MATHILDE LE GAL DE KERANGAL

Pitch sensitivity and context

Internship report

Supervisor: Alain de Cheveigné

August 7, 2015

Table of Contents

Acknowledgments

At this point of my research, I would like to thank Alain de Cheveigné for his admirable patience, motivation, continuous support, and great knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not imaging having a better advisor for this project.

I am very grateful to all the students and members of the LSP for their encouragements, advice and contribution to a great atmosphere in the lab.

This project was a great learning and human experience and I truly enjoyed working in auditory research.

Abstract

Sounds in our environment are often perceived as having a pitch. Pitch is fundamental to appreciate musical melodies, to segregate sound sources and to analyze auditory scenes. Different procedures can be used to measure pitch discrimination performance. In one classic procedure, the listener is presented with a sequence of two tones, the first tone is the standard and the second the target, and the listener must judge if the pitch goes up or down. Performance is quantified as a function of the frequency difference between these last two tones, ignoring the history of tones presented on previous trials.

The current study has two aims. The first is to validate a new procedure in which the listener is presented with a sequence of tones and must make a judgment after each tone (rather than after each tone pair as in the classic procedure), as to whether its pitch is higher or lower than that of the previous tone. The second aim is to determine whether pitch judgments depend on a longer history of tones presented prior to each judgment, and if so to quantify this dependency.
The new procedure (judgment after each new tone) has two potential advantages over classic pitch procedures. One is that the experimental yield is higher as there are more judgments per unit of time. The other is that each tone is followed by a judgment, rather than each tone pair (or triplet or quadruplet), so that it is easier to analyze the effects of earlier tones in a systematic way. With the classic procedure (pair of tones followed by a judgment) the analysis would need to distinguish between effects of tones playing different roles whereas in the new procedure every tone plays the same role. The uniform nature of the procedure is useful for our second aim, exploring context effects, and may also be of use for studies involving EEG or MEG.
Most studies interpret pitch discrimination performance as a function of the frequency difference between the last two, three or four tones (depending whether a two-, three-, or four-interval procedure is used) and ignore the effect of previous tones. However it has long been known that performance is degraded if the frequency of standard tones is roved rather than fixed [6], implying some influence from prior context. More recently, in [24], Raviv and Ahissar investigated the bias produced by recent history and concluded that the judgment in a two tone discrimination is built on a comparison between the last tone and an exponentially-decaying average of the just previous tone and past tones. In our study, this bias is studied in terms of pattern of tone presentation.
In our study, eighteen paid listeners participated in the experiments. Discrimination thresholds with the new procedure did not differ significantly from those from a classic procedure with comparable parameters. There is no indication that the new procedure entails a penalty. This result is of practical interest because the new procedure is more efficient (in terms of experimenter's or subject's time) than the classic one.
In addition to the two main procedures (classic and new), we ran several variant procedures for which the context was manipulated. In one variant of the classic procedure, the first tone of every pair was the same (1

kHz). This yielded the lowest thresholds. In a second variant of the classic procedure, the first tone of each pair was identical to the second tone of the previous pair. This also gave low thresholds. In a variant of the new task, near-threshold interval values alternated with relatively large intervals (1 semitone). This gave the largest thresholds. Together these results confirm that prior context affects pitch judgments. Furthermore, a detailed analysis of the pattern of responses over time for the new task revealed a systematic bias.

This study confirms that pitch discrimination performance is affected by prior context. It also validates a new procedure to measure the limits of pitch discrimination, possibly more efficient and better suited to certain studies than classic procedures.

CHAPTER 1

Introduction

1.1 Introduction

1.1.1 What is pitch ?

Many sounds surrounding us are perceived as having a pitch. Pitch does not refer to a physical attribute of sound, but rather to a sensation related to the periodicity of a sound waveform [21].
In music, sequences of pitches over time form melodies, and simultaneous combinations of pitches form the basis of harmony. Musical instruments, such as strings or winds, produce a pitch. In speech, vowels are "voiced" and can be connected to a pitch. This feature of sound can tell us about the characteristics of a speaker such as his/her gender, age or identity. Pitch also carries prosody in languages such as English, and gives lexical information in languages such as Mandarin. In auditory scene analysis, pitch contributes to grouping in sound source segregation: it provides a perceptual dimension along which different sources can be distinguished and followed over time [23].

Although pitch is one of the most studied topics in hearing research, mechanisms underlying this perceptual attribute remain sources of controversy and heated debates, and a precise definition is difficult to give. However, the definitions of pitch fall into two categories: those that refers to music, and those that avoid reference to music.
According to the American Standards Association (ASA (1960)) pitch is

”that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale.”

This definition relates explicitly pitch and music: pitch is not a physical attribute of a sound but an attribute of sensation.
More recently, the American National Standards gave a quite broad definition of pitch (ANSI, 1994, page 34):

"Pitch is that attribute of auditory sensation in terms of which sounds may be ordered in a scale extending from low to high. Pitch depends mainly on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus"

This definition underlines that pitch is linked to the frequency of a sound rather than to the intensity or its loudness, for example.
A definition that does not refers to music but based on a pure-tone comparison is given in [7]:

"A sound can be said to have a certain pitch if it can be reliably matched by adjusting the frequency of a pure tone of arbitrary amplitude"

Two tones generally have the same pitch if they share the same fundamental frequency (F0), despite differences in timbre and loudness. So, in the case of complex tones, the matching is made with a pure tone

having the same F0.

1.1.2 Parametric effects on the pitch of pure tones

Interactions between several physical parameters of a sound complicate the relationship between a physical stimulus and its perceived pitch. It is known that pitch depends mainly on the frequency, but intensity and duration influence pitch, although the effects are small.
In most of the following studies, pitch discrimination performance is determined by measuring the frequency difference limen or FDL. The FDL is the smallest detectable frequency difference between two tones, [19]. The FDL can be plotted in terms of absolute frequency difference (in Hertz) or as a proportion or percentage of the baseline frequency.

Pitch and frequency

Pitch varies with pure-tone frequency, but this variation is not linear: a given change in frequency may not produce the same change in the magnitude of pitch. This assessment was studied by Stevens [26] establishing the subjective "mel scale", see Fig 1.1. In this scale 1000 mels was fixed arbitrarily at 1000 Hz. In his experiments, listeners were required to adjust the frequency of a comparison tone until the pitch sounded half that of the standard.
The mel scale underlines that pitch varies with frequency but is put into question. Musical intervals are defined as function of a frequency ratio (for example an octave is a doubling in frequency), and most musicians would not claim that an octave, a fifth or a semitone sound larger or smaller than another depending on frequency. It is probably for that reason that the mel scale did not become as popular as the comparable sone scale for loudness [8].

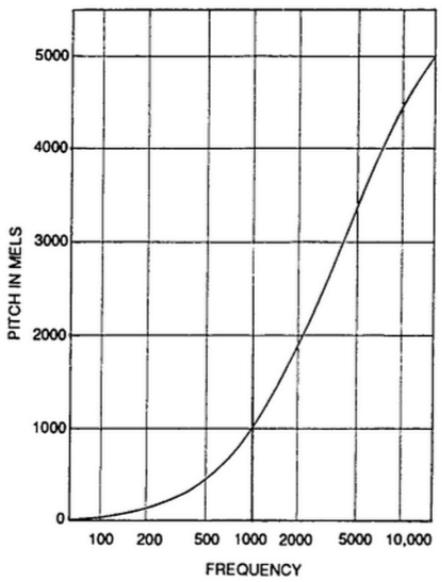

Figure 1.1: The relation of pitch in mels to the frequency of a pure tone in Hz, from [8].

Pitch is not evoked on all the audible range of frequency extending from 20 Hz to 20 kHz for young listeners. As defined by Plack in [22], "the existence region of pitch" extends from about 30 Hz to approximately 5000 Hz. Pitch is evoked in a bounded frequency range and, in this range, listeners' performance to

discriminate pitch varies as a function of frequency. As shown in Fig 1.2(a), discrimination is best in the 0.5 to 2 kHz range, and degrades at higher frequencies.

Tone duration and Level

Pure tone frequency discrimination thresholds depend also on duration, the dependency being greatest at low and high frequencies (Fig 1.2(b) from [18]). The dependency on level is of a different type (value rather than discriminability) and usually rather small, [17], [8].

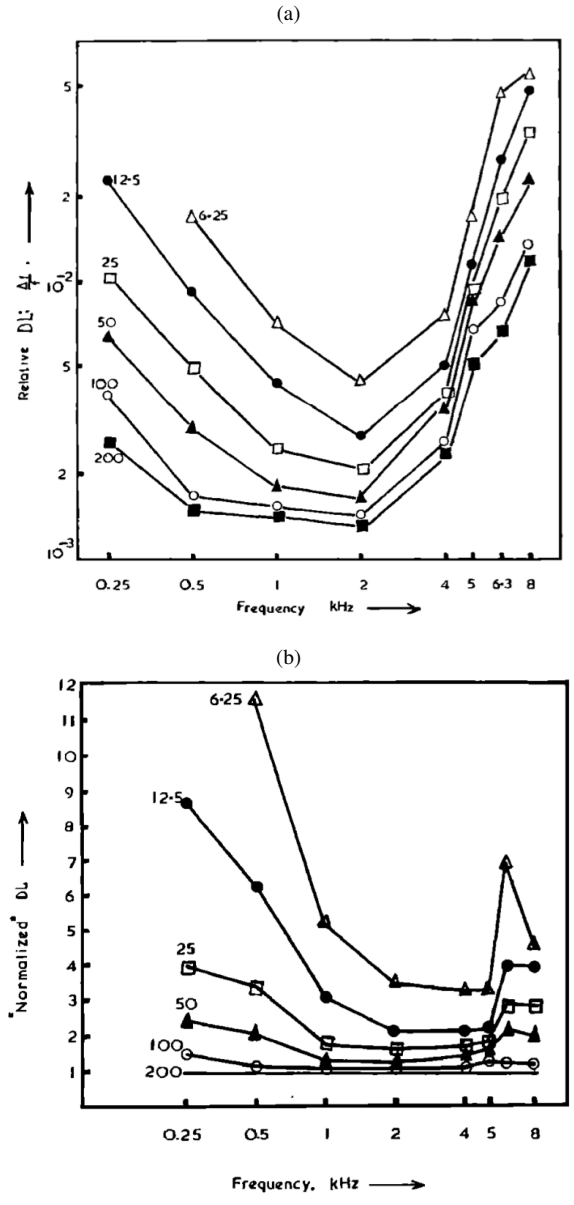

Figure 1.2: Data from the same subject for both figures, from [18] :
-Fig 1.2(a) : Plot of the relative frequency difference limen as a function of frequency. The parameter is tone's duration in milliseconds.
-Fig 1.2(b) : Plot of the frequency DL normalized by dividing the DL by the DL for a duration of 200 ms.

1.2 Variability of performances between subjects

A large inter-listener variability is often observed in pitch discrimination [14]. The difference of performance between groups of subjects can be partly explained by musical expertise.

Pitch plays a fundamental role in most forms of music because pitch variations over time create melodies and chords. Musicians are used to evaluate sounds along the pitch dimension each time they play their instrument. Accordingly, it is expected that musicians show substantially enhanced performance in pitch discrimination tasks, compared to non-musicians. This intuition is partly confirmed in the study [16] that investigates the influence of musical and psychoacoustical training on pitch discrimination using pure and complex tones.
Although some of the non-musicians are in the same range of thresholds as the musicians, Fig 1.3 (from [16]) makes apparent the large variability of results in the non-musician's group. For pure tones, nearly half of the non-musicians had thresholds in the same range as the musicians but in the case of complex tones, this proportion is closer to a third.
Most of the time, musical training helps at being more precise at discriminating pitch, which is observed by having smaller FDL.

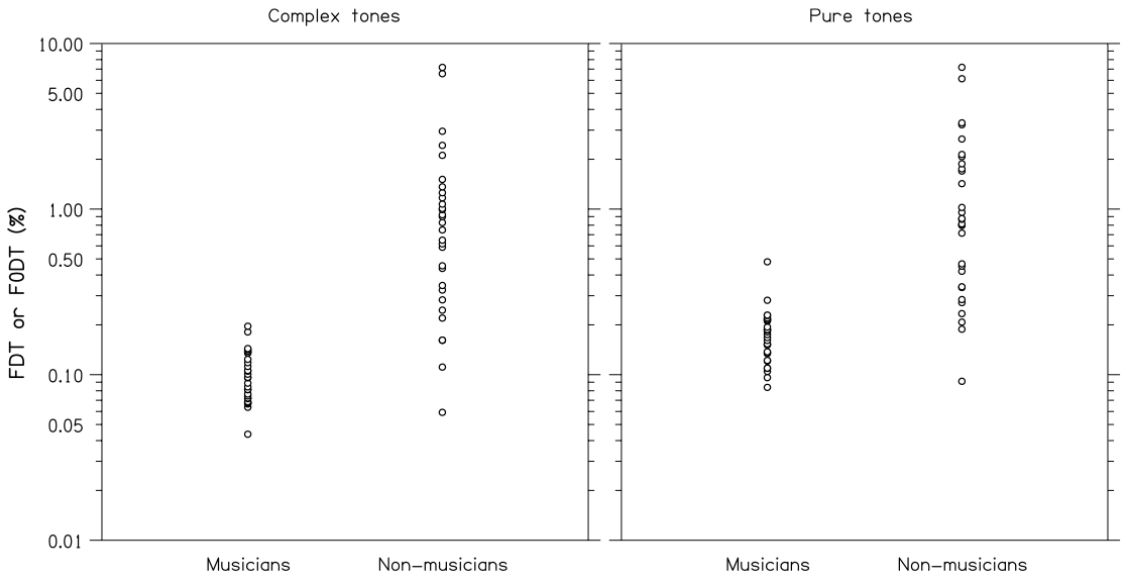

Figure 1.3: Individual FDTs for pure tones (right hand) and F0DTs for complex tones (left hand) in musicians and non-musicians. Each data point corresponds to an individual listener, and was computed as the geometric mean of the 10 threshold measurements, [16]

Some studies investigate the case of "pitch direction impaired" listeners [25], [14], [15]. Those listeners were able to detect a change in pitch but were not able to indicate the direction of the change. In [25], two tasks were presented, the first was to detect frequency changes, the second was to identify the direction of frequency changes. For the so called "pitch direction impaired" listeners, the identification FDL was from 4.7 to 9.2 times larger than the detection FDL.

1.3 Context effects on pitch discrimination

In most studies, pitch discrimination performance is evaluated as a function of the frequency difference between the last two, three or four tones (depending whether a two-, three-, or four-interval procedure is used) without taking into account the effect of previous tones. Judgments are considered independent between each other but it has been noticed that the previous tone history can have significant effects on judgment and therefore on performance.

This phenomenon has been observed quite early by Harris (see [6]). Harris investigated the effect of inter-tone duration on performance using a fixed or roved frequency of the standard tone. The fixed standard stimulus had a frequency of 1 kHz, the roved frequency was varying from 950 to 1050 Hz, in 5-cps steps. The fixed standard was always presented first. The results are presented in Fig 1.4 and show a considerable difference in effect of elapsed time. With the fixed standard, it appears that performance is little affected by the passage of time, as it declines by less than 1 Hz after 15 seconds. On the other hand, with the roving standard, the performance deteriorates rapidly, after 3 seconds the performance in this condition is comparable to the one of the fixed standard after more than 15 seconds. Roving the frequency of the standard tone is then detrimental compared to fixing it.

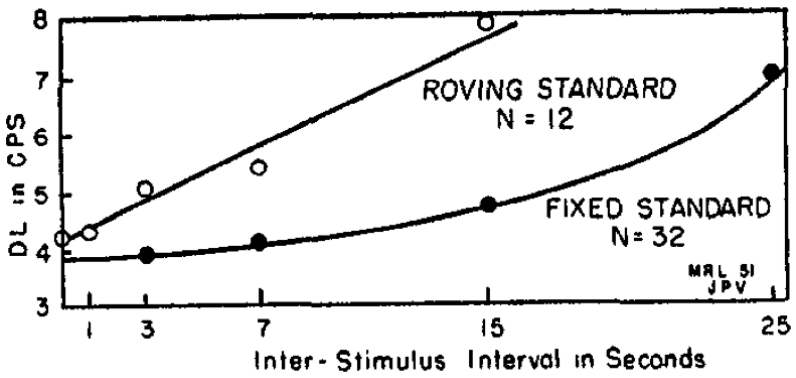

Figure 1.4: Pitch sensitivity at 1 kHz. An effect of inter-stimulus duration and roving frequency standard is observed. (CPS = circle per second = Hz) from [6].

The roving standard effect has also been observed in [3]. In this study, a standard tone (S) was presented, followed by an interference interval (I), followed by a comparison tone (C). The I interval was either a tone or a blank (silence). The results are presented in Fig 1.5. The trends are consistent with the one presented by Harris in [6] and confirm the worsened performance with a roving S. It also shows that inserting a tone instead of a silence between two tones to be compared deteriorates the performance.

More recently, this phenomenon has been observed in [25] when Semal and Demany were studying the case of "pitch direction impaired" listeners. The authors stress that roving did not affect all subjects the same way, and, as expected, the "direction impaired" listeners' performance was much more affected by roving than for the other listeners. For fixed standards, the "direction impaired" listeners showed better performance, even if they still reach higher thresholds than the other listeners. To explain this aspect of the study, the authors suggested a learning hypothesis: when the first element of each pair is always the same, direction impaired listeners can learn to label upward and downward changes, but when the first element differs widely from trial to trial, this learning is difficult or impossible.

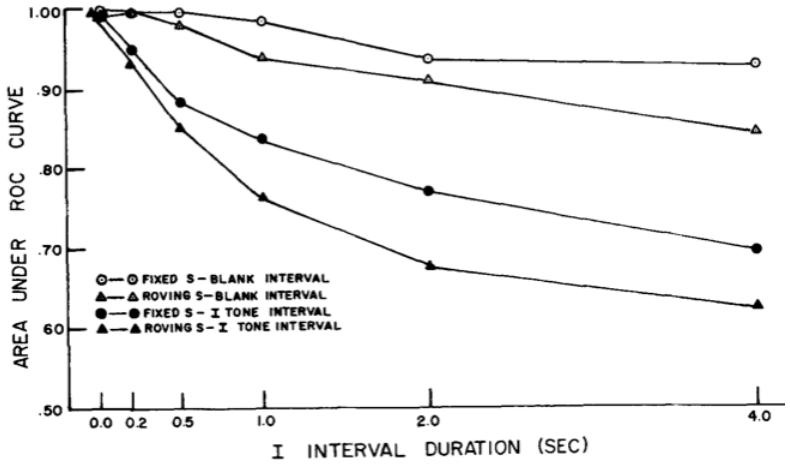

Figure 1.5: Area under the receiving operating characteristic (ROC) curve, average for four listeners, as a function of I-interval duration for the fixed S and Roved S, and for I-blank interval and I-tone-interval, from [3].

With listeners having poor discrimination performances, Amitay in [1] found that stimulus variability was deleterious for learning. In this study, the subjects were trained for either fixed standard, roved standard (the frequency varied around 1 kHz and took the values 900 Hz, 950 Hz, 1000 Hz, 1050 Hz or 1100 Hz), or widely-roved standard (the frequency was taking the values 570 Hz, 840 Hz, 1170 Hz, 1600 Hz, or 2150 Hz). The learning transfer across tasks was investigated. The "good" listeners had comparable thresholds regardless of the training condition except for one: higher thresholds on a roving condition were obtained if the listeners were trained on that condition rather than on a fixed condition. For "poor" listeners, introducing variability in the stimulus during training (roved or widely-roved condition) slowed learning. This study highlighted that learning depends on the interaction between individual abilities and the variability of the training set.

The differences of performance between roved and fixed standard can be interpreted as a "long-term template" effect present in the case of fixed standard. The improvement in thresholds, with fixed standard, are attributed to an increase accuracy of the internal representation of the repeated tone. However, in [20], one of the experiments presented a fixed frequency tone, but at the second position of the pair, and a worsened performance was noticed compared to when the fixed frequency was presented at the first position. The authors suggested then that presenting a fixed tone was not sufficient to achieve the best discrimination performance. The order of presentation of the fixed tone has an impact on performance, but more generally, it can be suspected that the order of presentation of tones regardless of roved or fixed standard can affect performance.

Raviv et al in [24], investigated the effect of recent history on pitch perception. They took the example of a 2AFC procedure: in a 2AFC procedure, when the magnitudes of two stimuli are small relative to the distribution of stimuli used in an experiment, participants tend to respond that the first stimulus was larger whereas they tend to respond that the second stimulus was larger when the magnitudes of the two stimuli are relatively large. This effect is known as the "contraction bias". These types of bias, that are a consequence of a certain context preceding the judgment, are therefore expected in pitch discrimination experiments. In order to explain these effects, Raviv et al in [24], proposed a model in which the decision in a two-tone

discrimination is based on a comparison between the second tone and an exponentially-decaying average of the first tone and past tones: the contribution of tones preceding the judgment is then taken into account.

Other context effects have been observed. Marmel and Tillmann, in [13], showed that discrimination performance in the case of small mistunings was better when the to-be-compared tones were related to the melodic context in terms of tonality. The ambiguity of Shepard tones separated by a triton has been investigated by Chambers and Pressnitzer in [4] . In this study, the tone preceeding the triton interval had a significant impact on the perceived direction of the triton interval, even for short-duration context tones.

1.4 Measuring pitch discrimination

Different methods can be used to measure pitch discrimination performance by determining the FDL. The method of adjustments is one of them and is evoked in the definition of pitch given by Hartmann in [7] : it consists in adjusting the pitch of a comparison tone by matching it with the pitch of a standard. Other methods are more often used, and generally two main approaches are considered: the adaptive and non-adaptive ones.

Method of constant stimuli

The non-adaptive approach is often called the method of constant stimuli (MSC). To determine the subject's FDL, the method tests a range of pitch differences and repeats this measure a certain number of times. In this method, the distribution at various pitch differences is specified in advance. The method of constant stimuli allows the full psychometric function to be measured.
In the case of pitch discrimination experiments, as the variability of performance between subjects is large, the range of tests of pitch differences has to be large as well to cover this intersubject variability. This raises the question of the efficiency of this method. The fewer the trials required to estimate the subject's FDL, the better, but in the case of pitch discrimination tasks, the MCS might be time consuming due to the high number of trials needed. That is why adaptive procedures have been argued to be more efficient than non-adaptive ones. They permit the outcome of previous trials to be used to place future trials at efficient testing locations.

Adaptive methods

Difference limens can be estimated using adaptive procedures, in which the frequency difference between tones varies as a function of the responses of the listener. In adaptive forced-choice procedures, the listener is asked to indicate which of the multiple sequential presentations on each trial was a target stimulus. For example, in a same-different task, the listener hears a certain number of pairs of tones and must indicate in which pair the tones' pitch changed. 3 or 4 intervals can be presented to the subject in a 3-4 Alternative forced choice (AFC) procedure. In the case of our study, the subject is asked to compare pitch changes between two tones by indicating if the pitch went up or down, we are then in the case of a 2AFC procedure. The adaptive aspect of the procedure changes the step between the frequency tones as a function of the subject's answer: the step increases when the subject's answer is incorrect and is reduced when the subject answers correctly.

The differences between adaptive procedures depend on experimental variables which include the initial starting value of the stimulus, the amount of difference between stimulus values presented (the step size),

the process that guides the sequence of presentation levels on each trial (the tracking algorithm), and the decision for ending the process (the stopping rule), [11].

The staircase procedure, inspired by the early works of Bekesy, searches the subject's threshold through a combination of increasing and decreasing stimulus steps, responding to negative and positive subject responses. This method use generally the previous or more responses within the adaptive tracks to select the new trial placement.
Simple up–down staircases call for a reduction in frequency difference when the subject's response is correct and an increase in frequency difference when the response is incorrect. Both the correct and the incorrect response sequences consist of one trial, and the difference in frequency carried by the track moves after each response, targeting the 50% correct performance. In order to target a higher percentage of correct answers on the psychometric function, transformed up-down staircases are used. In these methods, the sequence for an upward movement stays at one incorrect response but the sequence for a downward movement may be two or more correct responses. A two-down one-up procedure targets the 70.7% correct performance on the psychometric function, and a three-down one-up, the 79.4%: the frequency difference does change after each incorrect answer but changes after only ,respectively, two or three correct answers (see Fig 1.6).

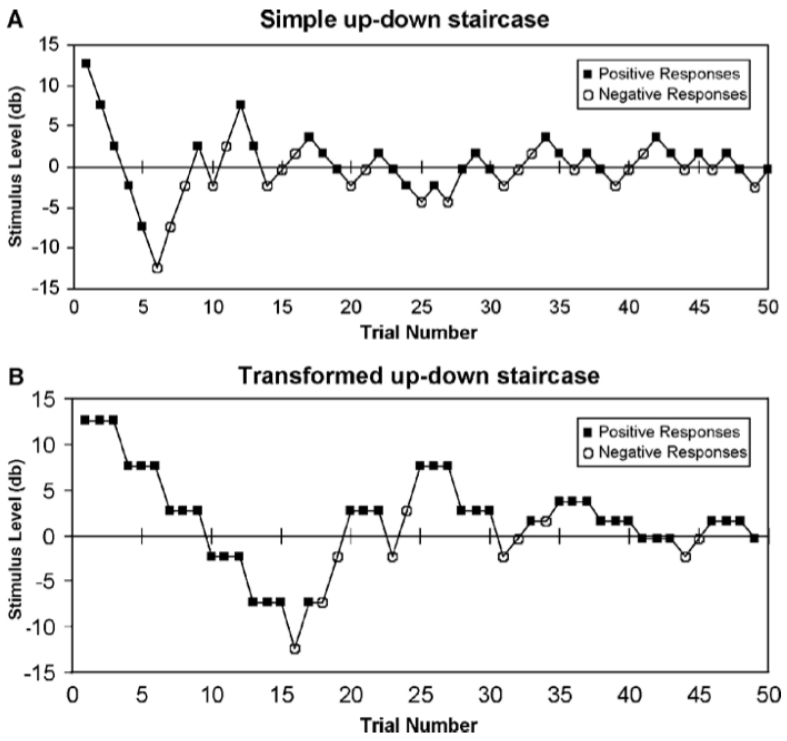

Figure 1.6: Example of a study of sound level sensitivity using adaptive tracks following a staircase procedure. (A) Simple up–down staircase; (B) transformed up–down staircase, following a three-down, one-up algorithm, from [11]

In our study, the procedure used is the one suggested by Kaernbach in [9], called simple up–down weighted procedure (in our case a simple one-up-three-down weighted procedure). In Kaernbach's simple up–down weighted procedure, the frequency difference is changed after every trial according to the ratio of up to down steps. For example, in order to target the 75% correct answers, the frequency difference should be changed upward after an incorrect response but 1/3 downward after a correct response.

In adaptive procedures the pattern of tone frequencies over time is highly constrained by the adaptive rule. This constraint is relaxed by the use of multiple parallel tracks ([12]). The more varied tonal context allowed by parallel tracks is useful for our study of context effects. Here we use a 4-track procedure.

1.5 This study

Our study has two main goals.
The first is to validate a new procedure in which the listener is presented with a sequence of tones and must make a judgment after each tone rather that after each tone pair as in classic procedures.
The second goal is to investigate context effects in terms of performance but also in terms of systematic bias (i.e. tendency to systematically respond in a particular direction) as suggested by [24]. Indeed, detrimental effects of context might be explained in terms of bias rather than degradation of discriminability itself. The investigation will be made to determine whether pitch judgments depend on a longer history of tones presented prior to each judgment, and if so to quantify this dependency.
This study consists of a series of psychophysical experiments, to address our two main goals, and one experiment involving both psychophysics and EEG.

CHAPTER 2

Methods

The aim of these experiments is: first, to validate a new method to measure pitch discrimination performance and compare the performance given by this task to classic tasks; second, to study the effects of context and memory on pitch discrimination performance. In order to achieve these goals, five tasks were designed and run in 3 sessions, two of which were also replicated using electroencephalography (EEG) in a fourth session.

2.1 Subjects

15 subjects took part in the experiments, 10 women and 5 men aged from 21 to 27 years old. 7 of them were musicians. One non-musician woman was excluded from the analysis because of her abnormally poor performance on the tasks. All the participants filled a questionnaire about their musical expertise in order to divide the subjects into two groups, a musician and a non musician group (See Appendix A). To be qualified as a musician, the subject had to have play an instrument for at least 5 years and still have a regular practice, they are or were part of a music school or conservatory.
All participants performed the 3 psychophysics sessions, they had no history of audiological or neurological disorders, and reported normal hearing which was confirmed by an audiogram done in the frequency band used in the experiments. The procedures were approved by the CERES ethics committee (IRB 20131100001072), the participants were informed of the procedures and a written consent was obtained from each of them.

2.2 Stimuli

In all the experiments, the sounds presented are pure tones 100 ms long with 10 ms cosine inramp and outramp. Each stimulus consisted of one tone in Task1 and Task3, two tones with 500 ms inter-onset interval in Tasks2, 4 and 5. Each tone or tone pair was presented 500 ms after the subject's response to the previous tone or tone pair. The frequencies were determined according to an adaptive rule (Sect. 2.4).

2.3 Tasks

In Task1 and Task3, the subject was presented with an ongoing sequence of tones, and had to respond after each tone whether it was higher vs lower in pitch than the previous one. The time between two tones

depended on the response time of the subjects, but they were asked to answer as fast as they could in order not to forget the pitch of the previous tone. In Task2, Task4 and Task5, the subject had to answer after each pair of tones.

In Task1, the frequency interval preceding each tone had a random sign and a magnitude determined according to an adaptive procedure (Sect 2.4) (Fig 2.1(a)). In Task3 the frequency interval preceding every other tone had a random sign and a magnitude fixed at 1 semitone (6%) (Fig 2.1(c)). In Task2 the frequency interval between tones of a pair had a random sign and a magnitude determined by the adaptive procedure (Sect 2.4). The frequency interval preceding the first tone of a pair and the second tone of the preceding pair was determined by the same adaptive procedure, see Fig 2.1(b). In Task5 the frequency of the first tone of a pair was equal to that of the second tone of the preceding pair, see Fig 2.1(e). In Task4 the frequency of the first tone was fixed at 1 kHz (fixed standard), see Fig 2.1(d).

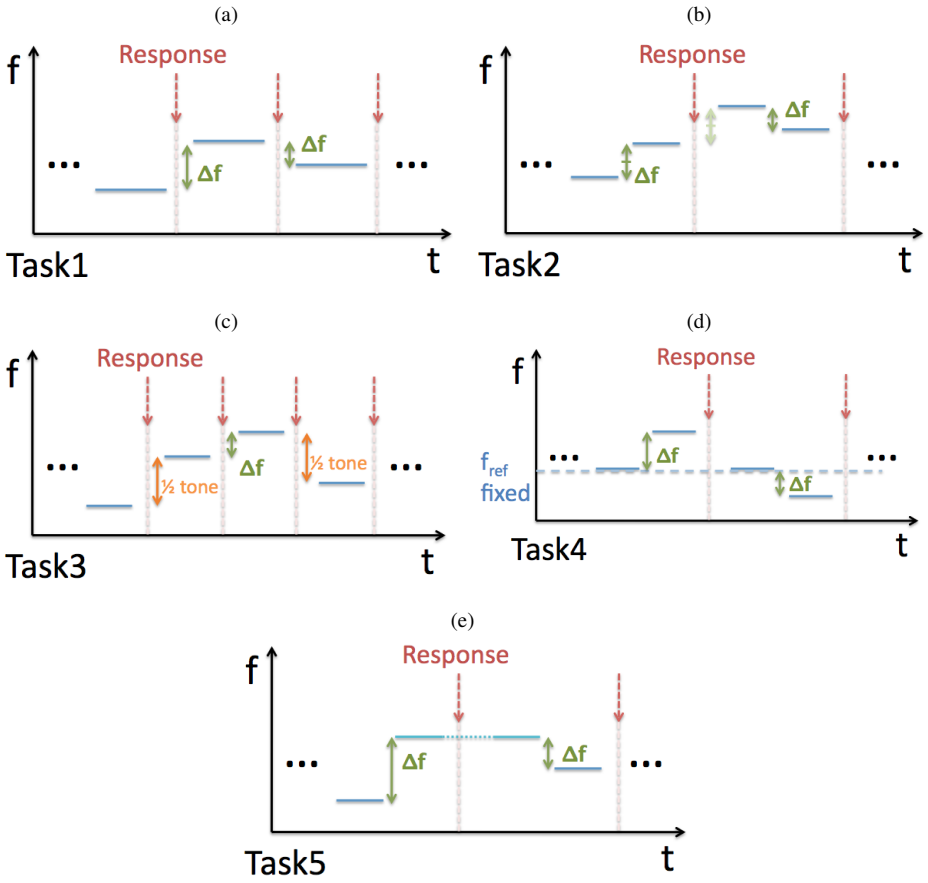

Figure 2.1: Summary of tasks. In Task1 and Task3 the subject responds after each tone. In Tasks 2, 4 and 5 the subject responds after each tone pair. The frequency interval between tones is determined adaptively (see text).

The performance of the tasks were organized in 3 sessions in each of which they performed 2 tasks (among the 5). Each session lasted approximately one hour, including instructions and training. Each session included 6 blocks, in each block the subject performed one task, the tasks were interleaved over blocks (see 2.2). In the first session, the subjects performed Task1 and Task2, in the second they performed Task4 and Task5, and in the third and last session, they performed Task1 and Task3 .

The sessions all started with a training block where the participants had to perform both task, each with half the number of trials as the main blocks. This was to ensure that subjects get used to the procedure and understand what feature of the sound they had to listen to. 120 trials per Task were presented, except for Task3, where 240 trials were presented.

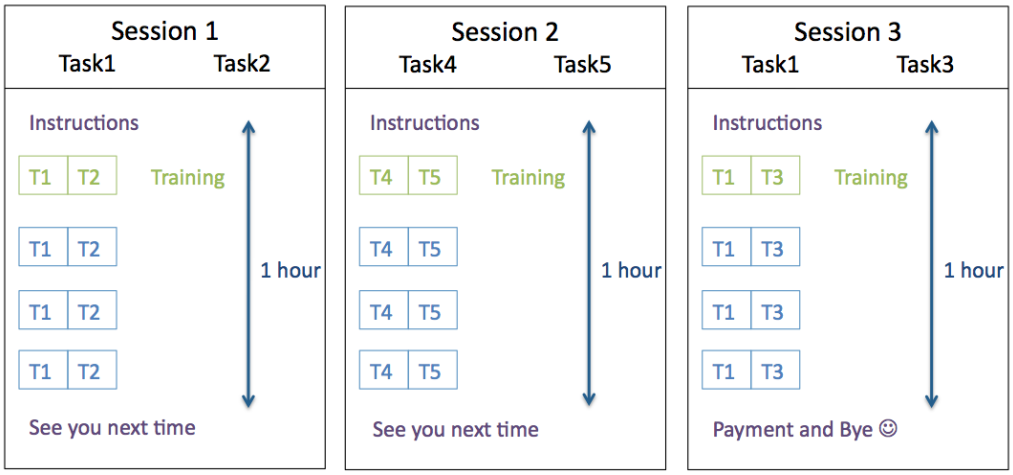

Figure 2.2: Organization of the sessions of experiments. Each session started with a training, two tasks were performed per session. All subjects had the same organization of sessions and tasks.

2.4 Adaptative procedure

All tasks used a 2-alternative forced choice weighted up-down procedure (from [9]). After every wrong answer, the ΔF becomes 2 times larger, and after every right answer the ΔF becomes $\frac{1}{2^{\frac{1}{3}}}$ times smaller: the procedure followed a weighted one-up three-down algorithm to track the 79.4 % correct point on the psychometric function.
Four independent tracks were interleaved [12]. The motivation for interleaved tracks was to reduce the serial correlation of interval sizes. This is important for the second goal of our study, which is to characterize serial context effects. Interleaved tracks also make it easier to characterize fluctuations of discrimination performance over time ([12]).
All experimental blocks started with a single track. After the first wrong answer (after at least 12 answers) the procedure switched to 4 independent interleaved tracks: the DF on each trial was adjusted based on the response to the DF presented 4 trials before it (Fig 2.3).

The frequency steps were randomly up or down with a bias to keep them close to 1 kHz. This bias was depending on the proportion $ratio = \frac{Freq - InitialFreq}{InitialFreq}$: if $ratio$ was bigger than a random number between 0 and 1, the frequency step turned up, in the other case, the frequency step turned down.

2.5 Apparatus

All stimuli were created using MATLAB software at a sampling rate of 44.1 kHz and and a dynamic range of 16 bit. They were played out via a Meridian Explorer2 sound card. Sounds were delivered diotically through Beyerdynamic DT 770 pro 250 ohm headphones, and presented at a listening level of 70 dB SPL.

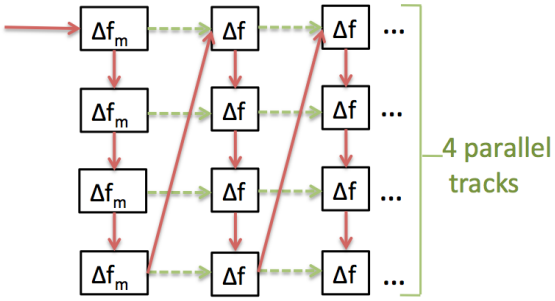

Figure 2.3: Adaptive procedure used in the experiments: Four parallel tracks are involved. The red path shows the order of presentation to the subject of the different frequency steps: the adaptation of ΔF, due to a wrong or right response every four trials, is represented by the green dotted arrows.
The procedure starts with one track and after the first mistake of the participant, represented by Δf_m, it switches to 4 tracks.

Listeners were tested individually in a double-walled sound attenuating booth placed in the basement of ENS (29 rue d'Ulm, Paris). Subjects used the keyboard to answer.

2.6 Analysis

2.6.1 Measures of performance

Stimulus parameters, subjects responses, and timing of stimulus and response were saved.

The data were analyzed in two ways.

In the first, the pattern of ΔFs was used to calculate a frequency discrimination threshold. The data obtained were fit to a psychometric function and the threshold was taken as the 79.4% correct response point. In the case of a 2 AFC method, the equation of the psychometric function is :

$$f(x) = 0.5 + 0.5 * \frac{1}{1 + \exp(-s * (x - x0))} \tag{2.1}$$

$x0$ is the threshold at 75% correct answer, s is the slope of the sigmoide at this point, the data x corresponds to the log2 of ΔF.

This analysis evaluates the subject's discrimination performance in terms of correct answers, as in classic pitch studies. The performance was measured in terms of FDL. Six blocks were performed per session, one FDL was determined per block, yielding 3 FDLs for each of the 2 tasks performed in that session.

In the second analysis, the probability of response on each trial (UP vs DOWN) was analyzed as a function of ΔF and of the preceding context (F, ΔF, previous responses, etc.) to reveal context effects. This analysis is of interest to observe possible systematic biases.

2.6.2 Statistics

Per session, three FDL per tasks were obtained, a geometrical mean was made over the three FDL. In order to compare the difference in performance between tasks, statistical tests such as ttests, repeated measure ANOVA, one-way ANOVA and two-ways ANOVA were made. Each test used to analyze the significance of an effect is specified in chapter 3.

2.7 EEG

The EEG recording was performed with the collaboration and help of Dorothée Arzounian.

2.7.1 Subjects

Three men and four women, aged from 22 to 24 years, participated to this experiment. Three of them already did the previous behavioral experiments, so they were trained for the tasks, the last two subjects were naive and never did the tasks before.

2.7.2 Tasks

The session was divided into 3 blocks, in which the subject performed Task1 (120 trials), Task2 (120 trials) and Task1 again (1080 trials). A training was made before the first two blocks, it was half long (60 trials) for the subjects who already knew the tasks and was the same duration (120 trials) in the case of the naive subjects. Block3 (Task1) was designed to be relatively long (9 times the other blocks) in order to reveal eventual order effects.

The EEG was recorded while participants were performing the tasks, but not during training.

2.7.3 Apparatus

The EEG was recorded during the entire session with a ActiveTwo BioSemi® system. 64 channels positioned according to the standard 10/20 system, plus 2 additional channels positioned on the mastoids (see 2.4) and 6 EOG channels (IO1, IO2, SO1, SO2, EO1, EO2) were sampled at a 2048-Hz sampling rate.

Figure 2.4: Biosemi system using 64 electrodes placed according to the modified 10-20 system, together with 8 additional electrodes placed on the mastoids and circumocular positions.

All stimuli were created using MATLAB software at a sampling rate of 44.1 kHz and 16 bit, they were played out via a RME fireface 800 sound card. Sounds were delivered diotically through ER-3A ETYMOTIC research Inc. 10 ohms insert earphones. Listeners were tested individually in a double-walled sound attenuating booth. Subjects used the keyboard to answer.

2.8 EEG analysis

Details of the analysis techniques and results are shown in the Appendix.

CHAPTER 3

Results

Fourteen paid subjects performed 3 sessions of 6 blocks each. In each session, two tasks were presented in alternation, 3 blocks for each task. In each block, 120 responses were recorded. Each session lasted approximately one hour.

3.1 Overall performance

The results of the three sessions are presented and compared to highlight different effects influencing pitch discrimination performance. The overall discrimination thresholds are shown in Fig 3.1. Each colored line corresponds to the performance of one subject and each point corresponds to the geometrical mean over three FDL. Each FDL was derived from 120 responses. The black curve shows the mean results over the subjects.

Thresholds values are overall consistent with the literature ([14]) Thresholds of good subjects (0.2 %) are similar to those reported in classic papers on pitch discrimination ([18]).
The spread across subjects is also consistent with the literature. However, the largest thresholds are still smaller than the average reported by [24].
A repeated measure ANOVA ($F(2.23,28.97)=6.757$, $p < 0.05$) reveals a significant effect of Task. Indeed, as seen in Fig 3.1, this effect seems to be quite quite large. Moreover, the traces in Fig 3.1 seem roughly parallel, and then the observed trends are consistent between subjects as reflected by the absence of a significant interaction (two way ANOVA, $F(2,78)=0.45$, $p>0.05$).

3.2 Musician vs non-musicians

Pitch discrimination is an ability used by musicians each time they play their instrument. As seen in Fig 3.2, the average thresholds across musician subjects are lower than the average thresholds across non-musician ones. A two-way ANOVA shows that the effect of task is significant ($F(5,72)=3.19$, $p<0.05$), and the effect of group is also significant ($F(1,72)=26.26$, $p<0.05$), but no interaction is present ($F(5,72)=1.19$, $p>0.05$) as confirmed by the fact that the two curves on Fig 3.2 seem quite parallel.
Musicians are more trained than non musicians and are more likely to show better pitch discrimination performances.

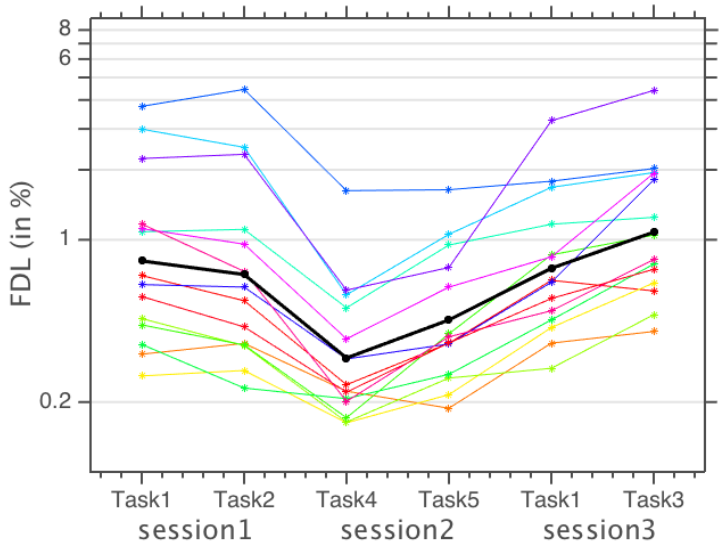

Figure 3.1: Frequency discrimination thresholds for each task in each session. Each colored line correspond to one subject, the black one represents the mean thresholds over the subjects.

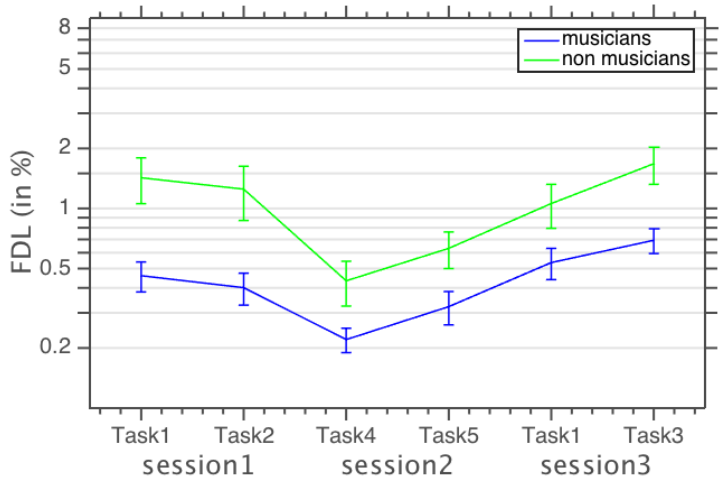

Figure 3.2: Frequency discrimination thresholds for musicians and non musicians over all the experiments. The effect of task is significant ($F(5,72)=3.19$, $p<0.05$) as well as the effect of group ($F(1,72)=26.26$, $p<0.05$). Error bars indicate the standard error of the mean over subjects.

3.3 Serial effects

Figure 3.3 shows the FDLs for each individual block of each task within each session. A two way repeated measure ANOVA shows that there is no significant effect of the repetition ($F(1.29,16.77)=3.9$, $p>0.05$), but the effect of task is confirmed ($F(2.19,28.44)=6.27$,$p<0.05$), there is no significant interaction between those two effects ($F(3.33,43.27)=1.262$,$p>0.05$). However, even if effects of training or fatigue don't pop out from the statistical tests, they cannot be totally excluded. Indeed, subjects can be affected differently by the opposites effects of fatigue and training, and then the inter subjects variability prevents from observing a general trend.

Task1 was presented in Session1 (3 blocks) and again in Session3 (3 blocks). One might have expected an improvement in performance due to training, but no significant difference is observed (Repeated measure

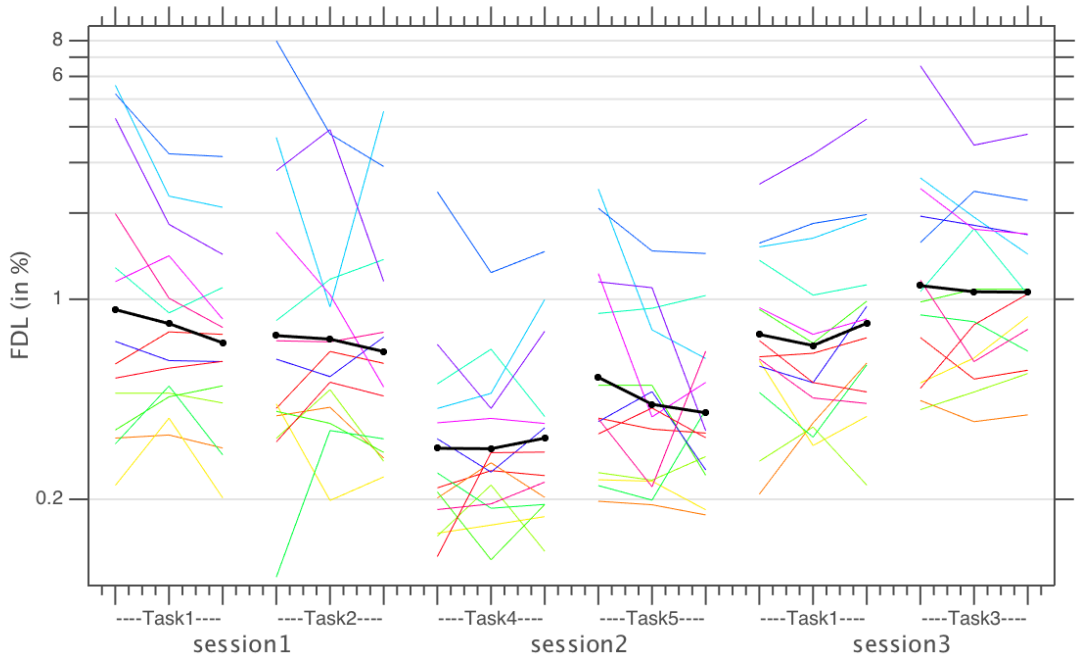

Figure 3.3: Thresholds over all 3 blocks of each task within each session: each task was repeated 3 times in each session. Each colored line correspond to one subject, the black one represents the mean thresholds over the subjects.

ANOVA P(1,13)=0.913,p>0.05) (Fig 3.4).

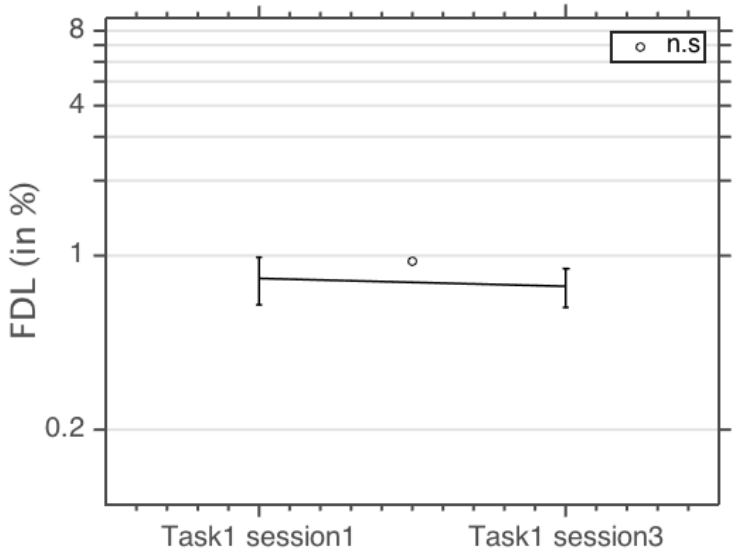

Figure 3.4: "Thresholds for Task1 in sessions 1 and 3. Error bars indicate the standard error of the mean over subjects.

3.4 New task vs classic task

During the first session, Task1 and Task2 were run . The thresholds obtained don't present significant differences (Repeated Measure ANOVA, F(1,13)=0.7079, p=0.4153 > 0.05).

The absence of a difference between Task1 and Task2 offers an answer to the first question addressed by this study. We see no systematic difference between the tasks, either in the mean FDLs (Fig 3.5)) or subject-specific results (Fig 3.3 , differences between tasks are of similar magnitude as differences between blocks of the same task), or the non-significant outcome of the statistical test. Lack of significance does not imply absence of any effect ([10]), as it can also merely reflect lack of statistical power. However, we note that other effects were significant. This gives weight to our conclusion that the new task is a good substitute for the old.

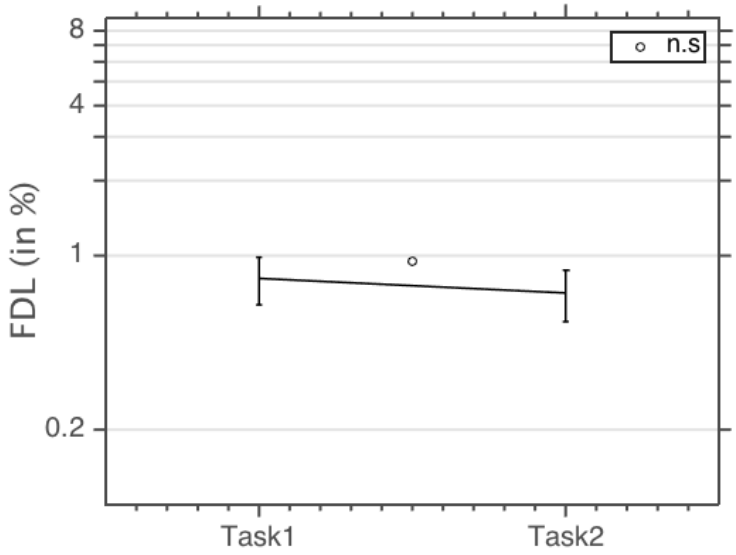

Figure 3.5: Comparison between the results of Task1 and Task2: mean thresholds over the subjects and their standard error associated

3.5 Context effects

A second aim of our study was to characterize the effect of prior context on pitch discrimination judgments.

Context effects emerge from three aspects of the data: (a) differences in FDL between tasks with different context patterns, (b) trial-by-trial performance (percent correct) as a function of prior stimulation, (c) trial-by-trial bias as a function of prior stimulation.

3.5.1 Anchor effects in the two-tone procedure.

Task2, Task4 and Task5 all used the same two-tone-per-trial procedure, but differed in how the first tone of each pair was chosen relative to previous tones. Figure 3.6 shows the FDLs for each Task. FDLs for Task4 (fixed standard) were significantly lower than for Task2 (Repeated Measure ANOVA, F(1,13)=9.301, p<0.05) or Task5 (Repeated measure ANOVA F(1,13)=15.571, p<0.05). FDLs for Task5 were lower than for Task2 (Repeated Measure ANOVA, F(1,13)=5.771, p<0.05). These results demonstrate an effect of prior context on discrimination performance.

The design of Task4 and Task5 suggests that the performance improvement in Task4 and Task5 compared to Task2 follows from "anchor effects". The first anchor effect is illustrated by Task4 and results from the fixation of the standard tone's frequency, the second is illustrated by Task5 and results from repeating the frequency between two consecutive pairs.

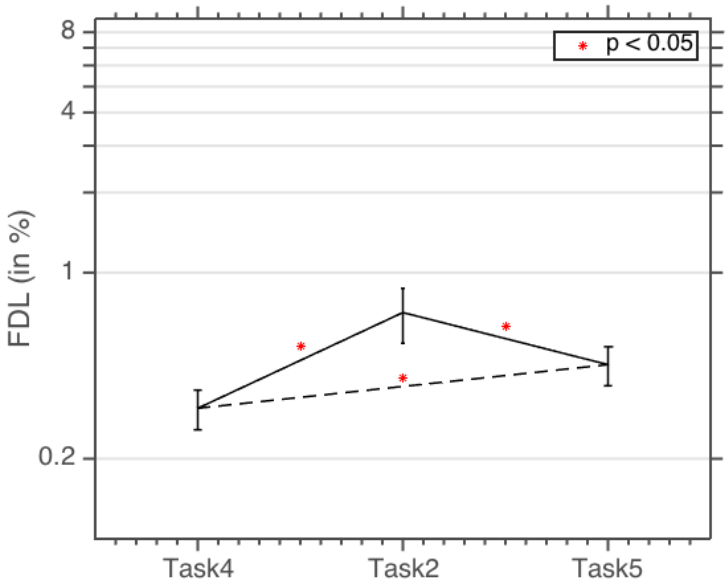

Figure 3.6: Comparison between Task2, Task4 and Task5: mean thresholds across subjects and their standard error associated

If fixing the standard frequency improves the performance, applying a large roving on the standard frequency should worsen it, the next subsection studies the effect of alternating large and small frequency steps.

3.5.2 Effect of a large DF interval in the single-tone procedure

Task1 and Task3 use the same kind of procedure, but in Task3 one over two intervals is a semitone. The results (Fig 3.7) show significant differences (repeated measure ANOVA $F(1,13)=12.085$, $p<0.05$): the thresholds resulting from Task3 are higher than the ones resulting from Task1. Then alternating small steps and large steps worsen the performance significantly, meaning that the subjects has difficulty to adapt to this alternation. There is a detrimental effect of preceding large frequency jump.

3.5.3 Trial-by-trial analysis in terms of percent correct.

In order to study if the pattern of prior tone frequencies biases the performance, the data were separated into two patterns (Fig 3.8). For the congruent pattern, the pitch goes in the same direction consecutively (up up or down down), for the opposed pattern, the pitch goes in different directions (up down or down up). This analysis studies the effect of the previous interval (irrelevant interval for response) on the present judgment (relevant interval for response), see Fig 3.8.

In this analysis, the previous interval is the pitch shift just preceding the judgment, the results of Task5 will be treated separately from the others as there was no pitch shift just preceding the judgment (the target

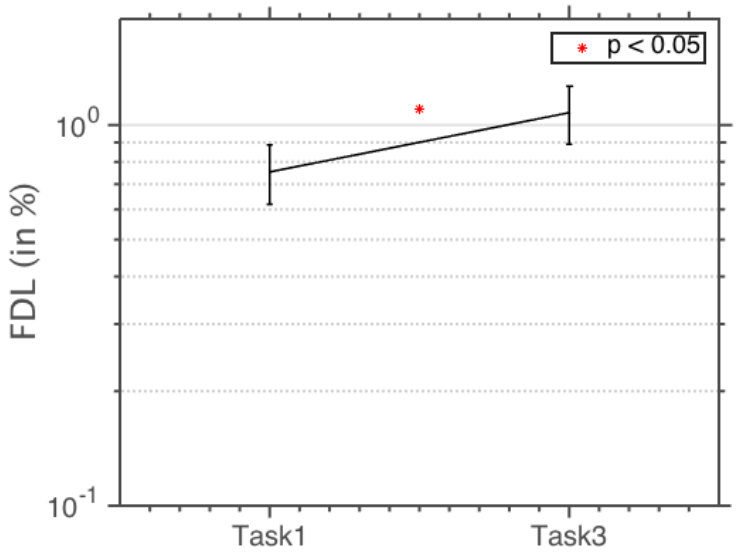

Figure 3.7: Comparison between the results of Task1 and Task3 during session3: mean thresholds and their standard error associated

of the previous pair is repeated).

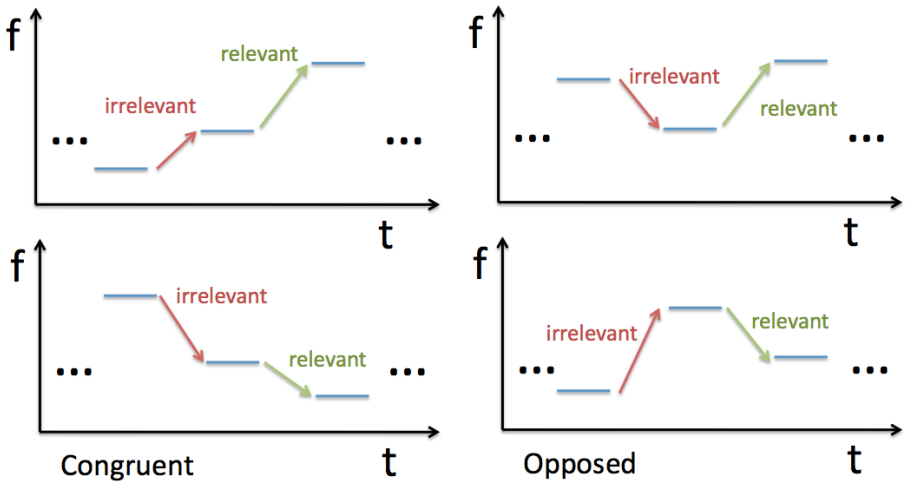

Figure 3.8: Separation of data into two direction patterns.
- For the congruent pattern, the pitch goes in the same direction,two times up, or two times down .
- For the opposed pattern, the pitch goes in different directions, up down, or down up.
The green arrow represents the interval in which the judgment is of interest, the red arrow represents the irrelevant interval which might influence this judgment

Most studies report context effects in terms of performance differences (as reflected by FDLs) between conditions, as we did in the previous paragraphs. In this paragraph, we extend the analysis to look at trial-by-trial effects.

To measure the thresholds corresponding to each pattern, the data were separated according to the congruent or opposed pattern. We divided the responses into two sets according to pattern, and fit each with a sigmoidal psychometric function. To limit the number of free parameters the sigmoids were constrained to have the same slope parameter, and differed thus only by their threshold.

An one-way repeated measure ANOVA was made on the thresholds obtained by the fit of two psychometric function on the two set of data corresponding to each pattern, it revealed that the listeners are better in the case of the congruent pattern.

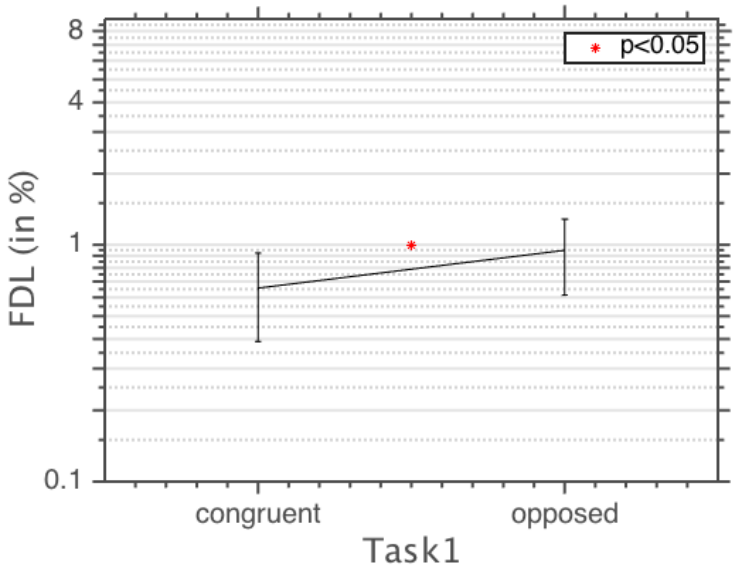

Figure 3.9: Comparison of thresholds obtained depending on the congruent or opposed patterns (Fig3.8). Those are the mean thresholds across subjects with standard errors.

3.5.4 Trial-by-trial analysis in term of response bias

Poor FDLs are usually thought to reflect imprecise sensory encoding, but they could also result from a context-dependent response bias.

The data were separated into two sets depending on whether the shift preceding the relevant judgment was up or down (2 conditions), see Fig 3.10.

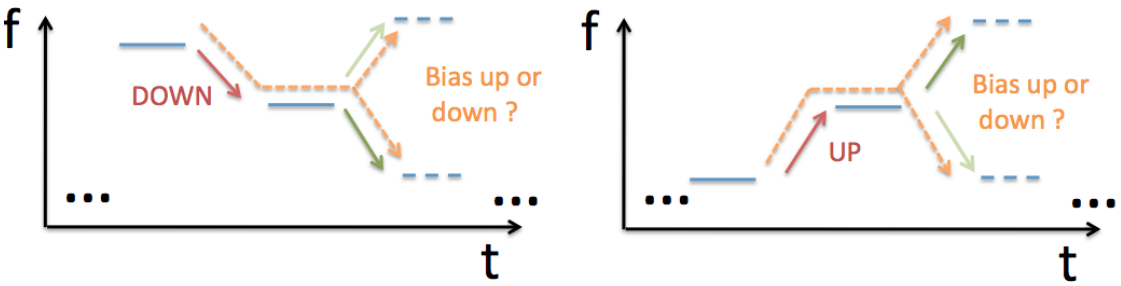

Figure 3.10: Instead of examining the opposed or congruent patterns, the preceding interval influence in term of bias is examined.

In the absence of bias, the response should depend only on the ΔF, and for $\Delta F = 0$ the listener should respond up or down with equal probability and the psychometric function should be centered. In the presence of bias the function should shift right or left depending on the sign of the bias, see Fig 3.11.

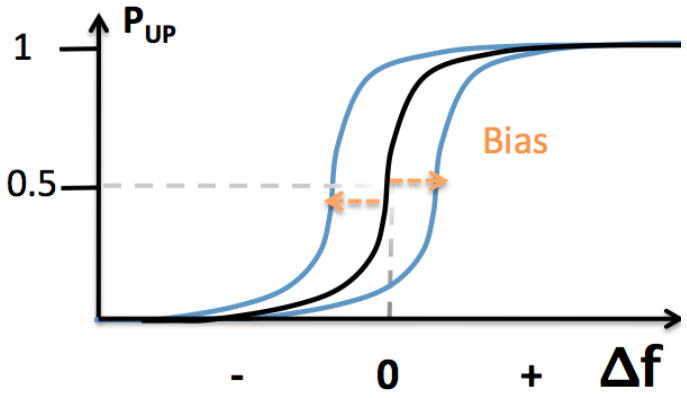

Figure 3.11: Psychometric function in term of probability of UP responses.

An analysis subject by subject was made in order to study whether the two sets of data are significantly different per subject. A bootstrap method (using 1000 iteration) was used in order to make this test. For this subject by subject analysis, the behavior results from the third block of the EEG measurement were also used. The result of this test are stored in the Table 3.1.
The bias presented in the table 3.1 are the difference between the 50% probability point of up answers with a previous interval going up and the 50% probability point of up answers with a previous interval going down. Graphically, the bias is positive when the psychometric function with previous up interval shifts right and the one with previous down interval shifts left, and negative in the other situation (fig 3.10).

Task1	Session 1-3				
	bias (in%)	p-value	thresholds (in%)		musicianship
	1.35	0.001	2.70	*	nm
	1.29	0.001	2.58	*	nm
	0.91	0.	0.97	*	nm
	0.63	0.036	2.24	*	nm
	0.27	0.	0.33	*	m
	0.27	0.	0.40	*	m
	0.23	0.049	0.76	*	nm
	0.11	0.035	0.35	*	m
	-0.17	0.	0.56	*	nm
	0.22	0.861	1.12		m
	0.13	0.096	0.60		m
	0.	0.484	0.32		m
	-0.12	0.891	0.65		nm
	-0.28	0.99	0.68		m
EEG DATA	Block 3				
	4.8	0.	3.7	*	nm
	0.56	0.	0.497	*	m
	0.27	0.	0.34	*	m
	0.16	0.006	0.44	*	m
	0.11	0.003	0.26	*	m
	0.069	0.097	0.32		nm
	0.058	0.164	0.41		m

Table 3.1: Table recapitulating the bias observed subject per subject. The star (*) indicates an alpha value of 5%

A big variability between subjects is observed: some subjects show significant differences and therefore are affected by the bias whereas other are less or not affected at all. The magnitude of the bias is also subject dependent. The presence of the bias does not seem to depend on musical expertise. However, it can be observed that the subjects who are more affected by the bias have high thresholds.

CHAPTER 4

Discussion

This study aims first at validating a new continuous procedure and second, at studying context effects that arise from previous tone history. Tasks using a classic procedure were compared to tasks using the new continuous procedure. Different parameters were manipulated to study context effects, and analysis in terms of performance and bias were made.
In a classic procedure, fixing the frequency of the standard tone (Task4) produces the best performances for all subjects. Relative to this, having the same frequency for the standard and for the target of the previous pair (Task5) worsens the performance, but a roving standard (Task2) is even more detrimental. Concerning the continuous task (Task1), the performances are comparable to the one given by the classic task with roving standard (Task2), but alternating small and large frequency steps (Task3) is detrimental for performance.

The continuous procedure has some advantages for studying context effects. The task is less time consuming than the classic procedure : only one sound is presented per judgment and the listeners have to answer as fast as they can, so in terms of timing the task is optimized. All the tones are followed by a judgment so this procedure is more adapted to study the effect of previous tone history. Moreover, as all the intervals between the tones are judged, this method is useful for EEG measurements. Preliminary analysis of EEG data recorded while subject performed classic and new tasks showed comparable responses (see Appendix B).

Context effects and bias suggest that some listeners integrate information from previous stimuli to form their estimate of standard tone pitch. Depending on how the stimulus frequencies are varied (roved or fixed standard), this is either beneficial or detrimental. Listeners vary in the extent to which they can minimize the bias from prior stimuli.

The gradual improvement of performances in Task2 Task5 and Task4 can be interpreted in terms of sensory trace mode and context coding mode ([5]).

In Task4, the standard was repeated on every trial so the listeners, thanks to the repetition, build an internal reference corresponding to the standard tone and compare it to the target tone. The repetition of the standard helps to have a more accurate internal representation in order to achieve the lowest thresholds. This internal representation thus built results from the use by the listener of a context coding mode, [14]. When operating with this mode, the listener performance is limited by the variability affected to the reference described as 'perceptual anchor' by Braida in [2].
In Task5, the target of one pair is the standard of the next, so the listeners use this repetition to start building an internal representation of the reference in order to succeed in the comparison task. However, the tone frequency on which the reference is built on is only repeated once and changes every trial, unlike in Task4

where the standard the reference on which is built on is fixed. The listeners still use the context coding mode, but, compared to Task4, the accuracy achieved to build the internal reference is reduced and this explains why the thresholds achieved are higher in Task5 than in Task4.
In the case of Task2, the listeners cannot use an internal reference as the standard is roved at every trial. The listeners have to use another strategy, not linked to context coding mode, but to sensory-trace mode. In the sensory-trace mode described in [5], a 'trace' linked to the observation is formed and diminishes over time. When operating in this mode, the sensitivity is limited by two types of noise: one linked to the observation itself (sensation noise) and another linked to the trace diminution over time due to decay it self or interference of other sources (memory noise) ([14]) . The worsened performance from Task5 to Task2 can be explained by the switch of mode from context to sensory-trace.
Same effects are observed in [20]: when the frequency of the standard was fixed, the thresholds were lower than when it was roved. However the thresholds obtained in that study are much higher than ours. In other words, our study provides a replication of that study with thresholds closer to those expected from previous literature on pitch discrimination.
In Task1, no tone is repeated, so, as in Task2, the listener cannot build an internal reference to compare to the other tones. This can also explain why Task1 and Task2 have comparable results (even if the absence of significant difference between the thresholds does not imply that no effect of task is present).

The frequency range used during a task is determined by the adaptive procedure, the 4 parallel tracks is used to obtain a wider range of context possibilities. With this procedure, the listener cannot anticipate the next stimulus as well as in the case of a 1 track adaptive procedure. However, in Task1 and Task2, even if the listener cannot predict the stimulus accurately, the frequency range used during the procedure is less and less wide (in function of the listener performance), so the listener can anticipate the presence of the next tone in a certain frequency range. This stimulus predictability might help the listener to do the task.
In the case of Task3, the alternation of small and large frequency steps might alter the performance due to a lack of possible stimulus expectation. The worsen performance in the case of Task3 compared to Task1 can be interpreted as a lack of predictability on the next stimulus because the frequency range used is much larger in Task3 than in Task1.

The trial-by-trial analysis reveals that the order of tones presentation can biases subjects answers. In terms of correct answers, there is an effect of tone pattern on the listeners responses. When the data are separated in congruent or opposed patterns, the thresholds corresponding to the congruent patterns are lower than the ones corresponding to the opposed pattern. This is confirmed by an analysis in terms of up answers: some subjects tend to answer up when the previous interval was also going up and tend to answer down in the other case. This effect highlights that the hypothesis that judgment are independent from pair to pair is wrong.
Those effects of systematic bias have already been observed in [24]: the 'contraction' bias is consistent with the pattern effects observed in our study. Indeed, in the case of the congruent pattern, when listener hear a tone going in one direction, the next tone will be compared to the mean of the previous tones with exponentially decaying weights, and then, the listener will tend to hear the tone going in the same direction. This has also been noticed in [25]: when the changes of the previous and present intervals are opposed in direction, the bias is stronger.

However, the bias does not affect all subjects. Moreover, the subjects affected by this bias cannot be separated into musician and non-musician groups. The musician criterion is not enough to explain this

variability of bias effects between subjects (and musicial training does not explain the absence of bias). The absence of bias effect can be interpreted as an ability to separate the judgments from one pair to the other and manage not to take into account the previous interval in the response.

The effect of the bias is not present in all subjects and this variability cannot be simply explained by a musical background. It would be interesting to investigate the criteria which determine the presence of bias in some subjects rather than in others.
Further investigations would be required to understand better the processes used in Task1 vs Task2, as no significant differences in performance have been found in our study. It would also be interesting to use Task1 design to study other auditory dimension (loudness for example), or even to generalize its use in other perception fields, such as vision.

Conclusions

Our findings suggest that in a pitch discrimination task, previous tones' history influences performance significantly.
The listeners reach the lowest thresholds when the frequency of the standard is fixed: this phenomenon can be described as an anchor effect which results from the building of an internal reference of the standard to be compared o the target. The repetition of the standard makes the internal reference more accurate and the comparison to the target tones is easier. When the frequency of the standard is the same as the target of the previous pair, the same process is used but less accurately (as the repetition occurs just once). When the frequency of the standard is roved, the performance of the listener is worsened.

Using the continuous procedure, an alternation of small and large frequency steps is significantly detrimental for performance: because in this condition the stimulus is less predictable, the listeners have more difficulty to anticipate the next pitch change, and then obtain lower threshold.
Focusing on the continuous task, a trial-by-trial analysis shows the presence of systematical bias: some subjects tend to answer that the pitch went up when the previous shift heard was up, and tend to answer that the pitch went down in the other case. However, this bias is not observed in all subjects. The subjects affected and non affected by the bias cannot simply be separated as a function of their musical expertise. However, it can be observed that larger bias values are obtained when subjects have larger thresholds.

The new continuous procedure used in this study obtains comparable thresholds as the classic task with roved standard. As it is more efficient, it may be a useful substitute for the classic task.

Appendices

Appendix A

This Appendix presents the first analysis of the EEG data. I only did the EEG measurements, this analysis was made by my supervisor Alain de Cheveigné and Dorothée Arzounian.

Analysis of EEG recordings.

The following is a preliminary analysis of the EEG data. It focuses on the stimulus-evoked responses, and the difference in these responses between Task1 (one tone per trial, response after each tone) and Task2 (two tones per trial, response after each pair).

Summary of recording methods.

EEG data were acquired concurrently with a psychophysical experiment that probed pitch discrimination performance. The recording session consisted of 3 blocks of approximately 5 minutes, 8 minutes and 25 minutes. Blocks 1 and 3 involved the same task (Task1), Block 2 involved a different task (Task2).
Subjects were 7 normal hearing individuals, (3 male), age range 22-24 years. Informed written consent was obtained under ethical approval IRB 20131100001072 (CERES).
Stimuli consisted of pure tones (100 ms, nominal frequency 1 kHz). In Blocks 1 and the tones were presented individually, in Block 2 the tones were presented by pairs, IOI=500ms. The frequencies of the tones were adjusted adaptively based on subject's responses, and the spacing between tones (Task1) or tone pairs (Task2) was self paced, each new tone or tone pair arriving 500 ms after the subject's response to the previous tone or tone pair. See methods for further details.
EEG were recorded with a Biosemi system using 64 electrodes placed according to the modified 10-20 system, together with 8 additional electrodes placed on the mastoids and circumocular positions (see Picture A.1). Sampling rate was 2048 Hz.

Analysis.

For each recording block, the data consisted of a 72-channel time series together with triggers indicating tone onsets (Blocks 1 and 3) or tone-pair onsets (Block 2). Together with each EEG file was a log file containing stimulus and response timestamps, and behavioral response data.
EEG signals were re-referenced by subtracting from each channel signal 0.99 times the average over channels. A factor of 0.99 was used rather than 1 to ensure that the data matrix retains full rank. Each channel was fit by a 10-th order polynomial, with a temporal weighting mask that was zero except during the 200ms

Figure A.1: Biosemi system using 64 electrodes placed according to the modified 10-20 system, together with 8 additional electrodes placed on the mastoids and circumocular positions.

preceding each stimulus (baseline). This polynomial was subtracted to remove slow trends. This particular weighting procedure helps ensure that the prestimulus portion of each trial has roughly the same amplitude. To further attenuate slow fluctuations, data were high-pass filtered (2nd-order Butterworth, 0.1 Hz cutoff). Prior to applying this filter the mean of the data over the first 10 s was subtracted to minimize the filter response to an initial step. Data were low-pass filtered by convolution with a square window of size 1/50Hz to fully suppress line-frequency interference including all harmonics. Data were further low-pass filtered by convolution with a 5-sample square window and downsampled to 409.6 Hz to save computation time. A square impulse response FIR filter minimizes smearing of temporal features (at the possible expense of spectral characteristics).
The data were then reorganized into a 3D matrix (time * channels * trials) based on the stimulus triggers, including a 200 ms pretrigger interval. Outlier trials were removed on the basis of their Euclidean distance (after removal of the DC component) from the average over trials for that block. Trials with distant greater than K times the standard deviation over all trials were rejected in two steps, with K=3 and K=2. Typically 5-15 % of trials were rejected for each block. After outlier rejection, the mean over each block (DC) was subtracted from each channel.
To suppress noise unrelated to stimulus evoked activity, the Denoising Source Separation (DSS) algorithm was applied to the data. DSS, which is related to the Common Spatial Pattern algorithm (Koles et al. 1990), finds the linear combinations of channels that maximize the ratio of repeatable power to total power (de Cheveigné and Simon 2008; de Cheveigné and Parra 2014). DSS involves the joint diagonalization of two covariance matrices: C0 covariance of the raw data, and C1 covariance of the data averaged over trials. Here, a solution common to all 3 blocks was obtained by calculating C1 as the sum over blocks of covariance matrices of data trial-averaged for each block.

Results.

Figure A.2 shows the average over trials of Block3, for one particular subject, of all channel signals after preprocessing (detrending, filtering, outlier removal) but before application of DSS. Values preceding the stimulus onset (t=0) are close to zero as a result of baseline correction. They then deviate from zero in two bursts, corresponding to two successive stimuli (the IOI, determined by the self-paced responses of the subject, was usually close to 1s). Closer examination of the trial-averaged waveforms suggests that they are strongly contaminated by alpha activity (circa 10 Hz).

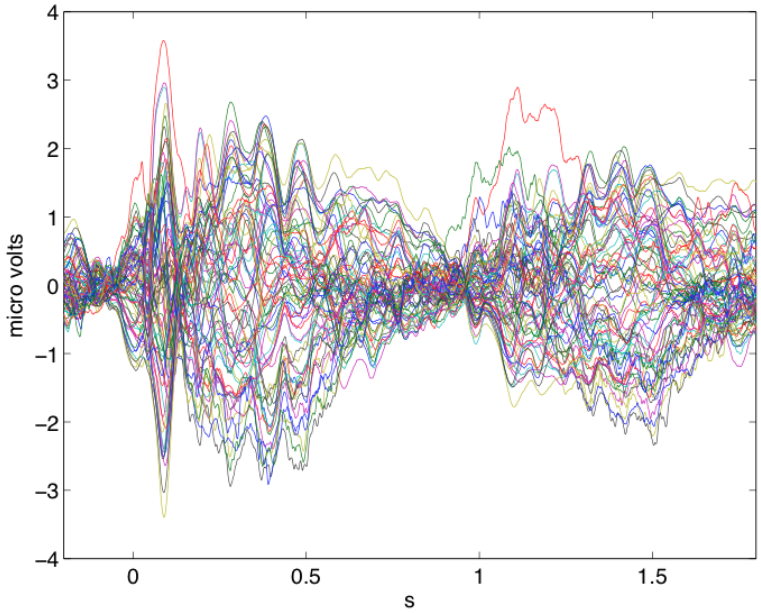

Figure A.2: Average over trials of the EEG after preprocessing (trend removal, highpass and lowpass filtering, outlier trial removal). Each trace is the signal from one electrode.

The data were baseline corrected by removing the mean over the 200 ms before the stimulus. To better characterize this responses, several different analyses were performed, each revealing a different aspect of the response.
(1) First, the DSS algorithm was applied to epochs of duration 2s starting 0.2 s before stimulus onset. Fig. A.3 (left) shows the average over trials of time course of the first DSS component for one subject, for each of the three blocks. The first DSS component represents the linear combination of channels with the greatest possible ratio of repeatable power to total power. For each block the component is highly reproducible (compare the excursion of the blue line to the width of the gray band). The value leaves the baseline level at about 150 ms post-onset and slowly returns to zero after 1s (Block1 and Block3) or 1.7 s (Block2).
For Block 1 and Block3 (single tone per trial, response after each tone) the signal returns to zero after about 1s, before deflecting once more in response to the next tone. For Block2 (two tones per trial, IOI=500 ms, response after second tone) the response extends over a longer interval and is of smaller amplitude. The topography (Fig. A.3 middle) is typical of late auditory responses (or of temporal response functions measured for speech). Similar responses were observed for all subjects (Fig. A.3, right).

(2) The previously found slow component was projected out of the data, and the data were again submitted to a DSS analysis, this time focusing on a 0.4 s epoch starting 0.2 s before the stimulus onset. Fig. A.4 (left) shows the average over trials of time course of the first DSS component for one subject, for the three blocks. This component is once again highly reproducible (compare excursion of blue line with thickness of gray band). For Block1 and Block3 the tone onset triggers a single complex deflection. For Block 2 the onset of the second tone triggers a second, smaller deflection approximately 0.5 s after the first. The associated topography (Fig. A.4 middle) is typical of early auditory responses.

In addition to this component, there are several others, also quite reproducible, with time-courses also locked to the onset of the stimulus (Fig. A.4 right). These multiple components span a subspace of stimulus-evoked activity, reflecting the activity of multiple sources with different time-course and spatial

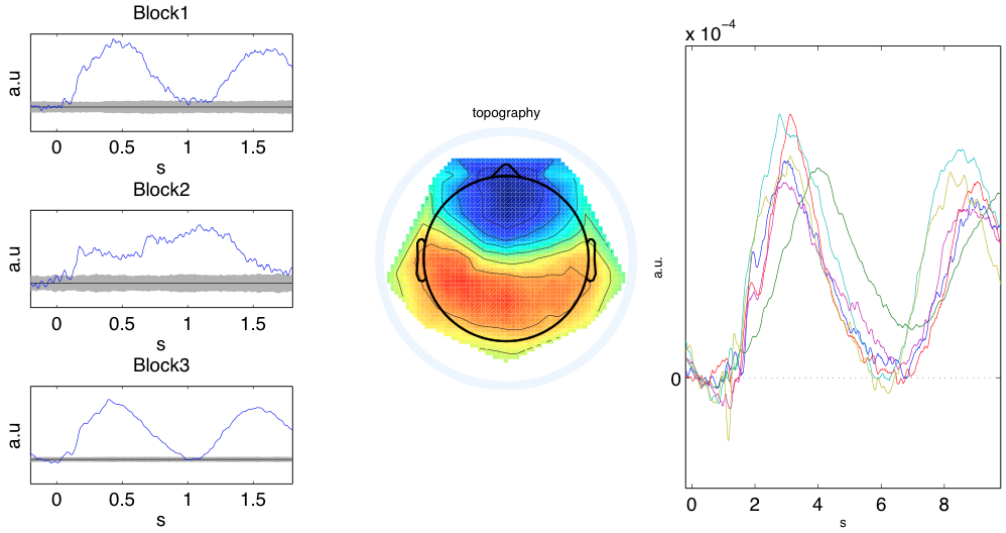

Figure A.3: Left: trial-averaged time course of the first component of a DSS analysis applied to epochs of duration 2 s starting 0.2 s before stimulus onset for one subject, for each block. The blue line represents the trial average, and the gray band represents +/- 2 SD of a bootstrap resampling of the mean. Middle: associated topography. Right: average over Block3 for 6 subjects.

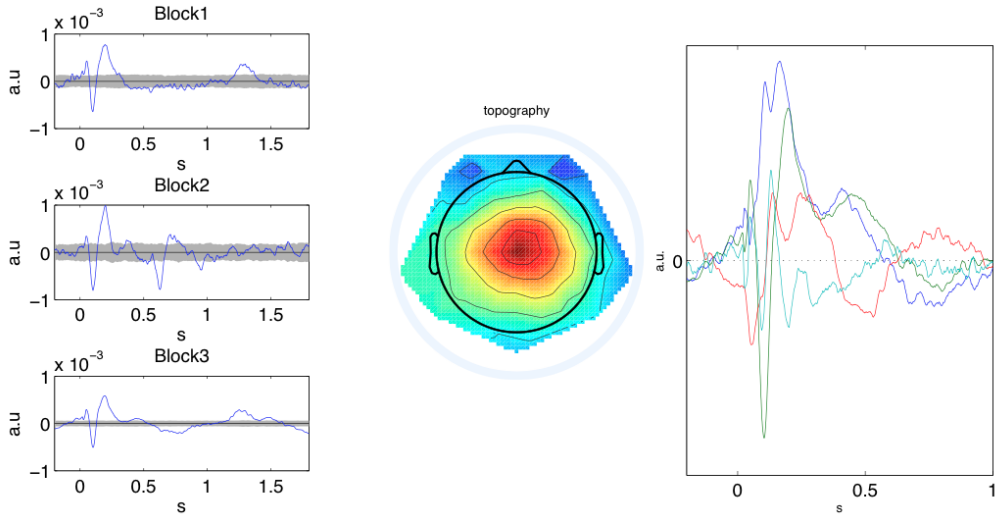

Figure A.4: Left: trial-averaged time course of the first component of a DSS analysis applied to epochs of duration 0.4 s starting 0.2 s before stimulus onset, for each block, for one subject (ADC-MLGDK-150619-pd01). Middle: corresponding topography. Right: time course of the first 4 DSS components for Block3.

characteristics. However there is not a one-to-one correspondance between sources and components.

Summary of the analysis:

A reproducible EEG stimulus-locked response is observed on each trial for both tasks. A component analysis technique (DSS) finds several components (linear combinations of electrode channels) that are highly reproducible. The first (most reproducible) component has a slowly varying time course that lasts the duration of the trial (approximately 1 s for Task1, approximately 1.7s for Task2). Subsequent components demonstrate a more phasic response. This response is temporally localized after the onset of the tone (Task1) or the onsets of both tones (Task2). The EEG response is simpler (more compact) for Task1 than Task2.

Additional analyses remain to be performed:

- Evoked-response analysis time-locked to the button press, to estimate the contribution of a motor contribution to the stimulus-evoked responses observed.

Appendix B

The subjects had to fill a questionnaire before the experiment in order to judge their musicianship. The questionnaire was in french.

âge:

Plus haut diplôme obtenu ? (entourez la mention concernée)

Baccalauréat

licence

master

doctorat

Avez-vous des problèmes d'audition ? (entourez la mention concernée)

Oui

Non

je ne sais pas

Etes-vous capable de dire d'un intervalle entre deux notes de musique consécutives s'il monte ou s'il descend (l'intervalle minimum étant le demi ton)? (entourez la mention concernée)

je ne peux pas

je peux difficilement

je peux mais pas toujours

je peux facilement

je peux sans aucun effort

Pratiquez-vous ou avez-vous pratiqué un ou des instument(s) de musique ? Si oui le ou lesquels ?

Combien d'années (tous instruments confondus) ?

Si vous avez arrêté de pratiquer, depuis combien de temps avez vous arrêté ?

Si vous avez toujours une pratique musicale régulière, combien de temps par semaine ?

Avez-vous reçu une formation musicale académique (école de musique, CNR, CNSM...) ?

Oui

Non •

Bibliography

[1] Sygal Amitay, David JC Hawkey, and David R Moore. Auditory frequency discrimination learning is affected by stimulus variability. *Perception & psychophysics*, 67(4):691–698, 2005.

[2] LD Braida, JS Lim, JE Berliner, NI Durlach, WM Rabinowitz, and SR Purks. Intensity perception. xiii. perceptual anchor model of context-coding. *The Journal of the Acoustical Society of America*, 76(3):722–731, 1984.

[3] Alan R Bull and Lola L Cuddy. Recognition memory for pitch of fixed and roving stimulus tones. *Perception & Psychophysics*, 11(1):105–109, 1972.

[4] Claire Chambers and Daniel Pressnitzer. Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, & Psychophysics*, 76(5):1271–1279, 2014.

[5] Nathaniel I Durlach and Louis D Braida. Intensity perception. i. preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America*, 46(2B):372–383, 1969.

[6] J Donald Harris. The decline of pitch discrimination with time. *Journal of Experimental Psychology*, 43(2):96, 1952.

[7] William M. Hartmann. *Signals, sound, and sensation*. Springer Science & Business Media, 1997.

[8] Adrianus JM Houtsma. *Pitch perception*. Academic Press San Diego, London, 1995.

[9] Christian Kaernbach. Simple adaptive testing with the weighted up-down method. *Attention, Perception, & Psychophysics*, 49(3):227–229, 1991.

[10] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

[11] Marjorie R Leek. Adaptive procedures in psychophysical research. *Perception & psychophysics*, 63(8):1279–1292, 2001.

[12] Marjorie R Leek, Thomas E Hanna, and Lynne Marshall. An interleaved tracking procedure to monitor unstable psychometric functions. *The Journal of the Acoustical Society of America*, 90(3):1385–1397, 1991.

[13] F Marmel, B Tillmann, and WJ Dowling. Tonal expectations influence pitch perception. *Perception & Psychophysics*, 70(5):841–852, 2008.

[14] Samuel Mathias. Individual differences in pitch perception. 2010.

[15] Samuel R Mathias, Peter J Bailey, Catherine Semal, and Laurent Demany. A note about insensitivity to pitch-change direction. *The Journal of the Acoustical Society of America*, 130(4):EL129–EL134, 2011.

[16] Christophe Micheyl, Karine Delhommeau, Xavier Perrot, and Andrew J Oxenham. Influence of musical and psychoacoustical training on pitch discrimination. *Hearing research*, 219(1):36–47, 2006.

[17] Christophe Micheyl, Li Xiao, and Andrew J Oxenham. Characterizing the dependence of pure-tone frequency difference limens on frequency, duration, and level. *Hearing research*, 292(1):1–13, 2012.

[18] BCJ Moore. Frequency difference limens for short-duration tones. *The Journal of the Acoustical Society of America*, 54(3):610–619, 1973.

[19] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[20] Mor Nahum, Luba Daikhin, Yedida Lubin, Yamit Cohen, and Merav Ahissar. From comparison to classification: A cortical tool for boosting perception. *The Journal of Neuroscience*, 30(3):1128–1136, 2010.

[21] Andrew J Oxenham. Pitch perception. *The Journal of Neuroscience*, 32(39):13335–13338, 2012.

[22] Christopher J Plack. *The sense of hearing*. Psychology Press, 2013.

[23] Christopher J Plack, Andrew J Oxenham, and Richard R Fay. *Pitch: neural coding and perception*, volume 24. Springer Science & Business Media, 2006.

[24] Ofri Raviv, Merav Ahissar, and Yonatan Loewenstein. How recent history affects perception: the normative approach and its heuristic approximation. 2012.

[25] Catherine Semal and Laurent Demany. Individual differences in the sensitivity to pitch direction. *The Journal of the Acoustical Society of America*, 120(6):3907–3915, 2006.

[26] Stanley S Stevens and John Volkmann. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, pages 329–353, 1940.