and the quality of life of people suffering from hearing loss (MacDonald, 2013). Research on improving spectral resolution, and thus perception with CIs, can be divided into two groups. One looks into completely new stimulation of the auditory nerve, such as optogenetics - using light, rather than electricity to stimulate the nerve (Thompson et al., 2020; Dieter, Duque-Afonso, Rankovic, Jeschke, & Moser, 2019). The second group tries to improve the sound encoding strategy used that transforms sound to electrical pulse train (a comprehensive overview of sound encoding strategies developed in the last years can be found in (Wouters, McDermott, & Francart, 2015)).

With the development of more complex speech encoding strategies came the development of CI performance tests. At the moment there are no objective neurophysiological markers or validated models that can predict the performance of different kinds of electrical stimulation (Wouters et al., 2015). Therefore, the assessment of speech encoding strategies is in principal based on behavioral performance measurements, or, in other words, perception tasks. Perception tasks in CI research are often called psychophysical tests, because they have a physical aspect - properties of sound for instance - and a psychological one - they measure how these physical properties are perceived. This combination between presentation of objective information and its perception, makes the tests more complex than would be expected. It also leads to confusing terminology (see Section 1.7) as physical and perceptive definitions are sometimes mixed. Think for instance of frequency and pitch, or intensity and loudness (in both cases the second term describes the perceptual interpretation of the first). To confuse matters further, one aspect of sound can have an impact on the perception of another aspect of sound, while physically they are separate. This means that a test can be good on paper, and yet not work as soon as human subjects are involved. A classical example of this is the relationship between loudness and pitch, which will be explained in the next section.

1.2 Pitch and its perception

Pitch is the perceptual interpretation of frequency. A simple tone consists of a periodic wave of which the rate of periodicity is expressed in frequency (Hz), travelling through air. This wave moves the ear drum (or tympanic membrane) which pushes and pulls at the ossicles that are attached on one side to the tympanic membrane and on the other to the oval window of the cochlea. Movement of the ossicles on the oval window creates waves in the cochlear fluid within the cochlea. This in turn moves the basilar membrane on which the hair cells lie. The ossicles function as a lever because the sound wave that previously travelled through air, now travels through cochlear liquid in the inner ear. The difference of impedance between these media requires some kind of impedance transformer to reduce reflection of sound energy. The hair cells send signals through the auditory nerve to the brain, where high frequency is perceived as high pitch and vice versa (Pickles, 1998).

Natural sounds are usually more complex than a pure periodic wave. They consist of multiple waves, stacked on top of each other. The lowest frequency of a series is called the fundamental frequency (F0) and

if the other waves are integral multiples of F0, they are called harmonics (Gelfand, 2017).

There are three main cues that convey pitch information to the auditory system. The first is spatial, according to which hair cells are excited on the auditory nerve (low frequencies at the apical end, high frequencies at the basal end (Drennan & Rubinstein, 2008)). This is the main information used in traditional CI stimulation methods (for instance CIS). Here, a Fourier analysis divides the signal into frequency bands that correspond to a specific area of the cochlea to stimulate. The 12-22 electrodes on the array (the exact amount depends on implant type and manufacturer) send pulse trains to the auditory nerve. In comparison, NH people have approximately 3500 inner hair cells that provide this information. This limited spatial resolution of electrodes on the array is not easily solved, as current spread leads to interfering signals when more electrodes are added (Boëx, De Balthasar, Kós, & Pelizzone, 2003). After sound has been sorted in frequency bands, each band undergoes a Hilbert transform that decomposes the time signal into the temporal envelope (ENV) and the temporal fine structure (TFS). The ENV follows lower, global frequencies, and the TFS higher frequencies of 500 Hz to 10 kHz, close to the center frequency of corresponding band (Moore, 2008; Wouters et al., 2015; Moon & Hong, 2014).

For NH people, a second way of pitch perception is when neurons synchronize their firing rate to match the frequency of a tone, which is possible as long as it remains below approximately 1500 Hz. This concept is known as phase-locking and can yield a strong perception of pitch (Wouters et al., 2015). A common accepted theory is that TFS is represented by phase-locking of neurons as long as the frequency of the signal remains below the maximum firing rate of the cells (Moon & Hong, 2014). If sound exceeds this frequency, only place-coding provides the perception of pitch (Rubinstein, 2004). For many years, the ENV was considered to be more important for the signal than the TFS. Therefore, the shape of the ENV was used in sound encoding to generate a shape of a pulse that matches the shape of the sound and TFS information was discarded. However, more and more studies demonstrate that TFS is needed for pitch perception (a.o. (Smith et al., 2002; Moore, 2008; Moon & Hong, 2014)).

The third and last cue the auditory system uses to perceive pitch is the periodicity of all harmonics when they are combined. Perception of this periodicity only works up to around 300-500 Hz (Carlyon & Deeks, 2002) and is quite weak when compared to TFS. Nevertheless, for complete pitch perception all three mechanisms are required as they each work best for different situations.

1.3 Music perception

Since the primary goal of CI development has been the restoration of speech understanding, this consequently reduced the focus on other aspects of listening such as music. Music requires the perception of three main aspects of sound: rhythm, timbre and pitch (Jiam et al., 2017). In this research, the focus will be on the latter. Pure tones excite a narrow region of the basilar membrane in the cochlea, but more complex, musical sounds consist of many harmonics and thus excite a larger region of the membrane (Drennan & Rubinstein, 2008). When a complex tone excites a larger region, the consequence is that differences between several complex tones become quite small when only place-pitch is considered (the specific excitation of hair cells based on their place on the basilar membrane), as each sound needs a large region of stimulation. Therefore, the temporal information is needed for complex-tone pitch perception. TFS not only present in place-pitch perception but also somewhat preserved by phase-locking of neurons (Verschooten et al., 2019). This has also been demonstrated by studies that presented subjects with chimaeric sounds; without proper TFS information, complex pitch perception becomes problematic (Smith et al., 2002).

To understand the confusing link between loudness and pitch, it is also necessary to know some things about loudness perception. For NH people, a louder sound will excite a broader region of the basilar membrane (Pickles, 1998). In the implant, this is often translated to the amount of charge delivered to the nerve, but spread of excitation also plays a role (Briaire & Frijns, 2006; He et al., 2020). This directly interferes with spectral resolution and thus pitch perception, but the effect goes two ways. Sounds that stimulate a larger area of the auditory nerve can be perceived louder than those who stimulate one very specific area. On top of that, there seems to be a dependency between intensity and perceived pitch, with increased intensity leading to the perception of higher tones in pitch (Arnoldner, Kaider, & Hamzavi, 2006).

1.4 Pitch testing

There are many tests in literature that measure a subjects ability to detect pitch.[1] Perception tasks commonly used include: speech understanding tasks, pitch discrimination and detection tasks, pitch direction tasks and many more. A downside of most of these tests is that they often use pure tone stimuli (some examples include: (Zhang et al., 2019; Wagner, Altindal, Plontke, & Rahne, 2021)) or simplified tones that contain less harmonics than a natural sound would (o.a. (Fitzgerald et al., 2007; Wang, Zhou, & Xu, 2011; Vaerenberg et al., 2011)). As explained in the section above, pure tone frequency detection is not representative of music perception. Although the second list of studies adds some harmonics to their pitch (mostly 3 on top of F0, e.g. in (Vaerenberg et al., 2011) the frequency range is 200-800Hz), this addition still does not fully encompass the complexity of natural sound.

One study was found that used complex, naturalistic stimuli that consisted of an F0 with 15 consecutive harmonics on top (Luo, Masterson, & Wu, 2014). They assessed pitch ranking - identifying which sound out of two has the higher pitch - and melodic interval ranking for both NH and CI subjects. This specific study found no significant difference in pitch ranking thresholds between NH and CI subjects, but based on high variability between subjects and previous studies conducted (Gfeller et al., 2002; Kang et al., 2009), they concluded this was likely due to these specific participants who performed exceptionally well.

The most natural music perception task would be to ask subjects to identify a piece of music. A downside of such a melody recognition task is that a piece of music contains many cues, of which pitch is only one. Often, rhythm can be used to recognize a piece. Research has demonstrated that rhythm perception of CI users is often as good as rhythm perception of people with normal hearing (NH), and thus rhythm will need to be removed if the goal of the task is melodic pitch perception (Nimmons et al., 2008). Removing rhythm from existing music is very unnatural. Nevertheless some experiments have been conducted with these so-called isochronous melody tests (Jung et al., 2010), but these tests are often too difficult for CI users. Another downside is that recognition of existing melodies (for instance nursery rhymes) requires knowledge of the chosen fragments, which makes if unclear what the test measures: pitch perception or the subjects familiarity with the chosen songs.

A test developed by Galvin, Fu & Nogaki called the melodic contour identification test (MCI) was developed to overcome these problems. Random melodies are generated that subjects need to recognize. To make the test more complex, they did not use pure tones but again a simplified natural combination of tones, consisting of three harmonics (Galvin III, Fu, & Nogaki, 2007). This test finds a way to combine clinical pitch perception with music perception, but it could be argued that the biggest difference between test and real life (the amount of harmonics in a natural sound), is still present. Temporal pitch perception is needed when place-pitch differences are small. However, when examining this paper more closely, there seems to still be a big difference in place-pitch between the different sounds. If the aim of a pitch test is to see if enough temporal information was provided, the stimuli need to be complex enough so that TFS is needed.

1.5 TFS testing

When specific TFS speech encoding strategies were developed, also specific TFS detection tests were formed. These tests include the TFS test of Moore and Sek (Moore & Sek, 2009), which has frequently been used and adapted (Füllgrabe & Moore, 2017). This test measures the highest frequency at which a phase shift can be detected. There are also tests that measure a subjects' ability to detect differences in spectral ripple density using TFS information. These tests do not focus on pitch perception.

Vaerenberg et al. developed the harmonic intonation and the disharmonic intonation tests, which were already mentioned in Section 1.4 (Vaerenberg et al., 2011). This tests asks subjects to discriminate between harmonic and disharmonic gliding tones[2], focusing on the frequency range associated with TFS. This test does not provide a full range of harmonics, but only the first three, to stay in the lower range. Indeed, it was found that most CI users have abnormal low pitch perception, which seems to confirm poor TFS processing (D'Alessandro et al., 2018). A significant correlation was found between temporal and music sensitivity, which further supports the idea that more low frequency and TFS cues could improve music perception (D'Alessandro, Ballantyne, Portanova, Greco, & Mancini, 2022).

[1] Contact the author for a full literature review of pitch perception tasks for CI users.

[2] A gliding tone is a tone that gradually changes its frequency over time.

1.6 Research problem

This paper presents the development of a new test that assesses the subjects' ability to detect differences between different gliding tones: the glide tone test (GTT). The detection of gliding tones requires perception of pitch over time. An advantage of the use of gliding tones is that they contain complex pitch information that is needed for music perception, without giving any unwanted cues that are sometimes present in other music related task such as the presence of a recognizable melody a rhythm or loudness cues. Complex stimuli that activate the full range of the electrode array of CI patients will be used. This ensures limited place-pitch differences between the stimuli, which means that temporal information (TFS) will be necessary to be able to perform the task. Hopefully, this will provide insight into the effectiveness of speech-encoding strategies to deliver complex sound spectra to their users.

This study will only examine NH subject to provide a benchmark of measurements that can later be used to compare CI patients to. This step ensures that the task developed will not be too difficult for CI patients, as they generally perform poorer than their NH counterparts on similar tests. Therefor the test developed should be not too difficult for the NH subjects.

The developed test should not take too long to complete, preferably under an hour, so that participants remain focused and it would be easy to use the test in the clinic.

In this paper the construction of stimuli (Section 2.1.1), development of the procedure (Section 2.1.2) and preliminary results using NH subjects (Section 3) are presented. The test and procedure are evaluated and future recommendations are made that will translate the test directly to the CI user.

1.7 Key terminology and abbreviations

Term	**Definition**
AFC	Alternative forced choice.
CI	Cochlear implant.
CIS	Continuous interleafed sampling - a common speech encoding strategy used in CIs.
complex tone	A tone that consist out of more than one frequency, generally multiple overtones are stacked on top of F0.
ENV	Envelope, slower variation in amplitude of speech signal over time.
frequency	Amount of oscillations per second, indicated in Hz, but perceived as pitch.
glide tone	A sound that changes in frequency over time.
glissando	Musical term for a note that changes its frequency over time
JND	Just noticeable difference. A term that indicates the last stimulus that can be distinguished by the participant.
loudest-correct	A percentage that indicates the relative amount of correct answers that were played at the loudest intensity when compared to the other two stimuli.
loudness	Perceived intensity level of a sound, often indicated in dB SPL.
naturalistic	The act of mimicking something found in nature.
NH	Normal hearing
pitch	The relative perception of frequencies.
place-pitch	Pitch detected by the region of the auditory nerve or basilar membrane stimulated.
prosody	Emotional pitch information in a voice, also known as speech intonation.
TFS	Temporal fine structure. Fast oscillations close to frequency band center.

Table 1: Terminology used in this paper, sorted alphabetically.

2 Methods

The glide tone test is a three alternative forced choice test (3-AFC) where subjects are presented with three stimuli, two of which are the same and one is different. It is up to the participant to identify the odd-one-out. The difference between the reference stimuli and the target stimuli will become smaller when correct answers are given in order to find the subjects just noticeable difference (JND).

2.1 Theoretical foundations

2.1.1 Stimuli

The stimuli used in this experiment are complex gliding tones with a fixed starting and ending frequency, but a change of exponent. A gliding tone is a sound that gradually changes its frequency over time. One of the reasons a glide tone was used, is the distinct correlation between pitch and perceived loudness (Arnoldner et al., 2006). In experiments where subjects have to identify pitch differences over changing frequency ranges (for instance the difference between 442Hz-884Hz and 442Hz-662Hz) loudness can be used as cue to differentiate between sounds, rather than pitch. Using gliding tones that have a constant starting and ending frequency should eliminate loudness cues that are related to pitch. In this experiment, all stimuli used will have a fixed *range* but a changing *exponent*.

If the periodic nature of sound is ignored for a moment, it is possible to view a static sound as a straight horizontal line and a gliding tone as a linear line with an incline (see Fig. 1). A basic formula of a linear function is $f = t^n$, where n indicates the exponent of the line. If n=1, the line is said to be linear. Figure 2 shows lines with different exponents to demonstrate different inclines used in the gliding tones. It can be seen that the beginning and end point of the lines converge, just as they will for the stimuli used in the research.

When choosing a gliding tone as stimulus for the experiment, it is essential that all subjects are able to detect a difference between the beginning and end frequency of the glide. A study using complex tones found that CI users can detect between 1 to 8 semitones difference between sounds (Kang et al., 2009). (In music, each octave is divided up into 12 semitones. Each key of a piano is one semitone separate from its' adjacent keys.) Gliding tones in this research all had an F0 that started at 442Hz and ended at 884Hz. This is a range of one octave, or 12 semitones, between two A's. This chosen range is based on music (A is the note on which orchestra's tune) and the interval is sufficiently large that it should be perceivable to the average CI user.

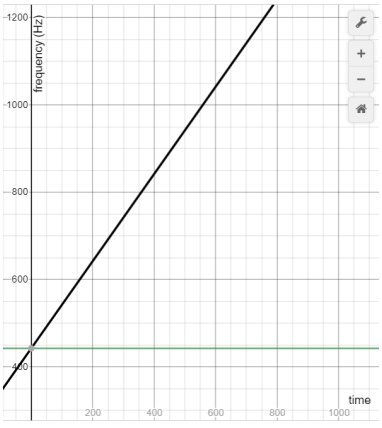

Figure 1: Example of a static and gliding sound. The x-axis shows time and the y-axis frequency.

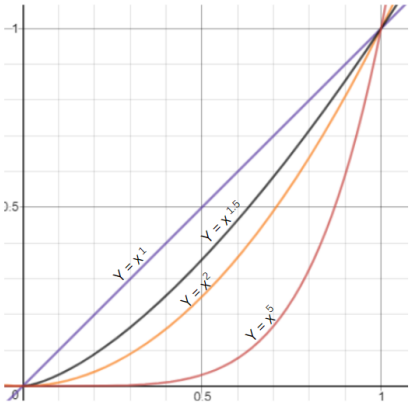

Figure 2: Example of gliding tones with different exponents.

The difference between the gliding tones generated was the exponent of the glide. As explained above, the exponent of a simple line can be represented with the letter n. However, sound consists of a string of (in this case) cosine waves that have to smoothly change frequency over time. This was achieved using the formula below, where F_{end} = the ending frequency of the glide (in Hz), F_{start} = the starting frequency of

the glide (in Hz), t = the duration of glide (in s), P = the slope of the glide (numerical value) and ϕ = the starting phase (in degrees). This formula was taken from the Matlab function chirp.m which generates a continuous swept frequency, or glide.

$$\beta = (F_{end} - F_{start}) * t^{-P}$$

$$y = cos((2\pi \frac{\beta}{(1+P)} \times t^{(1+P)}) + F_{start} \times t + \frac{\phi}{360})$$

The most important take-away from these formulas is that now P indicates the exponent of the glide, instead of n. In the generated stimuli, exponents ranged from P=1 (linear) to P=5 (higher degree of exponential behavior, see Fig. 2). The output of this formula is an array of normalised values that oscillate around 0; so the y-values of a cosine over time. Amplitude is not yet present in this output - this will be added later.

To ensure controlled stimuli that mimic natural sound, Matlab was first used to analyse a gliding tone played on a violin. Amplitude information of each harmonic at different time-points of the glide was extracted from the recording. Using the formula above, first an F0 was created and then twenty harmonics were added on top of F0 which started at 442Hz, so that F1 started at 884Hz, F2 at 1326Hz, etc. This created a matrix with 21 rows, as each harmonic was represented by one row. Towards the end of the glide of a violin, a constant resonant harmonic was present at 442Hz, which was added to this matrix as well, creating 22 rows in total. This matrix was then multiplied by the amplitude matrix extracted from the original violin glide, to mimic this spectrum.[3]

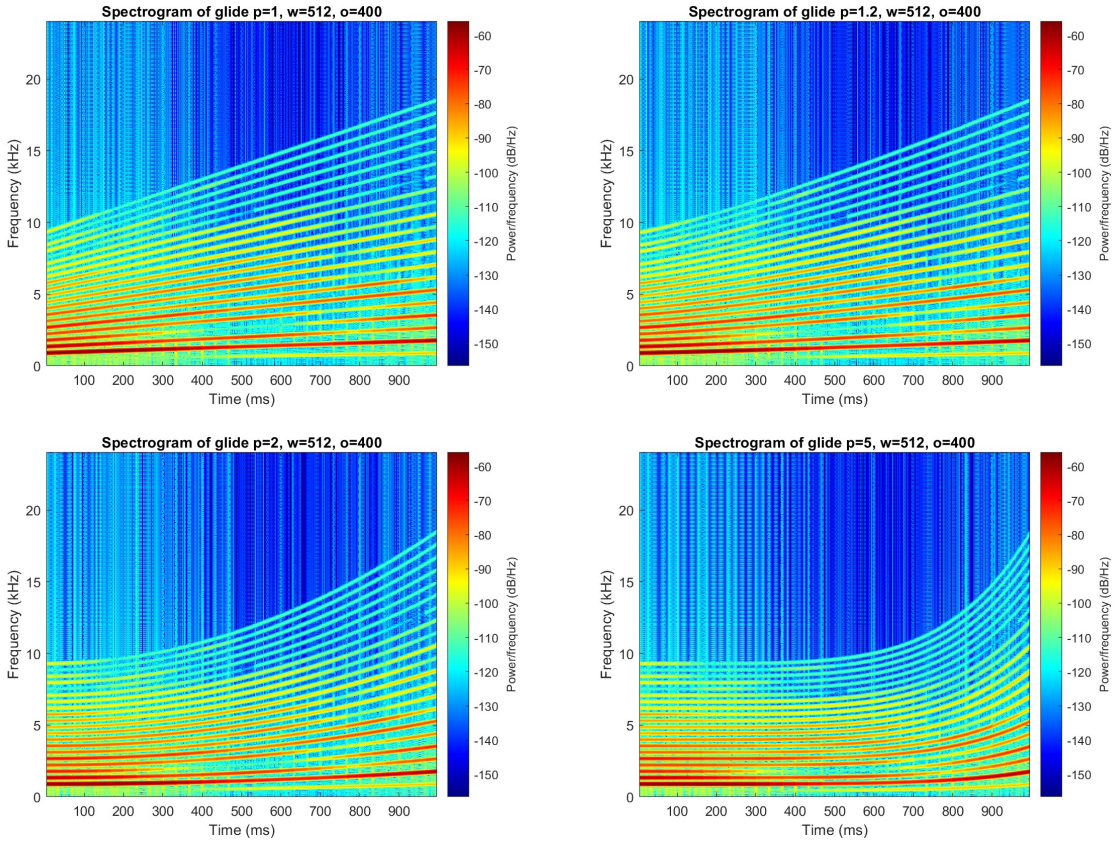

Figure 3: Spectrograms of four gliding tones with different exponents. Top left to right: P=1, P=1.2. Bottom left to right: P=2 and P=5. The x-axis shows time and the y-axis the frequency in kHz. The colours show the respective intensity of each overtone, with red being most intense and blue least.

[3]Contact the author for another essay that describes the analysis and generation of these stimuli in greater detail.

The replica was not perfect, as in the original sound more overtones were present. However, these would fall outside the range of the implant. This study aims to activate the whole range of the electrode array when CI users perform this test, which should easily be achieved with 20 harmonics (more on this later).

Figure 3 shows spectrograms of some examples of the generated gliding tones with different exponents. The weighted amplitudes of all harmonics are clearly visible in the changing colors, with red indicating a high intensity and blue a low intensity. Each glide lasted 1 second.

The ENT department of the hospital in Leiden has a specially developed model of the cochlea and the electrode array and speech encoding strategy (HiRes120) used by Advanced Bionics and their patients. Using this model, predictions could be made about the response of the array to the generated stimuli. Figure 4 shows electrodograms of the model, with on the y-axis the 16 electrodes of the array. The thickness of each line indicates the level of activation of each electrode. It can be seen that these stimuli indeed activate the whole array, and also which electrodes convey shifting frequency information. For instance, when looking at P=1 and P=5, there is a big shift in activation over the entire range of the array. For P=5 the activation of especially electrode 1 started a lot later than for P=1. In the higher electrodes, the stimulation of P=5 was more constant whereas P=1 shows more waveforms. This corresponds to the shapes of the spectrograms of Fig. 3.

When looking at P=1 and P=1.2 these lower electrodes already look more alike, although there is still a difference. Now both electrodograms show waves in the higher frequency-range, but these are still slightly shifted.

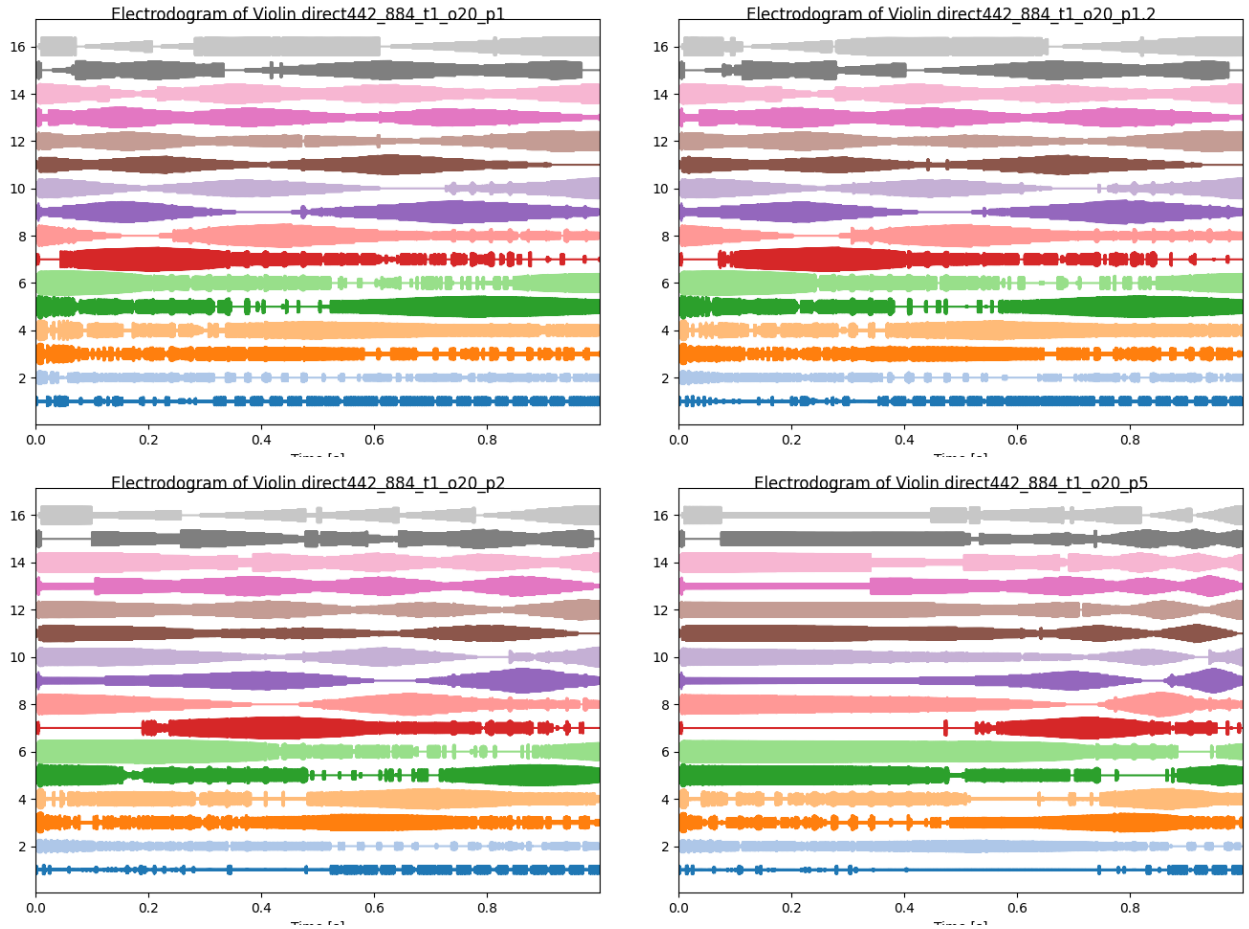

Figure 4: Electrodograms of four gliding tones with different exponents. From left to right: P=1, P=1.2 & P=2 and P=5. The y-axis shows the electrodes of the array and the x-axis time in seconds. A high electrode number corresponds to a band with higher frequencies.

The advantage of a gliding tone is that it avoids the possibility of musical recognition - an unwanted cue. Since the stimulus is a constant sound, there are no rhythmic differences between stimuli. It could be argued that a change of incline *over time* has an inherent rhythmic aspect. However, this would only be perceivable when pitch is used to detect the difference. Therefore, it can be argued that this proposed task is still a pitch task.

Another measure taken to prevent loudness cues was the addition of loudness roving of 10% to all stimuli. This roving was achieved by changing the amplitude table. This means that each stimuli will be played at a randomized intensity that will have a maximum of 10% difference to the other stimuli heard.

2.1.2 Procedure

The Glide Tone Test provides insight in the subjects' ability to detect a gradual change in pitch over time. It does this by finding the detection threshold of participants with a certain degree of accuracy. The higher the degree of accuracy, the higher the precision of the test. If the target is to reach a level where the subject will get the answer correct in 50% of the cases (known as a 50% point) a simple up-down procedure will suffice. A wrong answer will generate an easier stimulus and a correct answer a more difficult one. However, if a higher degree of accuracy (i.e. 70%) is desired, the procedure gets more complex. Now, stimuli will only be made more difficult when two correct answers in a row are provided (for the mathematics behind this, see Gelfand, Chapter 7).

In order to find the 70.7% point of each subject, another method described by (Gelfand, 2017) was used. Here, the test will increase in difficulty as long as correct answers are given by the subject. A wrong answer turns the direction of the test around - each stimulus gets easier and easier - until *two* correct answers in a row are given again. The examples given by Gelfand look for a dB detection threshold, but this same principle can be applied to find the JND between a linear and non-linear gliding tone.

A test that changes the stimulus presented to the participant based on their response, is called an adaptive procedure. Because the subjects are presented with 3 stimuli, this test is a 3-interval alternative forced choice experiment (forced choice because the answering is multiple choice).

To ensure speed of the test, a version of Parameter Estimation by Sequential Testing (PEST) was used. This adaptive procedure does not only change the direction of the stimulus like explained before, but also the step-size. Using larger steps in the beginning that become smaller with each direction change, ensures that the test swiftly goes to the region of interest (see Fig. 5). The rules of PEST were not strictly adhered to in this version, but rather used as guidelines to develop a fitting procedure for the GTT.

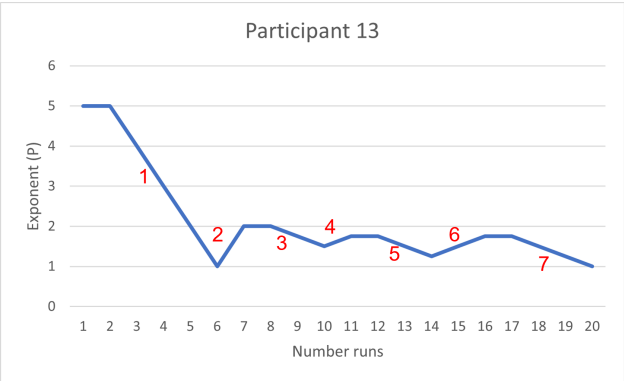

Figure 5: Example of psychometric curve, the red numbers indicate the number of direction changes. The x-axis shows the number of runs completed, the y-axis the exponent of the different sound. Please note that the reference stimulus had a P-value of 1, which means that values closer to P=1 indicate a better performance.

2.2 Development of test

Development of the test went hand in hand with its' evaluation, so that appropriate improvements to the test could be made when needed. This method led to one version of the test that used different stimuli than all other versions (Version 0). The following three versions do not differ in type of stimuli presented, but have different procedures used. This way accuracy of different procedures was assessed and possible learning effects were examined.

In this chapter, first the general procedure used in all versions will be discussed. Then the differences between different versions will be discussed. The evaluation methods will be explained in this chapter (see Section 2.3), and the results of these evaluations can be found in the next chapter (see Chapter 3).

2.2.1 Procedure

Although many aspects of the test were changed in different versions, some aspects remained the same throughout.

Software and hardware
All tests were conducted using the same laptop, with sounds played at a fixed volume in the free space of a quiet office. Playing the sound in free space enables CI users to follow exactly the same procedure as NH subjects, as both can listen to the stimuli as they normally would. Participants used a mouse to choose their answers. This test could be run on any computer that has Matlab installed on it. Excel was used to save the results.

Study design
At the start of each experiment, participants were informed about their participation in the experiment, and a hearing test was conducted to test for normal hearing ($<$20 dB hearing level in the frequency range of 125-8000 Hz on both ears). After receiving instructions, the participants performed a short practice run after which questions could be asked before the main task started. During both the practice run and the main task no feedback was given about the performance of the participants, but at the end of the experiments the results were shown when asked.

Task
The GTT is a three-interval three-alternative forced choice task, where each round contains two linear reference stimuli (P=1) and one non-linear stimulus ($P \neq 1$), presented in random order. Each stimulus was represented by a box on the screen that lit up while the stimulus played and was turned off after. Participants had to identify the odd sound by clicking on the matching box. Participants were instructed to only do this based on pitch and to neglect any loudness information they might detect. They were informed of the added loudness roving before the practice run so that sudden loudness changes between stimuli would not come as a surprise. All experiments started with the largest difference in P-value (or exponent) possible; P=1 as reference and P=5 as stimulus to be recognized. A correct answer would change the P-value of the odd stimulus to a value closer to P=1, and a wrong answer would have the opposite effect. It was hypothesized that stimuli with P-values that are closer to P=1 (for instance P=1.5) are more difficult to distinguish from P=1 than stimuli with a larger difference in P-value (such as P=4), and thus require better pitch perception skills.

The use of a combination of the two-up, one-down method with PEST (described more extensively in 2.1.2) meant that the participants had to do between 15-40 sets to find their JND at the 70.7% point.

2.2.2 Version 0

In the first version of the test, 3 participants were presented with stimuli that had a gliding tone of 4 seconds, with a 1 second on- and offset added to the sound. This made the total duration of the stimulus 6 seconds (see Table 2 and Fig. 6). Slopes varied from P=1 (linear and control) to P=5 with a smallest step-size of 0.25.

2.2.3 Version A

In this version the on- and offset were removed from the sound so that only the glide part of the sound remained (see Fig. 7). Stimuli now last 1 second and no break was allowed once the test had started. 15 participants completed this version of the experiment.

2.2.4 Version B

In order to see if a learning effect was present in this test, two participants conducted the experiment 4 times, spread over 4 different (not always consecutive) days. For more accurate measurements, smallest step-size was set to 0.2.

2.2.5 Version C

9 participants conducted the same test as version B, two times in a row after doing a practice run. This way the most accurate way of measuring could be assessed by comparing this to version A. Version C had the option to pause the test whenever they wanted.

2.2.6 Overview

	Version 0	Version A	Version B	Version C
Stimuli	t=6 seconds onset and offset	t=1 only glide	t=1 only glide	t=1 only glide
Procedure	Short practice run One full measurement No breaks	Short practice run One full measurement No breaks	Short practice run One full measurement done four days No breaks during experiment	Short practice run Two full measurements Breaks possible
Slope (P)	range = 1-5 stepsize = 0.25	range = 1-5 stepsize = 0.25	range = 1-5 stepsize = 0.2	range = 1-5 stepsize = 0.2
Estimated duration	30 min for 1 run	10 min for 1 run	10 min for 1 run	10 min for 1 run
Participants	3	15	2	9
Evaluation	Feedback Graphs	Feedback Graphs Calculated averages	Feedback learning effect analysis	Feedback Graphs Calculated averages

Table 2: Overview of versions of tests

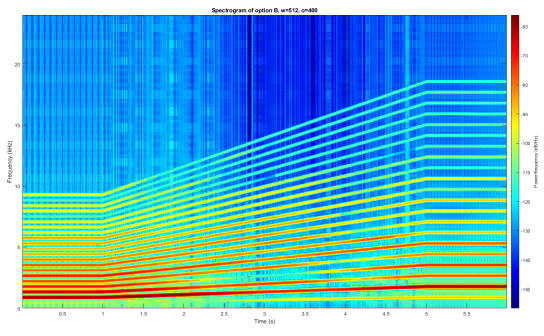

Figure 6: Spectrogram of a linear gliding tone used in version 0 of the experiment.

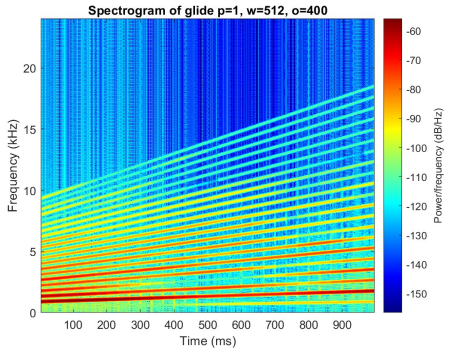

Figure 7: Spectrogram of a linear gliding tone used in version A, B and C of the experiment.

2.3 Evaluation

After each full experiment, participants were asked first for general feedback about the experiment and then more detailed questions about the interface, ease, clarity, duration of the test, etc. Psychometric curves of each run were made in Excel (see Fig. 5). These curves show for each run (x-axis) which P-value was presented to the subject (y-axis). As long as the line goes down, the subject has given a correct response (as the reference stimulus has a P-value of 1 and P-values closer to this will be harder to recognize than those with a larger difference in exponent), and an upward line indicates a wrong response. It can be seen in Fig. 5 that after a wrong response in run 6, two correct responses at the same level in run 7 and 8 are required before a more difficult stimulus was presented to the subject in run 8. This figure also shows the adaptive PEST procedure used; up until run 7, step-sizes are large with a difference of 1 per run. After run 8, the step-size is 0.5 and after run 10, the step-size became 0.25 per run. The red numbers count direction changes and show that the test stops automatically after 7 flips. Average scores were calculated by discarding the first two peaks and adding a last 'virtual' peak. Standard deviations were calculated using SPSS.

To investigate (significant) differences between groups, SPSS was used. This program was also used to look at trends and possible correlations in the data set.

2.3.1 Loudness

The stimuli themselves were developed in such a way that they should contain as little loudness cues as possible. Additionally, loudness roving was added to the experiment. The effect of loudness roving was investigated using SPSS. All stimuli were presented at a randomized intensity within a 10% range of one another. This means that there always was a loudest, a softest and a middle sound (volume-wise). This distribution was randomized. There was always one correct response, which was also randomized. This combination had as consequence that some participants would have had more loudest-correct answers than others (although the size of the group should have balanced this difference out if the group was big enough).

For example: one subject might have had 40% of the correct answers played at the loudest intensity, and another subject 15% of the correct answers. If it is assumed that the test become easier when the correct answer (or the non-linear stimulus) was the loudest one of the set, this would indicate that people with a high percentage loudest-correct answers would achieve better average scores. So a subject with 40% loudest-correct answers would score higher than a subject with 15%.

To examine this hypothesis more closely, a scatter-plot of the average scores per test versus the percentage of loudest-correct answers was made. This scatter-plot should show if any linearity between loudest-correct answer and average score achieved was present. The null hypothesis was that there would be no linear relationship between loudness and score (the desired outcome for this test). The alternative hypothesis was that there was a linear relationship between loudness and score. If there was a linear relationship, the direction of the slope would become relevant. A downward or negative linear relationship indicates that scores were closer to P=1 if the percentage of loudest-correct answer increased. In other words, subjects performed better if more of the correct answers were also the loudest answers. If the relationship is positive, the opposite is true: subjects performed worse when more of the correct answers were also the loudest ones. A regression analysis using SPSS investigated the significance of any observed linearity.

3 Results

3.1 Subjects

In total 26 participants took one of the versions of the experiment. All participants were of higher education (except for one who was at high school) and had NH. Subject 12 had 30 dB hearing loss. All subject demographics are shown in Table 3.

Version 0 used different stimuli than any of the other experiments. Therefor these participants and results will not be included in any of the analyses. Of the remaining 23 subjects, 16 were female and 10 were male, the age range 15-55 years, with a mean of 29.42 years (SD=8.08).

Subject	Age	Gender	Hearing	Education	Experience	Version
1	27	M	NH	conservatory	P	0
2	24	M	NH	university	A	0
3	26	M	NH	university	NM	0
4	26	M	NH	conservatory	P	A
5	23	F	NH	university	A	A
6	30	M	NH	university	NM	A
7	25	F	NH	university	NM	A
8	27	F	NH	university	NM	A
9	27	M	NH	university	NM	A
10	34	F	NH	university	NM	A
11	42	F	NH	university	NM	A
12	26	M	NH*	university	NM	A
13	27	M	NH	conservatory	P	A
14	26	F	NH	university	A	A
15	26	F	NH	university	NM	A
16	26	F	NH	university	NM	A
17	15	F	NH	high school	NM	A
18	55	F	NH	college education	NM	A
19	19	F	NH	conservatory	P	B
20	22	F	NH	conservatory	P	B
21	45	M	NH	university	NM	C
22	38	F	NH	university	NM	C
23	28	M	NH	university	NM	C
24	32	M	NH	university	NM	C
25	26	F	NH	university	NM	C
26	31	F	NH	college education	NM	C
27	33	M	NH	university	NM	C
28	28	M	NH	university	NM	C
29	28	F	NH	university	A	C

Table 3: Subject demographics. Subject 12 had 30 dB hearing loss on one side. NM = non-musician, A = amateur musician, P = professional musician, F = female, M = male, NH = normal hearing.

3.2 Results per version

3.2.1 Version 0

The overall feedback of this version of the experiment was that the stimuli were too long (t=6) and it was difficult to remember the first stimulus when the last one was playing. Participants started developing strategies to be able to answer, which makes the test more cognition than detection oriented. On top of that, the onset and offset of the sound provided the subjects with clues that were not wanted. Linear sounds had a harsh kink when the sound went from static to glide, which was not present in any of the non-linear sounds. Based on this feedback, it was decided to completely change the stimuli in any following versions of the experiment. In all following versions, on- and offset of stimuli were removed and the sounds were a lot shorter (t=1).

3.2.2 Version A

All 15 participants indicated that the test was clear and easy to understand, although some still had difficulty when the first stimulus offered was the non-linear sound. Some subjects requested a repeat or pause button.

All individual psychometric curves of Version A can be found in Fig. 9, and all averages and SD's can be found in Fig. 8. The average scores indicate the slope that with a 70.7% confidence can be identified as different when compared to a linear slope (P=1). This means that scores closer to 1 indicate a higher detection sensitivity then scores closer to 5.

The overall average score was P=2.2 (SD=1.00 or 4times step-size).[4] The musicians (n=4) scored an average of P=1.59, SD=0.277 (1step-size) and non-musicians (n=11) scored average P=2.42, SD=1.14 (4step-size). The difference between these groups was found to be significant ($p=0.046$), although the groups were quite unbalanced, with a ratio of 3:1 of non-musicians versus musicians.

Fig. 10 shows the relationship between average scores and the percentage of loudest-correct answers in the test. The graph shows a small downward relationship, which would mean that a higher percentage of loudest-correct answers improves the score of the test. However, the regression was not found to be significant (F=0.310 and $p=0.587$).

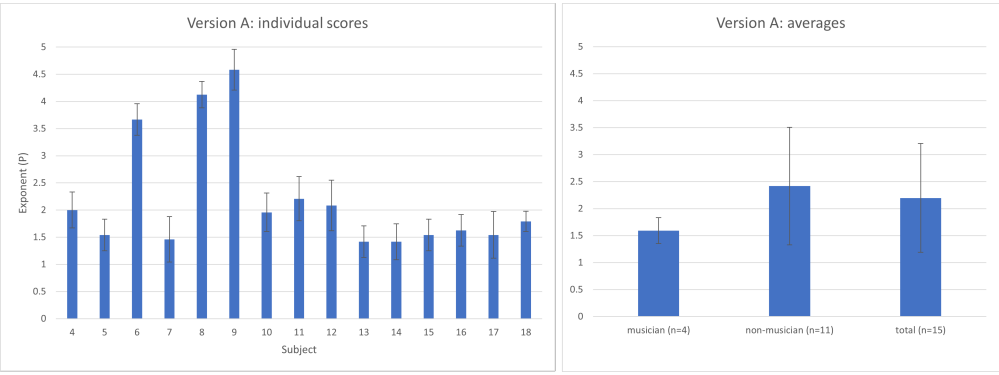

Figure 8: Average JND scores with SDs of Version A per subject, and group averages.

[4]Because the step-size determines the specificity of the test, it is more useful to think the deviation in amount of steps, than as a numerical value. In Version A, step-size was set at 0.25 see Tab. 2.

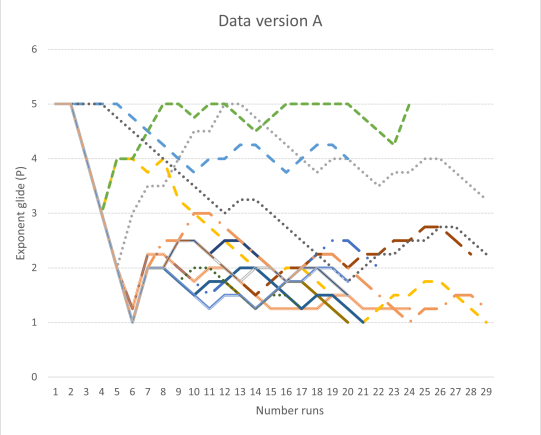

Figure 9: All psychometric curves of version A.

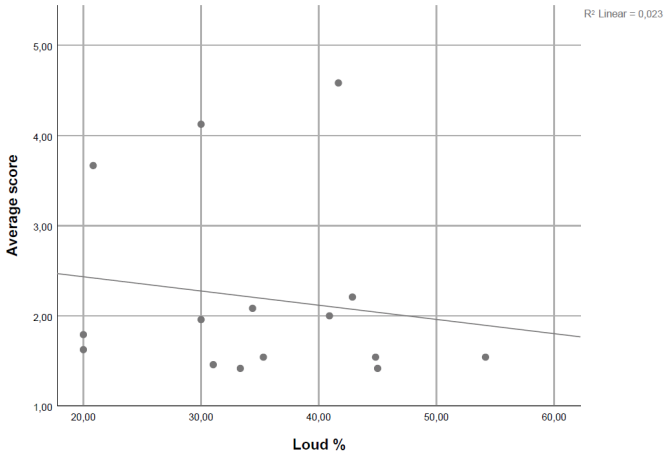

Figure 10: Scatterplot of average scores and the percentage of loudest-correct present in that test. Note that a score closer to 1 on the y-axis indicates a better performance of the task.

3.2.3 Version B

Two subjects completed version B of the test (with smallest step-size now set to P=0.2), which was conducted to investigate learning effects over four days of testing. On the last day S20 indicated that they were very tired and distracted after completion of the test, and both indicated boredom over time when asked for feedback. As can bee seen in Fig. 11 and 12, all scores lie relatively close to each other, with the lowest SD's for day 2 and highest SD's for day 1 (S19) and day 4 (S20) (see Tab. 4).

Because both subjects are musicians and n=2, there is no option of looking at a difference between groups.

To look at the relationship between loudest-correct percentage and correct answers, each separate trial was plotted (see Fig. 13). An ANOVA analysis showed no significant regression (F=1.196, *p=0.316*).

	Day 1	**Day 2**	**Day 3**	**Day 4**
Subject 19 average (SD)	1,28 (0.30)	1,24 (0.15)	1,32 (0.20)	1,44 (0.27)
Subject 20 average (SD)	1,32 (0.27)	1,44 (0.15)	1,44 (0.15)	1,6 (0.42)

Table 4: Average JND of P-value per day with (SD) of Version B.

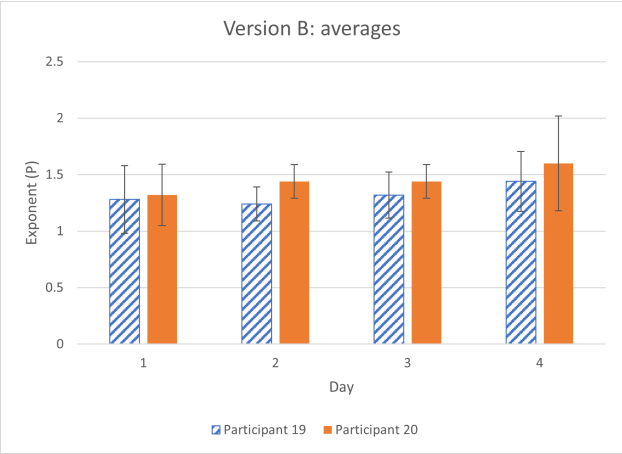

Figure 11: Version B: average scores with SD per day.

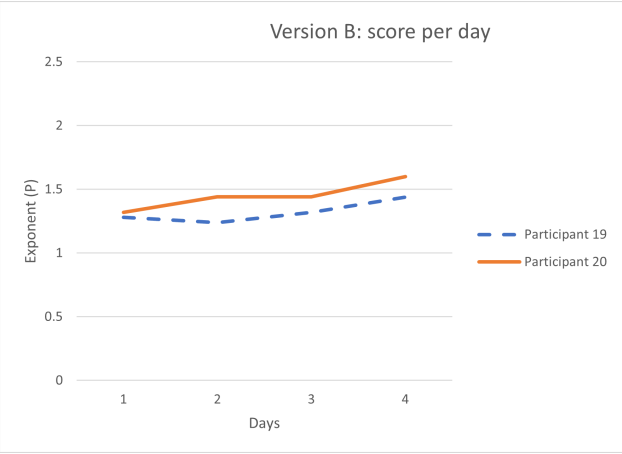

Figure 12: Version B: closer look at effect over time on slope.

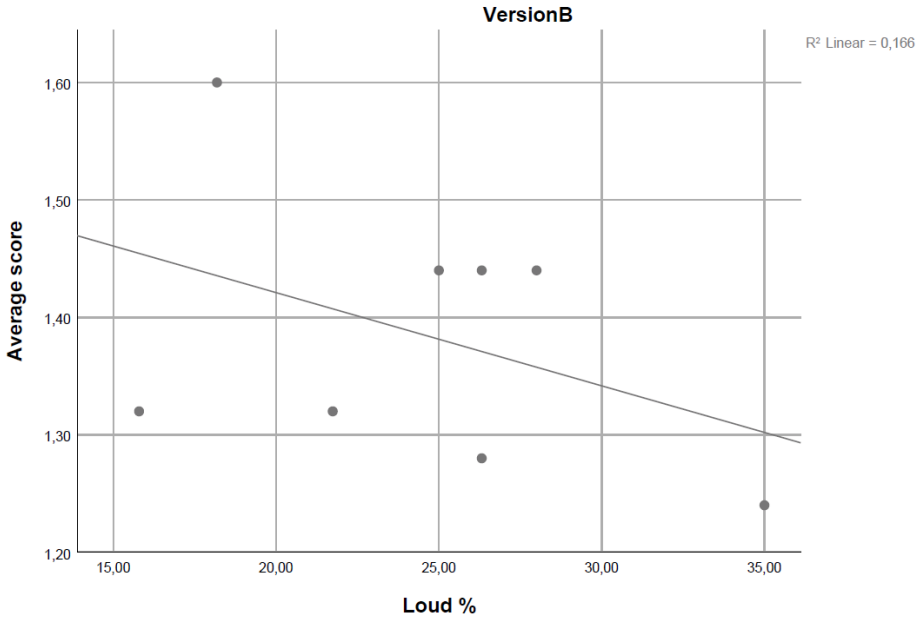

Figure 13: Scatterplot of average scores and the percentage of loudest-correct present in that test. Note that a score closer to 1 on the y-axis indicates a better performance of the task.

3.2.4 Version C

Version C of the test had the same smallest step-size as version B, but an added pause button. Participants were positive in their feedback about this feature. All subjects (n=9) completed the test twice in a row (see Fig. 14). Four participants had a higher second score, five a lower one. Scores per run can be found in Tab. 5 and the average score of both runs combined can be found in Fig. 15.

Because only one subject was a musician in the group, looking at significant differences between musicians and non-musicians is not possible.

Again each separate trial was plotted (see Fig. 16) to investigate the relationship between loudest-correct and scores. An ANOVA analysis showed a significant regression (F=8.303, $p=0.011$). This regression is positive, which means that a higher percentage of loudest-correct answers resulted in worse scoring of the test.

Subject	Run 1	Run 2	Difference
21	1.73 (0.38)	1.23 (0.27)	0.50
22	2.17 (0.21)	1.50 (0.30)	0.67
23	1.40 (0.35)	1.93 (0.32)	0.53
24	1.30 (0.19)	1.20 (0.16)	0.10
25	1.63 (0.21)	1.70 (0.19)	0.07
26	1.40 (0.20)	1.23 (0.18)	0.17
27	1.47 (0.15)	1.67 (0.19)	0.20
28	1.83 (0.18)	1.47 (0.32)	0.37
29	1.23 (0.14)	1.43 (0.14)	0.20

Table 5: Averages JND of P-value for both runs with (SD) and the difference between them for each subject that completed version C.

Figure 14: Version C: both average scores per subject.

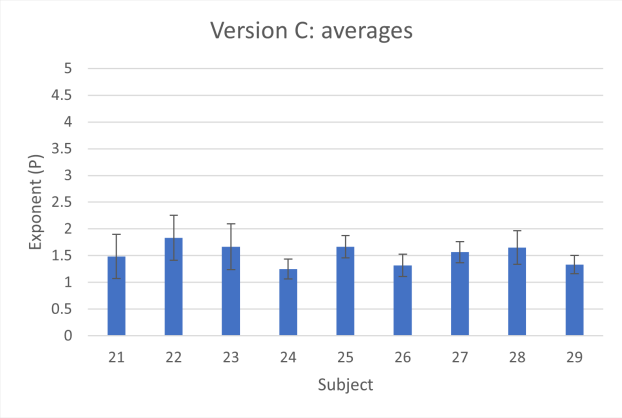

Figure 15: Version C: average scores per subject.

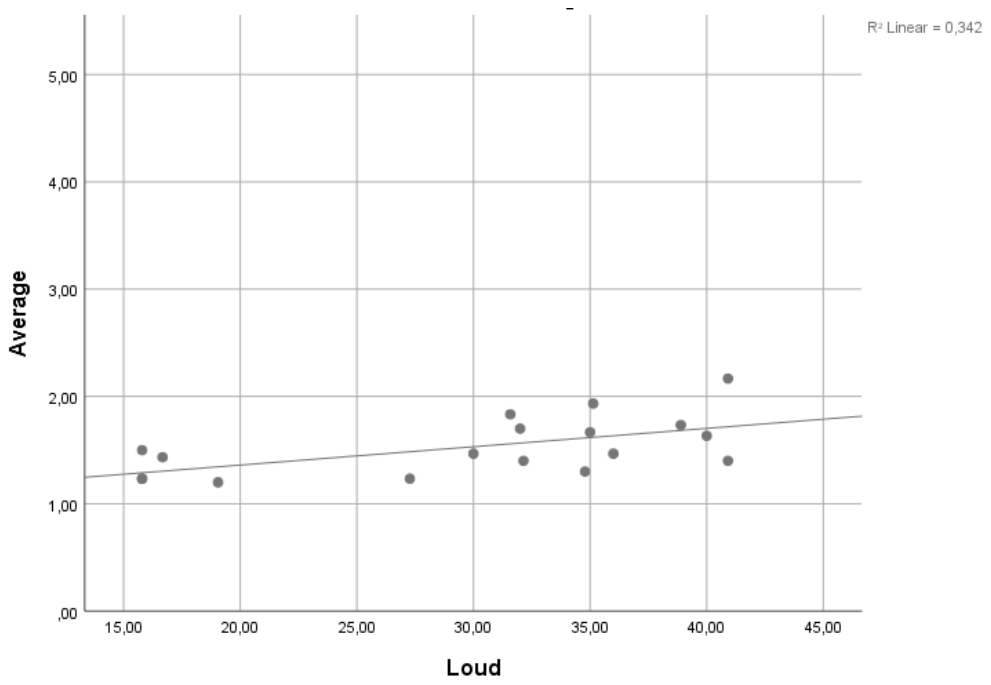

Figure 16: Scatterplot of average scores and the percentage of loudest-correct present in that test. Note that a score closer to 1 on the y-axis indicates a better performance of the task.

3.3 Overview

Five subjects were not able to distinguish exponents of P>2.00 with 70.7% confidence. The highest score was for S9 (P=4.58, SD=0.37 or 1+ step-size), the lowest score belonged to S24 (P=1.25, SD=0.18 or less than 1 step-size), which means that S24 performed best on this test. The only non-normal hearing subject was S12. All average scores can be found in Fig. 17.

An overview of all individual scores, their SD's and their SD's converted to step-size can be found in Figure 6.

There was no big difference between the procedures of version A, B and C of the test. B and C had exactly the same stimuli, and A only had a different step-size. It might therefor be interesting to look at all results combined. Fig. 18 shows all different runs that were completed (so one per participant for version A, 4 per participant for version B and 2 per participant for version C). An ANOVA analysis showed no significant regression (F=1.265, $p=0.268$).

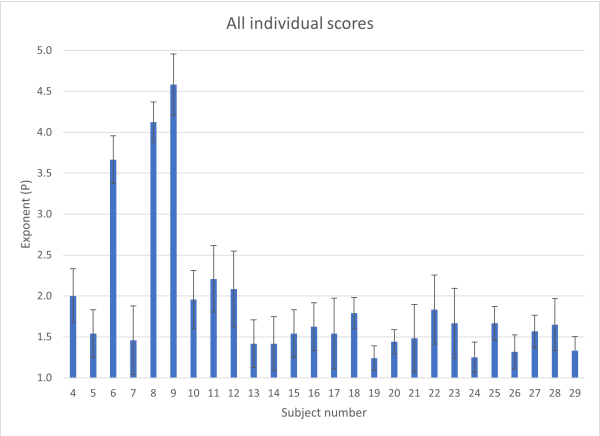

Figure 17: All averages per subject

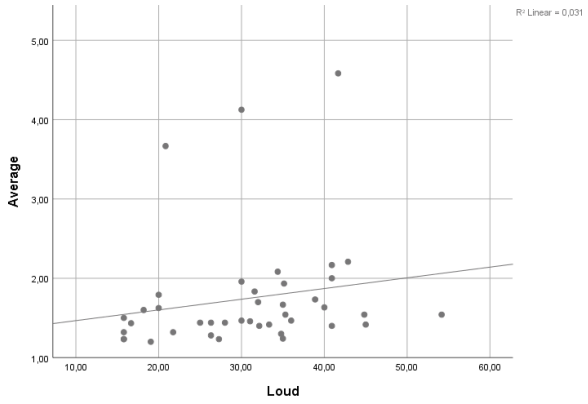

Figure 18: Scatterplot of average scores and the percentage of loudest-correct present in all tests. Note that a score closer to 1 on the y-axis indicates a better performance of the task.

Subject	Score	SD	SD in stepsize
4	2.00	0.33	1.33
5	1.54	0.29	1.17
6	3.67	0.29	1.17
7	1.46	0.42	1.67
8	4.13	0.24	0.98
9	4.58	0.37	1.50
10	1.96	0.35	1.41
11	2.21	0.41	1.62
12	2.08	0.46	1.85
13	1.42	0.29	1.17
14	1.42	0.33	1.33
15	1.54	0.29	1.17
16	1.63	0.29	1.17
17	1.54	0.43	1.72
18	1.79	0.19	0.75
19	1.24	0.15	0.75
20	1.44	0.15	0.75
21	1.48	0.41	2.06
22	1.83	0.42	2.11
23	1.67	0.43	2.13
24	1.25	0.18	0.92
25	1.67	0.21	1.03
26	1.32	0.21	1.04
27	1.57	0.20	0.99
28	1.65	0.32	1.59
29	1.33	0.17	0.85

Table 6: All average JND P-values and standard deviations of each subject. Please note that this table gives the combined averages for Version B and C.

4 Discussion

4.1 Interpreting results

4.1.1 Version 0

The three participants who took this version of the test indicated that it was too complex to do. Considering the demographics of these subjects (all of NH with the highest level of education) this was a worrying sign that did not bode well for the average CI user. On top of that, the unwanted cue that especially the onset of the sound added to the task made this version unusable for further experiments.

Because all following versions used different stimuli, the results obtained from version 0 will not be compared or discussed further.

4.1.2 Version A

All subjects that participated in this version of the test found it easy to understand - although not always easy to do. Most participants had clear psychometric curves that seemed to give a good indication of their score. However, there were some participants whose start of the test was strange. For some it took 4 tries to recognize the difference between P=1 and P=5, but after the initial difficulty, they scored really well (see Fig. 19). Sometimes the opposite was also the case, with almost maximum scores in the beginning that then went back to the lowest scores at the end (see Fig. 20). These curves indicate that maybe the first measurement is not always a good indication of a subjects' ability.

When looking at the standard deviations of this version, only 2 out of 15 participants had a SD that was smaller than one step-size (12.5% of the subjects.) All other participants had a SD that was between 1 and 2 steps (see Tab. 6). The one subject with hearing impairment (S12) did not score specifically low, but had the largest SD. There were three participants that are outliers (S6, S8 and S9) who scored worse than the other subjects. However, neither the demographic data, nor the standard deviation suggests that these data-points are invalid. These subjects were just less good at the task than most other participants, and were therefore not excluded from the analysis.

On average the musicians scored better than the non-musicians. This difference was significant ($p=0.046$), although the unbalanced group size needs to be remembered. What can also be observed is that the musicians scored closer to each other (average 1 step-size apart), whereas non-musicians have a bigger variability in their group (4 step-sizes apart). A conclusion can be that all musicians are more likely to score good at this test, but not all non-musicians are good at it.

As discussed in Section 2.3, regression analysis was conducted to see if loudness-roving affected the participants. It was assumed that the test would be easier when the loudest sound was the non-linear sound (called loudest-correct in this paper). This would mean that a higher percentage loudest-correct leads to better test scores. Fig. 10 shows that the percentage of loudest-correct was between 20-45 % approximately. Looking at this figure, there seemed to be no linear relationship between loudest-correct and score. ANOVA analysis confirmed that there was no linear correlation. However, when looking at the outliers, it can be seen that the highest-percent loudest-correct lead to the worst score. For the remainder of the group, there seemed to be no increase or decrease in performance based on the percentage loudest-correct given. This was the desired outcome of the test, since this seems to indicate that these participants did not use loudness as a cue.

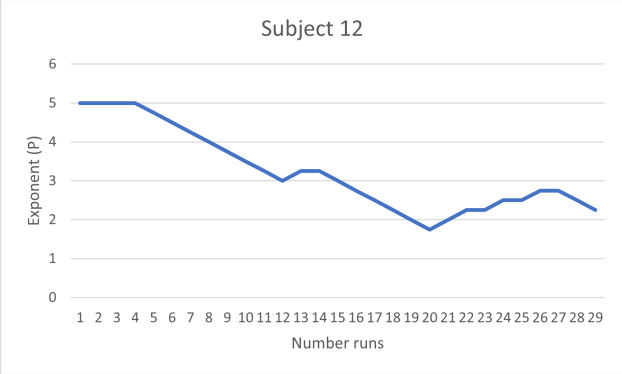

Figure 19: Individual psychometric curve of subject 12.

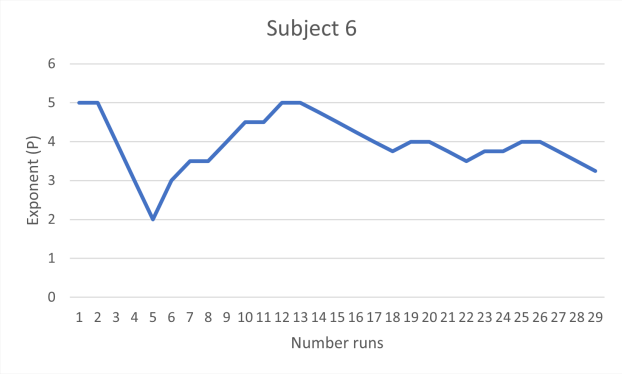

Figure 20: Individual psychometric curve of subject 6.

4.1.3 Version B

Version B was developed to investigate possible learning effects when the test is taken more than once. The main feedback of this version was that the test becomes boring with repetition, which was to be expected. Rather than seeing a gradual improvement with repetition over time (the expected learning effect), there seemed to be a gradual decrease of performance, but this was not significant (see Fig. 12). This decrease can be explained by the boredom of the subjects, who possibly were less focused as time went on. There seemed to be reduced SD between day 1 and day 2 for both participants (see Table 4). After that, the SD went back up again, which likely comes from diminished focus of participants and therefor a larger spread of results.

The loudest-correct versus average score plot shows a seemingly random distribution of points (see Fig. 13). No regression was found, which indicates that both subjects probably did not get influenced by the loudness roving in a way that affected their performance of the test.

4.1.4 Version C

Subjects said that they liked the pause button (although it was never used) and did not indicate that the test lasted too long when done twice. For some, it was motivating as they wanted to improve on their previous score, whereas others found it a bit boring the second time. This is reflected in Fig. 14, where 4 subjects improved their performance and 5 didn't. The test including practice and explanation never took more than 30 minutes in total.

Version B found no clear learning effect, but only the expected in-test variability. Therefor it was decided to calculate the score of version C by combining the data from both runs the participants completed.

This group did not have such clear outliers as present in version A. All participants scored between P=1.25 (S24) and P=1.83 (S22). This indicates that all subjects did very well on the test. It was expected that running the test twice would lead to more accurate and more reliable results. Remarkably, the expected improvement of accuracy was not found. In this version of the experiment, 3 of the participants (33.3%) had a SD smaller than one step-size, but also 33.3% had a SD of more than 2 step-sizes, which was never observed in version A or version B. Based on the assumption that for some repeating the test is motivating, while for others it becomes boring, the difference between both scores was looked at for these participants. It can be hypothesized that a worse score the second run, as well as a big difference between both runs, indicates boredom. However, there was no such connection found. Of the 3 subjects that had a high SD, two improved and one worsened with repetition. Subjects with low SD had a similar random division. It seems that repetition does not necessarily improve accuracy.

In this version of the test, a positive regression was found between loudest-correct and average scores (F=1.265, p=0.268). This seems to indicate that a higher percentage of loudest-correct answers presented to the subject, lead to a worse average final score. This is an interesting find, as it was hypothesized that more loudest-correct answers would make the test easier and therefor lead to better average scores. It is possible that the instruction the participants received to ignore any loudness cues they found lead to this upside-down effect; whenever the loudest answer was the correct answer, they were inclined to discard this option. However, when looking at the spread of data-points on the graph (see Fig. 16), the effect seems clear

but not large, as the spread of results is between P=1 and P=2 roughly. This means that although there might be a small effect, the participants all still scored very well on the test. Another option to keep in mind is that this regression found is a coincidence, as it was not found in the other two versions of the test which use the same stimuli.

4.1.5 Overview

Because the nature of the stimuli was the same - even though the step-size was not - it is possible to combine all results for a complete overview analysis. The main complication is that in different versions, subjects completed the test a different number of times. Version B has 4 iterations of results, version C 2 and version A only one. To not loose or cherry-pick desired data points, it would be best to include all iterations. The downside of this decision is that the presence of S19 and S20 becomes 4 times stronger than those of version A and 2 times stronger than those of version C. Due to variability between iterations, it is not recommendable to just multiply scores of version A by 4 to balance this out.

The best scoring subject was S19 in version B (P=1.24) and the worst scoring subject S9 in version A (P=4.58). However, the main purpose of this paper is not to assess the performance of the individual, but rather the performance of the test overall.

As mentioned in Section 3.3, and Fig. 18, no regression was found when all results are combined. In other words, there was no relationship between the loudest-correct percentage offered to the participants and their final average score.

4.2 Evaluation of test requirements

In the design of the used stimuli, the main requirements of the test were integrated. All stimuli used the whole electrode array to complicate place-pitch percepts. Rhythm cues should not be present. The fixed starting and ending frequency and added loudness roving should limit (hopefully eliminate) loudness cues, which was investigated. As mentioned before, in version A and B no regression was found between answers and loudness. Version C had a small regression between loudest-correct answers and scores, which means that the participants who were presented with *more* loudest-correct answers, scored worse on the test. With a small group, this effect could be a coincidence, and when all group data was combined, this significance of regression disappeared.

When developing a new test, the reliability of the test is essential. Reliability is high when repeated measurements yield the same results. This means that version A of the test does not say a lot about the reliability of the test. Version B and C provide more information on this and show that there are differences when the test is repeated, but these differences are relatively small. Version B has a maximum difference of approximately P=0.3 (the difference between day 1 and day 4 of S20, see Tab. 4), and version C a maximum difference of P=0.6 (see Tab. 5).

One of the best ways to assess the accuracy of the score found will be to look at the standard deviation of the score. The smaller the SD, the more reliable the score. Any SD smaller than 1 step-size is most desirable, as this means that even one step of the stimulation away from the final score would already overshoot the target. Version B provided the most accurate result with two SD's of 0.75 step-size. However, asking participants to come back four days in a row is not a desirable outcome of the test.

Surprisingly, version A (with only one iteration) had overall a smaller SD than version C. Version C had *more* SD below one step-size, but also more SD's above 2, which is not wanted. It is unclear why a double iteration sometimes leads to more accurate and sometimes to less accurate scores. It could be a specific feature of both groups (as the group sizes are small and of different size). If this is the case, more experiments with version C should give more high-accuracy scores. It is also possible that this variation is just a part of repetition, and that the scores of version A would be similar if all participants had done the experiment twice. Indeed, when looking at the SDs of the individual runs of version C (see Appendix Table 7), no SD in step-size is above 1.89 (S21, run 1), and there are 9 (out of 18) SD in step-size that are lower than 1. Therefor it can be concluded that the higher SD's of version C when compared to version A are the effect of inter-test variability. This however does not mean that just choosing the score with the lowest SD of version C will give a better indication of a subjects JND. It merely shows that people are not perfect and

will score different on the same task when they have to repeat it. This confirms that version C will give a better representation of the subjects' JND than version A - despite the sometimes higher SD values.

A procedural requirement was that the test should last under one hour to be used in the clinic or included in existing test batteries. The average time of version C of this test was 30 minutes.

It is recommended to use version C of the test in future experiments.

4.3 Strengths and weaknesses

The stimuli used in this experiment are more complex than most of the stimuli found in literature when it comes to the presence of harmonics. The added weights to the harmonics create a naturalistic sound, that makes the listening task more related to daily life and listening skills. This combination suggests that only place-pitch perception will not be enough to distinguish between different stimuli. This effect can be seen in the model predictions of the electrodes of Fig. 4. The use of a glide tone, rather than a melody makes the test very objective.

Results were processed as soon as they were available which enabled direct correction of unwanted features when necessary. Version 0 is a prime example of the benefit of such an approach. Only 3 subjects' results are now unusable, instead of the entire group. This process also meant that things like learning effect, breaks for participants, etc. could be assessed and added when needed. This resulted in a test that is easy to understand for subjects, but not always easy to do. The adaptive procedure ensures that every participant reaches the limit of their perception which meant that everyone, both high and low scorers, found the test tricky towards the end. This is exactly the purpose of the test, as it wants to measure the perceptual limit of slope detection.

The test itself provides a beautiful balance between controlled circumstances and naturalistic perception. Most music perception tests test more than just perception of pitch (for instance the recognition of specific melodies does not only require pitch perception but also knowledge of the fragments provided (Kong, Cruz, Jones, & Zeng, 2004)), and are thus hard to use as objective measurement. On the other hand, a lot of specific pitch perception tasks are so clinical that they might not necessarily reflect the skill-level of the subject in less-than-perfect circumstances of daily life. In this test, the combination of a gradual changing pitch with naturalistic overtones that synthetically created (and thus completely controlled by the researcher) provide a complex listening task that *only* measures skill without needed knowledge. Knowing a lot about children's songs will not make a subject better at glide tone perception. This enables the researchers to compare results between participants as is done for pitch perception tasks.

The use of gliding tones has another interesting benefit, as tones that gradually change their frequency are an integral part of tonal languages. These stimuli are of course more abstract than a language, so association of words can not be used to detect changes, as no words are spoken (this was a problem for Wong and Soli). Another benefit is the wide range of application - because the test does not test language perception but rather the mechanism needed to speak the language - non tonal-language speakers could participate in this test as well. This broadens the amount of available and suitable subjects for the test.

A downside of this research is that the group of subjects used is not a very representative group of society. All participants have a high degree of education and are either involved with similar research themselves in the ENT department, or are (professionally) trained musicians. This combination of participants is therefor probably one of the highest scoring groups that can be found. However, the positive of this is that, although participants scored really high, none of the participants obtained the maximum available score (P=1.1) which indicates a smaller likelihood that this limit will be reached.

Another downside of this research is that the experiment so far is only done by NH subjects and not yet by CI users. However, now a benchmark is found that can be used to compare CI patients to.

4.4 Future recommendations

Translation to CI

The most important next step in the assessment of this test will be to ask CI patients to participate in the experiment. The entire procedure of all versions can be copied directly to a CI patient, which allows for a comparison between results obtained in this experiment to new subjects. Considering the results of the model as well as the literature available, it is expected that the difference between P=1 and P=5 is detectable for CI users. However, if this is not the case, it will always be possible to add more levels above P=5.

It would be interesting to see how CI patients with different speech-encoding strategies perform in this test. The current hypothesis is that TFS information is needed to be able to distinguish between stimuli, but this has not yet been tested. Comparing speech-encoding strategies that provide different amounts of TFS might prove this hypothesis right or wrong.

Stimuli

It is possible to set the step-size of the experiment even smaller for the last part of the experiment. Because the test follows an adapted PEST procedure (see Section 2), the step-size in the beginning of the experiment is set larger than at the end. At the moment, version C has a smallest step-size of 0.2 but this can of course be set to 0.1 or any desired value to obtain higher accuracy. It needs to be remembered that this could mean that the test becomes a bit slower than before.

The current stimuli are based on the sound spectrum of a recorded violin. If more research into music perception is done, it could be interesting to also mimic the spectrum of other instruments such as a clarinet or a trombone (both instruments are capable of playing a glide tone, unlike a piano or a harp). If the spectrum of a human voice is mimicked, this test becomes more (tonal) language related.

Currently the F0 of the stimuli used was 442-884Hz, based on the central A used in music. However, when looking at the human voice, a lower frequency range becomes of interest. Gliding tones are an integral part of tonal languages, so shifting the gliding tone down and spreading it over a narrower range might provide interesting information for tonal language speakers. It is important to note that the F0 of a human voice (around 80-255 Hz (Fitch & Holbrook, 1970)) often falls outside the range of the implant. Only the overtones are encoded and provided to the listener.

Another change to the stimuli would be to change the direction of the glide from downward to upward (so for instance from 884-442Hz). This becomes more relevant when the translation to tonal languages is investigated.

Analysis

In the future experiments it is recommended to not only save the loudness, but also save the specific response of each subject per question (so A, B or C). This way it will be possible to see if there is a patient that will only click on the loudest answer or not. With the current information this analysis is not possible. The only thing that can be said is whether or not loudest-correct answers helped the subjects in their task.

When this test is given to new subjects, it might be interesting to combine it with existing psychophysical tests. The stimuli by design contain pitch information which changes over time. This makes it a spectrotemporal test. It can be argued that this combination of factors makes it unclear what is measured; is it the subjects pitch perception or the subjects time perception that is measured? By combining this test with other tests, more insight might be obtained about the dominant predictor for the ability to take this test. The prediction here is that that will be pitch perception, but it is important to verify this.

5 Conclusion

Pitch perception is crucial for functioning in society and quality of life. However, for many CI patients, complex pitch perception is a problem due to limited spectral resolution and insufficient fine structure information. Many new speech-encoding strategies have been developed over the last years to solve this problem using TFS. This also lead to the development of new TFS and music perception tasks, but many are either too subjective or too clinical.

The purpose of this paper was to develop a test that combines skills needed for naturalistic music perception with an objective task. Complex stimuli based on a gliding tone of a violin were developed in such a way that they contain no unwanted cues. Based on model predictions, the stimuli should activate the entire electrode array of the implant, which is exactly the type of stimulation that is difficult for CI users and requires TFS information.

The task was to distinguish different exponents of glide tones, using the shifting frequency as a cue. In this experiment, normal hearing subjects conducted different versions of the experiment. This way learning effects and test accuracy could be assessed. Evaluation of the test showed that the glide tone test (GTT) is quite fast (approximately 30 minutes), easy to understand and contains no unwanted cues. Regression analyses show no overall relationship between the added loudness roving and results, but more in depth analysis would be recommended for future experiments. JNDs were found for the NH subjects, which provides a benchmark against which results of CI users could be compared.

The GTT provides insight in the way complex tones are perceived by the subjects. The stimuli used are easily translated to tonal language research, which makes the test applicable for pitch and music and tonal language experiments, while remaining accessible to anyone.

6 Acknowledgements

I would like to thank my supervisors dr. ir. Richard Hendriks of the TU Delft and dr. ir. Jeroen Briaire of the LUMC for helping me with the whole project. Also a special thank you to the research group of the LUMC, and specifically to Bram Knipscheer and Savine Martens (for running my stimuli through the model). I would also like to thank my family and friends for proofreading several versions of this thesis and for asking obvious questions that I should have seen coming (but didn't).

7 Appendix

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
S4	6	1,50	2,50	2,0000	,35355
S5	6	1,25	2,00	1,5417	,29226
S6	6	3,25	4,00	3,6667	,30277
S7	6	1,00	2,00	1,4583	,43060
S8	6	3,75	4,50	4,1250	,30619
S9	6	4,00	5,00	4,5833	,40825
S10	6	1,50	2,50	1,9583	,36799
S11	6	1,50	2,75	2,2083	,43060
S12	6	1,50	2,75	2,0833	,49160
S13	6	1,00	1,75	1,4167	,30277
S14	6	1,00	2,00	1,4167	,34157
S15	6	1,25	2,00	1,5417	,29226
S16	6	1,25	2,00	1,6250	,34460
S17	6	1,00	2,25	1,5417	,43060
S18	6	1,50	2,00	1,7917	,18819
S19	20	1,00	1,80	1,3200	,25464
S20	20	1,00	2,20	1,4500	,29647
S21	12	1,00	2,20	1,4833	,43029
S22	12	1,00	2,40	1,8333	,44176
S23	12	1,00	2,60	1,6667	,44586
S24	12	1,00	1,60	1,2500	,19306
S25	12	1,20	2,00	1,6667	,21462
S26	12	1,00	1,80	1,3167	,21672
S27	12	1,20	2,00	1,5667	,20597
S28	12	1,00	2,00	1,6500	,33166
S29	12	1,00	1,60	1,3333	,17753
Valid N (listwise)	6				

Figure 21: All results per subject.

Subject	Run 1 SD	Run 1 SD step-size	Run 2 SD	Run 2 SD step-size
21	0.38	1.89	0.27	1.34
22	0.21	1.07	0.30	1.50
23	0.35	1.73	0.32	1.60
24	0.19	0.96	0.16	0.82
25	0.21	1.07	0.19	0.96
26	0.20	1.00	0.18	0.90
27	0.15	0.75	0.19	0.94
28	0.18	0.90	0.32	1.60
29	0.14	0.69	0.14	0.69

Table 7: Table that shows the SD's and the SD expressed in step-size for the individual runs of version C. The SD in step-size = SD/0.2

References

Arnoldner, C., Kaider, A., & Hamzavi, J. (2006). The role of intensity upon pitch perception in cochlear implant recipients. *The Laryngoscope*, *116*(10), 1760–1765.

Boëx, C., De Balthasar, C., Kós, M.-I., & Pelizzone, M. (2003). Electrical field interactions in different cochlear implant systems. *The Journal of the Acoustical Society of America*, *114*(4), 2049–2057.

Boisvert, I., Reis, M., Au, A., Cowan, R., & Dowell, R. C. (2020). Cochlear implantation outcomes in adults: A scoping review. *PLoS One*, *15*(5), e0232421.

Briaire, J. J., & Frijns, J. H. (2006). The consequences of neural degeneration regarding optimal cochlear implant position in scala tympani: a model approach. *Hearing research*, *214*(1-2), 17–27.

Carlyon, R. P., & Deeks, J. M. (2002). Limitations on rate discrimination. *The Journal of the Acoustical Society of America*, *112*(3), 1009–1025.

D'Alessandro, H. D., Ballantyne, D., Portanova, G., Greco, A., & Mancini, P. (2022). Temporal coding and music perception in bimodal listeners. *Auris Nasus Larynx*, *49*(2), 202–208.

Dieter, A., Duque-Afonso, C. J., Rankovic, V., Jeschke, M., & Moser, T. (2019). Near physiological spectral selectivity of cochlear optogenetics. *Nature communications*, *10*(1), 1–10.

Drennan, W. R., & Rubinstein, J. T. (2008). Music perception in cochlear implant users and its relationship with psychophysical capabilities. *Journal of rehabilitation research and development*, *45*(5), 779.

D'Alessandro, H. D., Ballantyne, D., Boyle, P. J., De Seta, E., DeVincentiis, M., & Mancini, P. (2018). Temporal fine structure processing, pitch, and speech perception in adult cochlear implant recipients. *Ear and hearing*, *39*(4), 679–686.

Fitch, J. L., & Holbrook, A. (1970). Modal vocal fundamental frequency of young adults. *Archives of Otolaryngology*, *92*(4), 379–382.

Fitzgerald, M. B., Shapiro, W. H., McDonald, P. D., Neuburger, H. S., Ashburn-Reed, S., Immerman, S., ... Svirsky, M. A. (2007). The effect of perimodiolar placement on speech perception and frequency discrimination by cochlear implant users. *Acta oto-laryngologica*, *127*(4), 378–383.

Füllgrabe, C., & Moore, B. C. (2017). Evaluation of a method for determining binaural sensitivity to temporal fine structure (tfs-af test) for older listeners with normal and impaired low-frequency hearing. *Trends in hearing*, *21*, 2331216517737230.

Galvin III, J. J., Fu, Q.-J., & Nogaki, G. (2007). Melodic contour identification by cochlear implant listeners. *Ear and hearing*, *28*(3), 302.

Gelfand, S. A. (2017). *Hearing: An introduction to psychological and physiological acoustics*. CRC Press.

Gfeller, K., Turner, C., Mehr, M., Woodworth, G., Fearn, R., Knutson, J. F., ... Stordahl, J. (2002). Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults. *Cochlear implants international*, *3*(1), 29–53.

He, S., Xu, L., Skidmore, J., Chao, X., Riggs, W. J., Wang, R., ... Warner, C. (2020). The effect of increasing pulse phase duration on neural responsiveness of the electrically-stimulated cochlear nerve. *Ear and hearing*, *41*(6), 1606.

Jiam, N., Caldwell, M., Deroche, M., Chatterjee, M., & Limb, C. (2017). Voice emotion perception and production in cochlear implant users. *Hearing Research*, *352*, 30–39.

Jung, K. H., Cho, Y.-S., Cho, J. K., Park, G. Y., Kim, E. Y., Hong, S. H., ... Rubinstein, J. T. (2010). Clinical assessment of music perception in korean cochlear implant listeners. *Acta Oto-Laryngologica*, *130*(6), 716–723.

Kang, R., Nimmons, G. L., Drennan, W., Longnion, J., Ruffin, C., Nie, K., ... Rubinstein, J. (2009). Development and validation of the university of washington clinical assessment of music perception test. *Ear and hearing*, *30*(4), 411.

Kong, Y.-Y., Cruz, R., Jones, J. A., & Zeng, F.-G. (2004). Music perception with temporal cues in acoustic and electric hearing. *Ear and hearing*, *25*(2), 173–185.

Looi, V., Rutledge, K., & Prvan, T. (2019). Music appreciation of adult hearing aid users and the impact of different levels of hearing loss. *Ear and Hearing*, *40*(3), 529–544.

Luo, X., Masterson, M. E., & Wu, C.-C. (2014). Melodic interval perception by normal-hearing listeners and cochlear implant users. *The Journal of the Acoustical Society of America*, *136*(4), 1831–1844.

MacDonald, R. A. (2013). Music, health, and well-being: A review. *International journal of qualitative studies on health and well-being*, *8*(1), 20635.

Moon, I. J., & Hong, S. H. (2014). What is temporal fine structure and why is it important? *Korean journal of audiology*, *18*(1), 1.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, *9*(4), 399–406.

Moore, B. C., & Sek, A. (2009). Development of a fast method for determining sensitivity to temporal fine structure. *International journal of audiology*, *48*(4), 161–171.

Mudry, A., & Mills, M. (2013). The early history of the cochlear implant: a retrospective. *JAMA Otolaryngology–Head & Neck Surgery*, *139*(5), 446–453.

Nimmons, G. L., Kang, R. S., Drennan, W. R., Longnion, J., Ruffin, C., Worman, T., ... Rubinstein, J. T. (2008). Clinical assessment of music perception in cochlear implant listeners. *Otology & Neurotology*, *29*(2), 149–155.

Pickles, J. (1998). *An introduction to the physiology of hearing*. Brill.

Rubinstein, J. T. (2004). How cochlear implants encode speech. *Current opinion in otolaryngology & head and neck surgery*, *12*(5), 444–448.

Shannon, R. V., Fu, Q.-J., & Galvin 3rd, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta oto-laryngologica. Supplementum*(552), 50–54.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Thompson, A. C., Wise, A. K., Hart, W. L., Needham, K., Fallon, J. B., Gunewardene, N., ... Richardson, R. T. (2020). Hybrid optogenetic and electrical stimulation for greater spatial resolution and temporal fidelity of cochlear activation. *Journal of Neural Engineering*, *17*(5), 056046.

Vaerenberg, B., Pascu, A., Del Bo, L., Schauwers, K., De Ceulaer, G., Daemers, K., ... Govaerts, P. J. (2011). Clinical assessment of pitch perception. *Otology & Neurotology*, *32*(5), 736–741.

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C., Joris, P. X., Heinz, M. G., & Plack, C. J. (2019). The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. *Hearing research*, *377*, 109–121.

Wagner, L., Altindal, R., Plontke, S. K., & Rahne, T. (2021). Pure tone discrimination with cochlear implants and filter-band spread. *Scientific Reports*, *11*(1), 1–8.

Wang, W., Zhou, N., & Xu, L. (2011). Musical pitch and lexical tone perception with cochlear implants. *International journal of audiology*, *50*(4), 270–278.

Wei, C.-G., Cao, K., & Zeng, F.-G. (2004). Mandarin tone recognition in cochlear-implant subjects. *Hearing research*, *197*(1-2), 87–95.

Winn, M. B., & O'Brien, G. (2022). Distortion of spectral ripples through cochlear implants has major implications for interpreting performance scores. *Ear and hearing*, *43*(3), 764–772.

Wong, L. L., & Soli, S. D. (2005). Development of the cantonese hearing in noise test (chint). *Ear and hearing*, *26*(3), 276–289.

Wouters, J., McDermott, H. J., & Francart, T. (2015). Sound coding in cochlear implants: From electric pulses to hearing. *IEEE Signal Processing Magazine*, *32*(2), 67–80.

Zhang, F., Underwood, G., McGuire, K., Liang, C., Moore, D. R., & Fu, Q.-J. (2019). Frequency change detection and speech perception in cochlear implant users. *Hearing research*, *379*, 12–20.