SEMANTIC CONGRUENCE and PERCEPTION

An experimental investigation.

Luciano Spilotros
Emmanuel Parzy
Medialogy, 10th semester
Spring 2010
Aalborg University Copenhagen
Supervisor: Luis E. Bruni

INTRODUCTION

Semantic congruence is a concept that has been in the last years more and more in the focus of perceptual studies. It starts to be considered as an important top-down factor of perception, emphasizing consequently the importance of memory and past experience in perceptual processes. A particular theory of perception from Hermann Von Helmholtz, the theory of *unconscious inference*, is also placing past experience at the heart of human perception.

In this study, it will be attempted to prove that the semantic congruence is compatible with the theory of *unconscious inference*, and can even be a central concept. It will be argued in this report that semantic congruence and the building of semantic links between stimuli is a main objective in perceptual processes. Semantic congruence will also be defined, and a review of the use of this concept across scientific fields will be done.

From an analysis of the previous perceptual researches on semantic congruence, it will also appear that research and experiments are still needed in order to clarify some effects of semantic congruence on perception. From an analysis of such previously conducted researches will emerge an idea of experiment, which will then be designed and implemented.

The aim of this study is also to attempt bringing our knowledge of media technologies to the area of cognitive science. This knowledge will permit to investigate on the best tools that can be used in order to observe perceptual and cognitive concepts, and maybe bring new ideas for further experiments.

PART I – ANALYSIS

CHAPTER 1 - BACKGROUND RESEARCH

The study of perception has a long history, starting from the theories of antic Greek philosophers to modern studies based on results gathered by means of brain imagery.

Many theories have been presented in attempts to explain the mechanisms that permit us to build a picture of the outside world. The conflicts between perception theories gave rise to questions that are still not completely answered: *do our senses give us an exact picture of what is surrounding us, or do we build such a picture from incomplete sense information in a sort of guess work? Is perception involving processes impermeable from other cognitive processes, such as emotional ones, or are they influenced by it?*

Recently, a growing amount of researchers are acknowledging a certain importance of top-down factors on cognition and perception. Among these factors, semantic congruence and its effects on perception are more and more in the focus of perceptual researches.

1.1 - PREVIOUS USES OF SEMANTIC CONGRUENCE

The terms of "semantic congruence", "semantic congruency" or "semantic congruity" have their roots in linguistics. In 1964, Pollio used the term semantic congruity as the proximity of words *"in semantic space"*. This semantic space was fixed by a set of bipolar scales to rate evaluation, potency and activity of words (Osgood and Tannenbaum, 1955). Later, the term was introduce to cognitive science to define an effect previously known as *"cross-over effect"* (Audley and Wallis, 1964), and known even earlier in studies on the effects of "*affective value-distance"* over reaction times in judgment tasks (Shipley et al., 1945; Dashiell, 1937).

The theory of the *"semantic congruity effect"* proposes that all stimuli are coded qualitatively over different dimensions (size, brightness, loudness, etc.). When a comparison is done between two stimuli over one of such qualitative dimensions, the reaction times from subjects are faster if the judgment being asked *("which one is brighter"* versus *"which one is darker"*) is *"congruent"* with the quality of the stimuli (relatively bright or dark). If the direction of judgment is incongruent with the quality of the stimuli, then the subject´s perceptual system has to first do a translation between the two dimensions (e.g. dark and bright), and that is the reason why reaction times appear to be longer.

Later, social psychology took the term *"principle of semantic congruity"* to define its own concept. According to Burke and Franzoi (1988), it refers to the principle that (environmental) "*settings must be recognized and labeled as salient features of social interaction. As such, they must be associated with affective meaning like social identities and behaviors"*. Further, they state that *"the relation between identities and situations, like the relation between identities and behaviors, exists through the common semantic dimensions and the human need for semantic consistency. We refer to this as the principle of semantic congruity"*.

Behind the socio-psychological definition of semantic congruity is thus the idea that the identity subjects are taking is tending to be congruent (in harmony, in agreement) with their surrounding environment, and vice-versa: this theory also considers that we modify our environment according to our current identity (see the definition of *identity* according to Burke and Franzoi (1988) for more details).

These concepts using the name of semantic congruity should not be confused with another use of the terms semantic congruence, congruency or congruity, the one of concern in this study. This specific concept has been lately more and more in the focus of perceptual and cognitive researches. To define this concept of semantic congruence (this term will here be used instead of congruity or congruency), it is necessary to look at both definitions of semantics and congruence.

1.2 - DEFINITION OF SEMANTIC CONGRUENCE FOR PERCEPTION STUDIES

Semantics is, along with pragmatics, the science of meanings. The term was defined by Bréal (1897), who introduced it as a name for his field of research, a subfield of linguistics, the science of languages. The term congruence, together with its derivations congruency and congruity, points out the notions of agreement, harmony, equivalence, or correspondence; it implies thus a partial relation of similarity, as opposed to total similarity or equivalence. Two congruent entities would thus have some similarities on some specific points, but would not be equivalent.

It must be underlined that congruence can also refer to concepts in mathematics, geometry, programming and neuro-linguistic programming; in all those specific frameworks, congruence seems nevertheless to always refer to this notion of harmony and correspondence at the origin of the term.

Semantic congruence could consequently be defined as a notion of agreement, harmony, equivalence or correspondence between the meanings of several

components. It appears that two types of semantic congruence can be differentiated: referring to equivalence and correspondence implies a categorization, a direct comparison between similar things. On the other hand, referring to the ideas of harmony and agreement tends to imply integration, unity and coherence between several “components”. It is this notion that is usually implied when one talk about semantic congruence in perception and cognitive science, and it will be consequently the one to which the term semantic congruence will refer in this study.

1.3 - SEMANTIC CONGRUENCE AS A MAIN FACTOR IN PERCEPTION

Defining semantic congruence as the harmony or agreement between the meanings of several stimuli permits to place it as a main factor of perception and multimodal integration.

Research on multimodal integration deals with the analysis of the factors responsible of the binding versus segregation of unimodal stimuli (Spence, 2007). This field of perceptual research has permitted the finding of a certain amount of factors that permit to modulate the perception of several stimuli as a single event or as several independent events. These factors are commonly divided in two types: *structural factors* that are physical characteristics of the stimuli (such as temporal or spatial correspondence between modalities), and *cognitive factors* that are involving processes from the subjects' memory, emotions, experience, etc (Spence, 2007; Suied et al., 2008; Welch and Warren, 1980; Koppen et al., 2008; Bertelson et al., 2000).

Among the structural factors are generally listed the spatiotemporal correspondence between stimuli (the relative position in time and in space of the stimuli), any temporal correlation between stimuli, and the relative salience of concurrent stimuli. The variation of such factors permit to some extent to modulate the perception of multimodal events as bound or independent. Nevertheless, it has been observed that in many cases, the perceptual system is able to ignore discrepancies between stimuli (up to a certain limit). Effects such as ventriloquism or visual capture are examples of such phenomena.

The study of multimodal integration could reach new perspectives if put under the light of a theory of perception inherited from Hermann Von Helmholtz (1821-1894). His theory of perception was driven by the assumption that the perceptual system uses past experience in order to infer the objects constituting a surrounding environment. The inference of the elements of the surrounding environment from “raw” sense information is processed, according to Helmholtz, in a problem-solving

kind of way. This theory, named theory of *unconscious inference*, consequently places past experience as the necessary resource for perception as we experience it. Helmholtz's theory was later revived by Adelbert Ames Jr. (1880-1955), who emphasized the fact that sense data is ambiguous and imprecise, that elements as basic for perception as shape size and distance are not directly accessible from the sense data but are better inferred from our experience, and perception can be defined as a kind of “guess”. As a proof, he led the experiment of the trapezoidal room, where perception of shape, size and distance is bias by the vision of a room with unconventional dimensions (Figure 1).

Figure 1: An example of a trapezoidal room experiment

The concept of semantic congruence can be added in this theory as the goal of the perceptual system in its real-time problem solving task. It will be here argued that the reason to be of this perceptual problem-solving is to reach a picture of the environment that “makes sense”, in respect with the subject's past experience. By “a picture that makes sense” is meant a picture where semantic congruence is achieved, where the meanings of all stimuli are in harmony which each others.

It has been proven, for example in the case of the ventriloquism effect, that if this harmony is not existing physically (i.e. a spatial discrepancy between auditory and visual stimuli), the perceptual system can distort the perception of those stimuli in order to reach a picture semantically congruent, and congruent (in harmony, in agreement) with the past experience of the subject (at this point, it can also be argued that ”semantically congruent” is nearly synonymous with ”congruent with one's past experience”).

This theory also allows considering a question that is still debated in perceptual research: *to what extent is perception learned, and are there innate perceptual processes?* Ames' trapezoidal room experiment proved that the perception of size, shape and distance, which can be considered as "low-level" perceptual processes, can be biased by past experience. The existence of purely innate perception is consequently put in doubt, or at least is proven to be possibly biased by higher knowledge and experience.

The phenomenon in the focus of this study, namely semantic congruence, is, as previously stated, a notion acquired through experience. Consequently, situations were semantic congruence influences specific percepts are proofs that these percepts are, at least partially, dependent on past experience.

To sum up, this study will consider perception as a real-time and continuous problem-solving task, where the sensory inputs (stimuli) have to be organized in picture where semantic congruence is achieved.

An example of study on semantic congruence in perception could then be based on challenging the perceptual problem-solving process. This could be done by modulating the semantic congruence/incongruence of stimuli, and analyze what is the resulting perceptual image (what does the subject perceive of the scene).

1.4 - PREVIOUS RESEARCHES ON SEMANTIC CONGRUENCE IN PERCEPTION

The psycholinguistic experiments of semantic congruence have evolved through time, following the technological advancements, and have been studied under the light of both behavioral and brain imaging paradigms. The trend for brain imaging studies is to try to define neural correlates to cognitive phenomena; this topic is not of interest in the present review and will be skipped (see Doehrmann and Naumer, 2008 for a review on this subject).

In this part, some of the experimental setups used in research will be presented to acknowledge the different trends and methods in this domain.

In 2003, Laurienti et al., in the field of cognitive-neuroscience, were the first to study semantic congruence out of the frame of language experiments, with multimodal cues (pictures of objects and animals, and related sounds). Their brain-imaging experiments showed that activity in certain brain parts is modulated by the semantic congruence (or *"contextual congruence"*, as they also call it) among crossmodal stimuli. A parenthesis must be made here to underline the fact that even

if Laurienti et al. (2003) are the first entry studying multimodal *"semantic congruence"*, stated as such, similar concepts have long before been studied under different names in multimodal perceptual research.

In 1953 Jackson (1953) studied the *"ventriloquism effect"* with both relatively abstract cues (lights and ringing bells) and cues based on everyday objects (kettle blowing steam and whistling sounds) that are supposedly carrying more semantics (see Figure 2). His observation was that the ventriloquism effect (the tendency of subjects to regroup auditory and visual cues into a single event even though they come from different spatial positions) was stronger with the semantically charged cues than with the abstract cues. In this experiment, Jackson did not manipulate semantic congruence by using cues going from semantically incongruent to congruent combinations (the spatial position of cues was actually the variable submitted to manipulation), it is more correct to say that the semantic "charge", or relative abstractness, was the semantic variable. Nevertheless, this experiment was a proof of the importance of semantic variables in multimodal integration, and more specifically in the intensity of the ventriloquism effect.

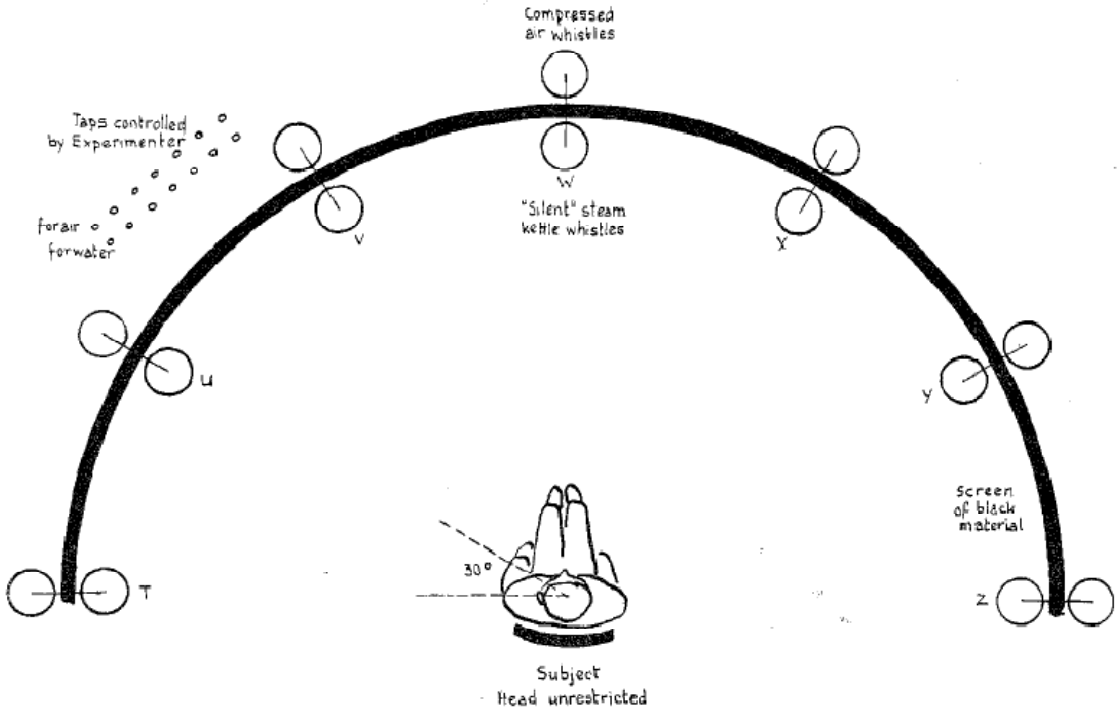


Figure 2: Jackson's (1953) experimental setup: The subject had to point to the direction from which was perceived the audio event; the angular deviation between the subject's choice and the real position of the audio event was measured

In 1980, in the frame of behavioral studies on perception, Welch and Warren (1980) listed the concept of *"unity assumption"* among the factors influencing crossmodal

bias. This unity assumption was defined as the assumption from a subject that several stimuli are corresponding to a single event. This concept, still present in some multimodal perception researches, is supposedly depending on factors such as the redundancy of physical properties (shape, size, motion, etc.) through the stimuli, the saliency of the stimuli, the subject´s experience with similar events, phenomenal causality, perceptual grouping, situational factors, or instructions such as the knowledge of the presence of a perceptive discrepancy (Welch and Warren, 1980; Warren et al., 1981; Vatakis and Spence, 2007; Vroomen, 1999).

The concept of *"unity assumption"* is obviously very close to the concept of semantic congruence. Rhetorically, the proof can be made by saying that the assumption that several percepts are linked into a single event (i.e. the unity assumption) is depending on the semantic harmony or correspondence of those percepts (the semantic congruence). Further, the semantic correspondence of those percepts is certainly depending on factors such as for example the redundancy of physical properties (shape, size, motion, etc.) through the stimuli, the saliency of the stimuli, the subject´s experience with similar events, phenomenal causality, perceptual grouping, situational factors, or instructions such as the knowledge of the presence of a perceptive discrepancy. All these are factors of the unity assumption, as they are listed in the literature on multimodal perception. It appears thus that even though it was not cited under this name, the effects of semantic congruence on multimodal interactions were already taken into consideration long before the use of the term by Laurienti et al. (2003) in his multimodal neuroscientific experiment.

More behavioral studies on the semantic congruence between multimodal non-linguistic cues have been conducted since. The studies from Murray et al. (2004) and Lehmann and Murray (2005) showed that memorization was more effective for semantically congruent audiovisual stimuli (line drawing pictures and related sounds) than with unimodal stimuli or audiovisual semantically incongruent stimuli.

Gallace and Spence (2006) studied how what we could call symbolic semantic congruence could influence the perception of relatively abstract cues. In their experiment, subjects had to evaluate the sizes of different disks, presented after that high or low frequency tones were played (the subjects were asked to ignore those tones) - Figure 3. A significant influence of the auditory tones over the reaction times was observed, a result that shows effects of semantic congruence even with relatively abstract cues.

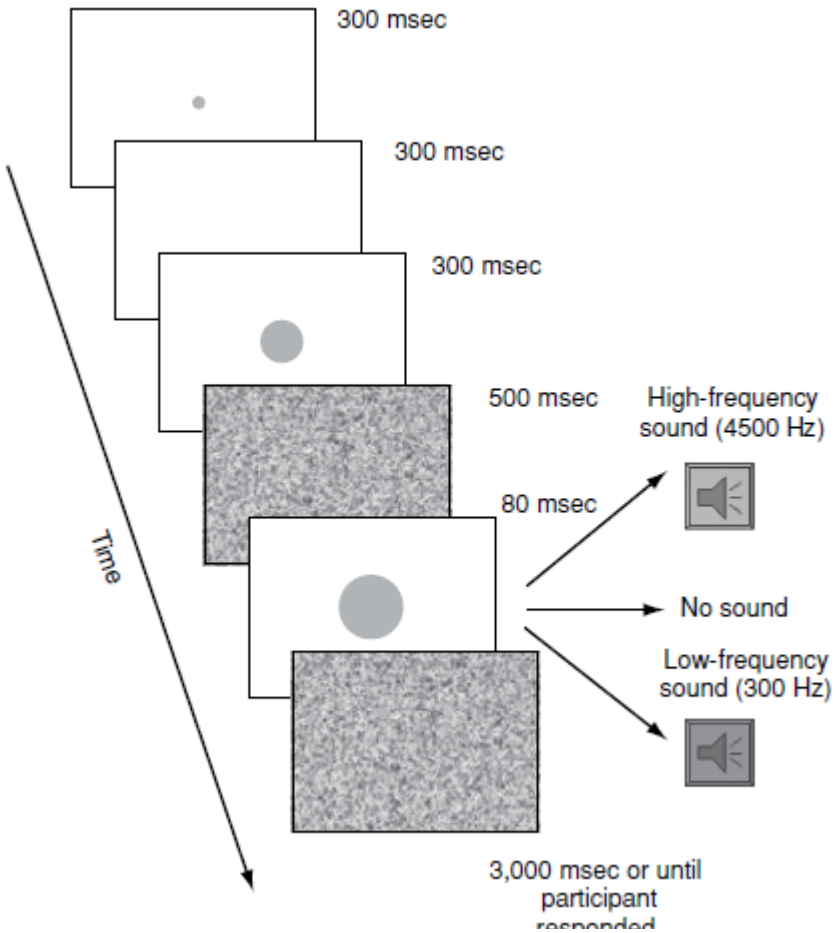


Figure 3: Gallace and Spence's (2006) experimental design: Subjects had to press one of two pedals according to whether they perceived the last circle as bigger as or smaller than the previous one; the reaction times were measured

Vatakis and Spence (2007) demonstrated that the gender congruence between faces and voices influences the integration between those visual and auditory cues (they describe their study as an experiment on the unity assumption) – see Figure 4. It is nevertheless not clear if the experiment reflects the effect of gender congruence or a priming effect with the subjects initially binding a specific voice with a specific face, as in the experimental process they were first shown the original faces and voices played together.

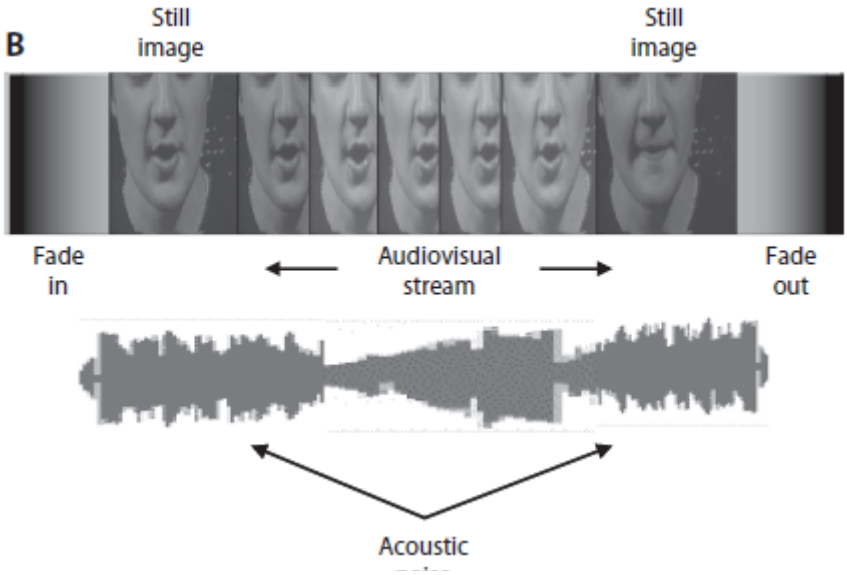


Figure 4: Vatakis and Spence's (2007) experimental design: The acoustic stream was variably shifted from the visual stream; the audio and visual could be gender congruent or incongruent; the measurement was based on the perception of a shift between audio and visual channels from the subjects, by asking them which between the sound or the visual was played first

Smith et al. (2007) also lead an experiment concerning gender perception: they observed that the gender perception of an androgynous face was significantly modulated by the presence of pure tones with frequencies in the range of either male or female speaking pitches (Figure 5). This result leads to the assumption that the perceptual system was seeking semantic congruence between auditory and visual cues.

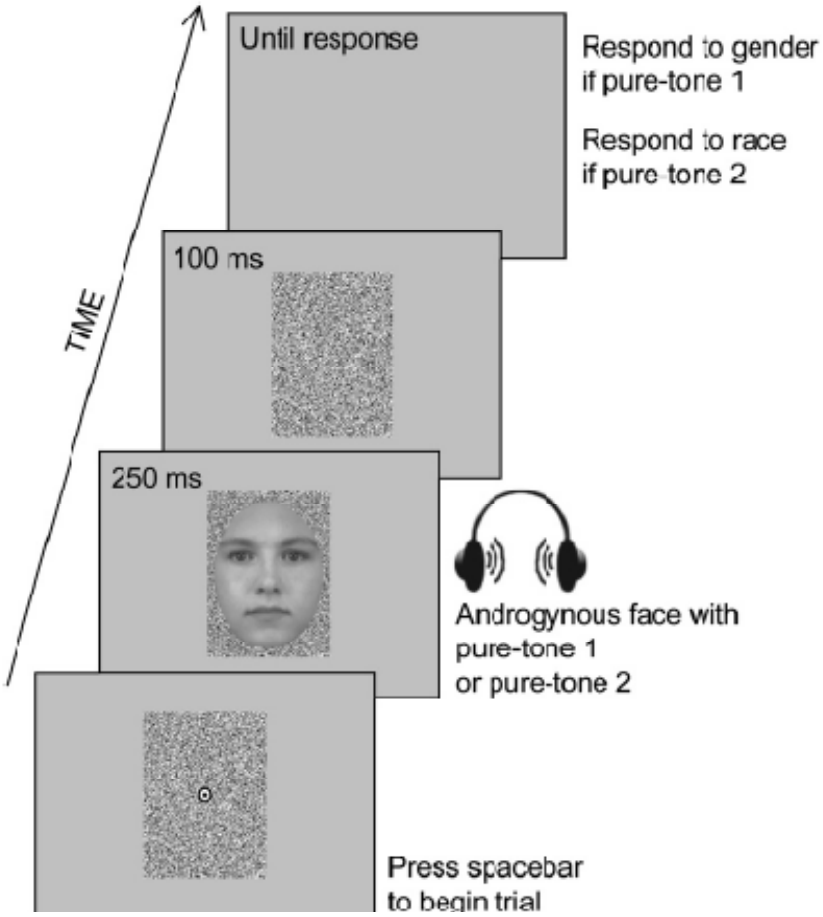


Figure 5: Smith et al.'s (2007) experimental procedure: Subjects were presented the face with a low or high frequency tone; depending on the tone, the subjects had to identify either the gender or the race of the face; the "gender" decisions according to the tone frequency was the variable of importance

Koppen et al. (2008) studied the effect of semantic congruence on the Colavita visual dominance effect, with cues as animal pictures and sounds. The *"Colavita effect"* demonstrated that subjects have a tendency to ignore auditory cues and rely more on vision when they are asked to do a speeded recognition of cues' modalities. They did not find any influence of semantic congruence on this effect, even though some other aspects of subjects' performance were modulated according to the congruence or incongruence between the auditory and visual cues.

Suied et al. (2008), in a study on the integration of auditory and visual information in the recognition of realistic objects (Figure 6), observed that in an object recognition task, reaction times are shorter for semantically congruent audiovisual stimuli than for semantically incongruent ones, and that spatial alignment of the stimuli did not have any significant effect. They also observed that when the stimulus was the unimodal auditory target stimulus (only the sound of the object to which the

subjects had to react to), the reaction times were not different than with the target auditory stimulus coupled with an incongruent visual stimulus; on the other hand, the fact of having a target visual stimulus coupled with an incongruent auditory stimulus increased significantly the reaction times compared to a situation with a unimodal visual target stimulus. The experimental setup used by Suied et al. (2008) was based on an immersive audiovisual virtual environment, this study having been also an argument for the efficiency of such display technologies for the observation of some perceptual effects.

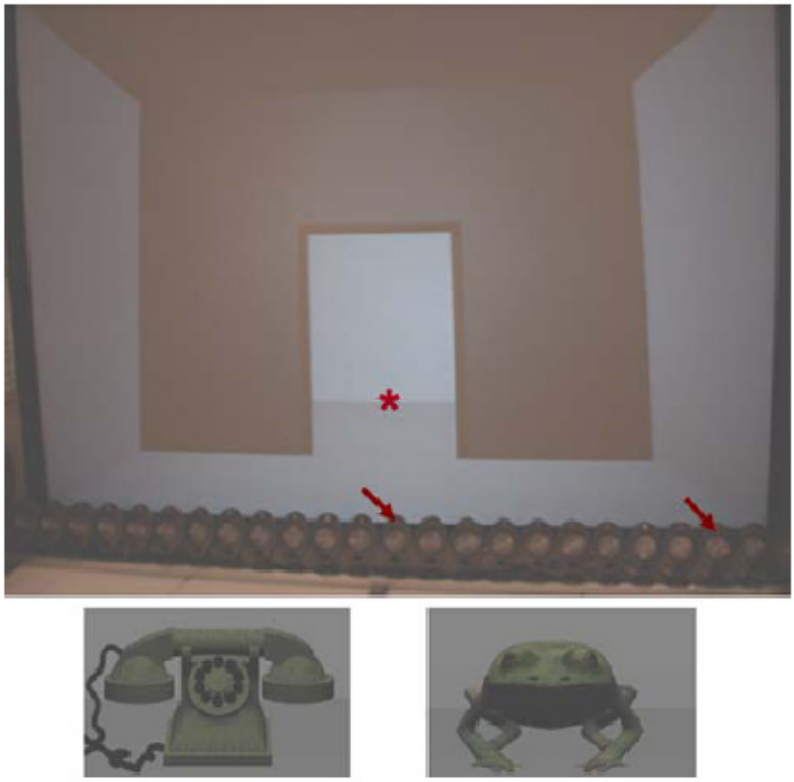


Figure 6: Suied et al.'s (2008) experimental design using an immersive audiovisual virtual environment: the red arrows correspond to the placement of the speakers, and the frog and phone models were the objects used for the recognition task. The subjects had to press a key if the target (frog or phone) was appearing either visually or auditorily; the reaction times were measured.

Yuval-Greenberg and Deouell (2008) conducted an object recognition experiment, based on audiovisual representations of animals as shown in Figure 7. Several variables were set: the auditory and visual cues could be congruent or incongruent and before each trial, it was asked to the subject to focus on one of the two modalities ("*What do you hear?"* versus "*What do you see?"*). The contrast of the visual cue was varied, going from a clearly visible picture to a slightly less visible one, and a delay between the stimuli and the subjects' response could be introduced. As predicted, the subjects were faster with congruent stimuli than incongruent ones.

They were also faster when they were asked to focus on the visual stimuli, even when a delay was introduced between the stimuli' presentation and the subjects' answer. However, this visual dominance disappeared when the low-contrast pictures were used, a proof that modality dominance is also depending on the quality of the stimuli.

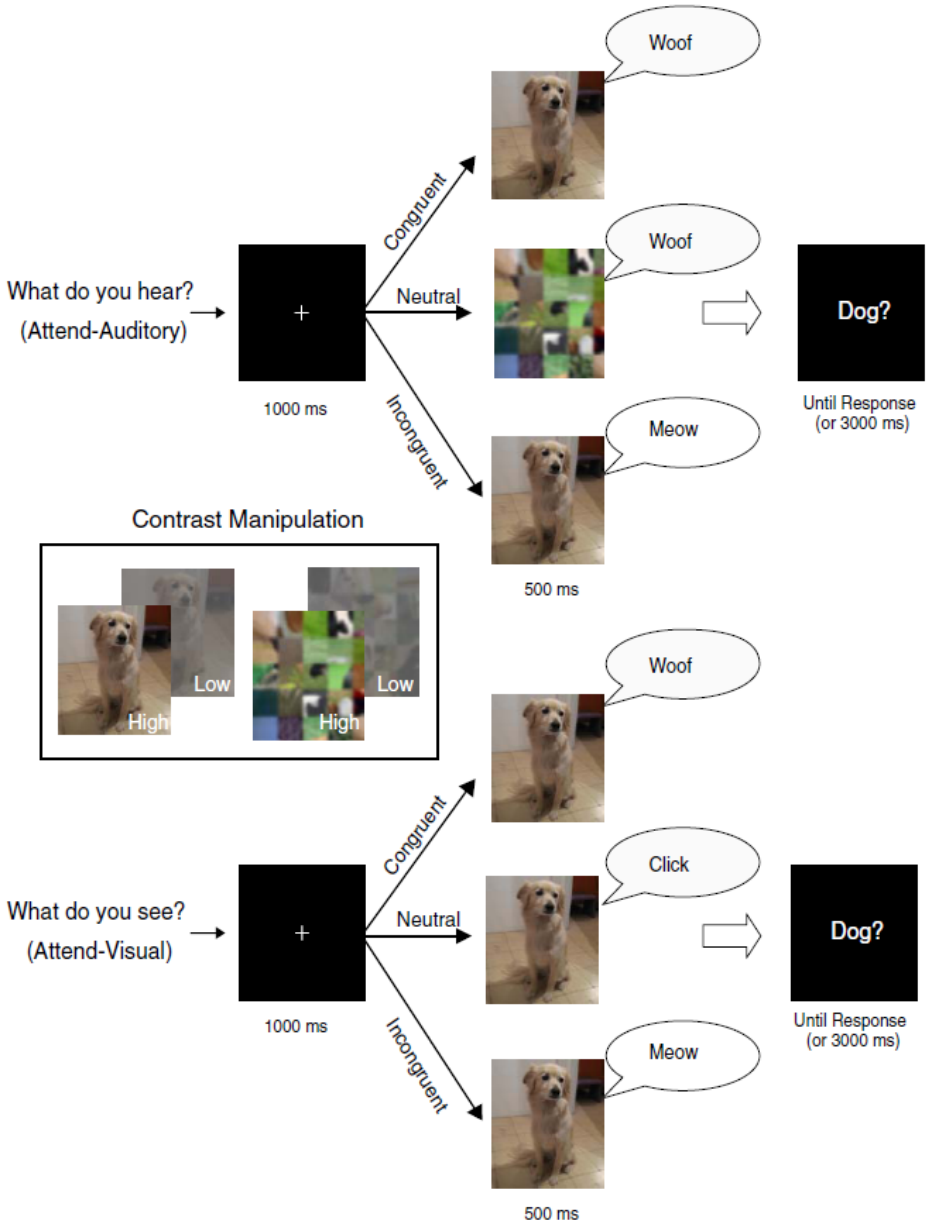


Figure 7: Yuval-Greenberg and Deouell's (2008) experimental procedure: In the "what do you see" procedure, the neutral condition involves a "meaningless" mosaic as a visual cue; in the "what do you hear" procedure, it involves an abstract click as auditory cue. The subjects were exposed to all possible combinations between audiovisual streams, contrast and time delay; the reaction times of right answers were measured

1.5 - SYNTHESIS

These different studies on semantic congruence are giving insights on different observable effects on perception tasks. Suied et al. (2008), Gallace and Spence (2006) and Yuval-Greenberg and Denouell (2008) showed that semantic congruence permits faster performances in object recognition. These observations are portraying how situations of semantic incongruence are complicating the perceptual "problem-solving" task, leading consequently to longer reaction times when incongruent stimuli have to be recognized.

Smith et al. (2007) results are giving clues concerning how the perceptual system tends to look for semantic congruence, even between stimuli that have relatively abstract relations (is this case, human androgynous faces and pure tones in the female or male frequency ranges). Vatakis and Spence's results (2007), on another side, proved that semantic congruence is effectively a factor of multimodal integration.

If the influence of semantic congruence and incongruence on subjects' reaction times in perceptual tasks has been relatively much documented, it is interesting to note that Jackson (1953) is the only entry studying effects of semantic congruence on a localization task. Though, as told previously, this work does not observe effects due to a variation from semantic congruence to semantic incongruence, but analyzes the amplitude of the ventriloquism effect with stimuli going from relatively abstract (light bulbs and bell sounds) to everyday-life objects and sounds (steaming kettle and whistle). A deeper analysis of the effects of semantic congruence and incongruence on the spatial localization of events would consequently be an interesting track for further knowledge on this topic: it is thus this track that will be followed in this experimental research.

The following parts of this report will be dedicated to the design and implementation of an experiment that will permit to analyze the influence of semantic congruence on the perception of the location of audiovisual events. In this experiment, the semantic congruence between visual stimuli and auditory stimuli coming from the same position will be varied. The assumption leading this experiment is that semantic congruence can be an important factor in the localization of auditory sources. Consequently, a first hypothesis is that subjects would be more precise in localizing audiovisual stimuli that have congruent semantics, instead of abstract stimuli that present not or little semantic relations between each others. Another hypothesis is that in the case where the auditory and visual stimuli, originating from the same position, are incongruent, the subjects will bias the perception of the

auditory source location and “move it” to a close visual object congruent with the sound.

Hypothesis 1: *Subjects would be more precise in localizing audio stimuli which is semantically congruent with the visual object located at the sound source, than localizing abstract audio stimuli that presents none or little semantic relations with the object located at the sound source*

Hypothesis 2: *In the case where the auditory and visual stimuli, originating from the same position, are incongruent, the subjects will bias the perception of the auditory source location and “move it” to a close visual object congruent with the sound*

If these assumptions prove being right, it would be an argument supporting Helmholtz's theory of perception based on unconscious inference, where past experience is a leading factor of perception.

PART II – EXPERIMENTAL DESIGN

CHAPTER 2 - EXPERIMENTAL DESIGN

The idea at the source of this experimental design is to place subjects in front of several visual objects distributed in space. The subjects will then be exposed to different sounds, played from the positions of the different objects (with the help of spatial audio techniques), in a way that these sounds will be perceived as coming from these objects.

The setup will consequently permit the experimenters to play different types of sounds at each object's positions: in this way the semantic congruence between the audio and the object at its source can be variated.

In order to allow the analysis of the hypothetical effects of semantic congruence and semantic incongruence on auditory localization, the experiment must involve three possible situations: a first one with abstract auditory and abstract visual stimuli, from which the results obtained will be used as a reference for the other configurations. Another one will involve semantically (non-abstract[1]) congruent stimuli, and the last one semantically (non-abstract) incongruent stimuli.

In each condition, the subject will be facing eight visual objects. In the abstract condition, these objects will be three-dimensional geometrical shapes (cubes and pyramids) with different colors, to ease the identification of these objects (Figure 8).

1 In this study, the adjective non-abstract will be used as synonym of concrete in order to define qualities of a stimulus, which are related to a specific object or everyday-life instances. Subjects, when exposed to such non-abstract stimuli, would supposedly call to their past experience with similar stimuli.

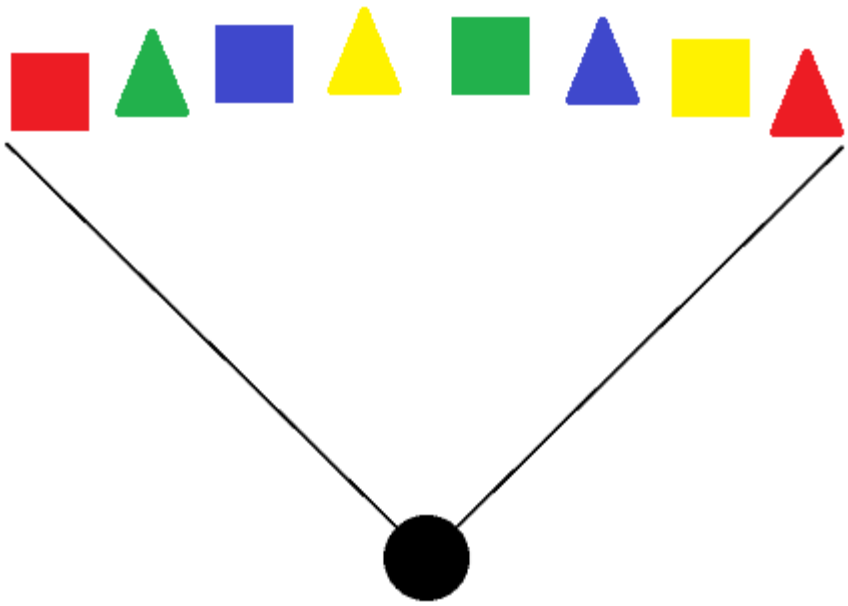


Figure 8: Abstract condition with colour cubes and pyramids as visual stimuli

In the semantically congruent and incongruent conditions, the objects will be four identical phones and four identical radios. A color sticker will be placed near each of them, again to ease the identification of the different objects.

During the experiment and with the help of three-dimensional audio techniques, sounds will be played from the positions where these eight objects are. For the abstract condition, these sounds will be also abstract (i.e. free from any semantic values or references), created will different synthesis techniques (granular synthesis for example). In the two other configurations, recorded phone sounds and fragments of radio broadcasts will be used. All the sounds have a length between five and six seconds.

2.1 - SPATIAL AUDIO TECHNIQUES

The experimental setup implies a spatial audio system that must make it possible to control eight virtual sound sources. Several spatial audio techniques can be listed in order to achieve such a setup.

A first possibility could be the use of binaural audio techniques. Binaural sound consists in audio with filtering corresponding to the natural filtering of shoulders, pinna and skull that occur in a natural listening condition. This filtering is variable depending on the position of the sound source (azimuth, elevation and distance from the listener). Consequently, audio binaurally processed is made to be listened through headphones, to avoid such a filtering to occur twice. Binaural sound can be

implemented through two different processes: it is possible to record sounds binaurally with the help of a mannequin with microphone placed in the ears. The materials used to manufacture such mannequin must reproduce the filtering characteristics of the corresponding parts of the human body.

Figure 9: Example of mannequin for binaural recording

It is also possible to compute binaural audio algorithmically. The above-described mannequins can be used to obtain a Head Related Impulse Response (or HRIR), representing the frequency filtering caused by the mannequin's body before the audio reaches the microphones, depending on the position of the sound source. This impulse response can be used to model a Head Related Transfer Function (HRTF), which can then be implemented in a filtering algorithm to process mono sounds. This last technique has the advantages of being more flexible than the method based on recording sessions with a mannequin (the algorithm can be used infinitely and spatialize any monophonic sound file).

A binaural algorithm from Brown and Duda (1997) was implemented and tested for the sake of this experimental design. It appeared unfortunately that the algorithm did not allow enough precision in the localization of audio sources. Consequently, such solution was dropped (see appendix 2 for more details on the implementation and testing of this solution).

Other spatial audio techniques are involving the use of speakers, such as the Ambisonic technique for example; nevertheless, it must be underlined that as only eight specific sound sources are needed in the experimental setup to be developed, it is possible to use an eight-channel audio system, with a speaker assigned to each channel and placed where each source is supposed to be. Such a system was implemented, and tested on five subjects who had to locate sounds played randomly at each of the eight source positions. These sources were signaled by numbers, to

ease their identification. Each subject executed sixteen trials, two per possible sources.

The result was considered satisfactory (*M* = 14, *2SD*=3.03) and it can be concluded that this setup configuration as being precise enough for the purpose of this experiment.

2.2 - PLACEMENT OF THE OBJECTS IN THE NON-ABSTRACT CONFIGURATIONS

In the two test configurations involving concrete stimuli (phones and radios), different placements of these objects are possible. Different possibilities will be here described, with assumptions on the perceptual effects that can possibly emerge from them. A first possibility could be to have a regular alternation of phones and radio as shown in Figure 10.

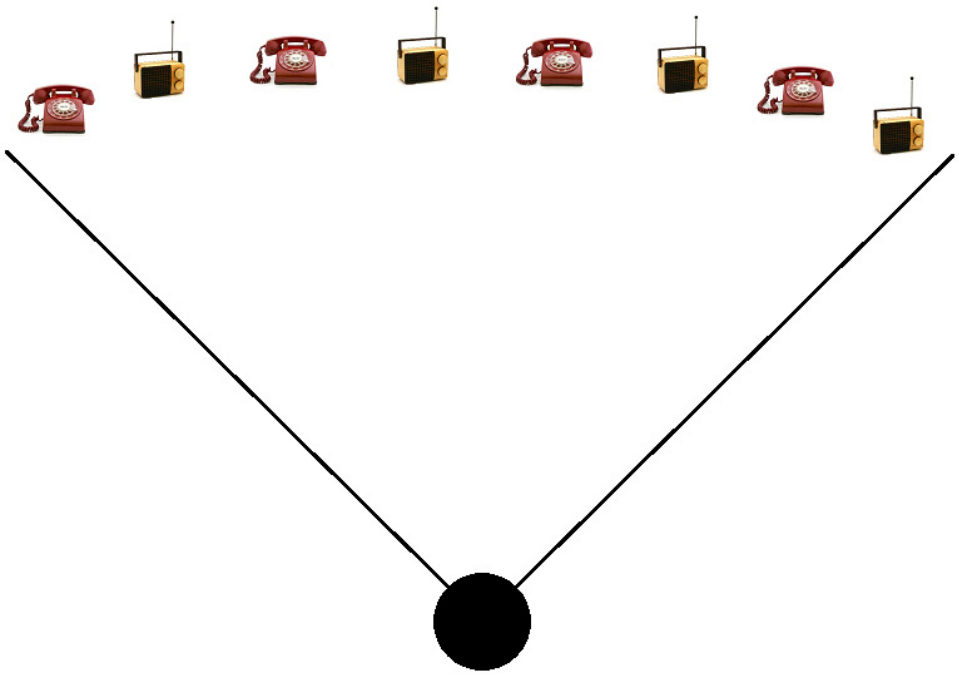


Figure 10: First possibility for the objects' positions, with a regular one by one alternation between phone and radios

The assumption can be made that this configuration will restrict the possible semantic incongruence effects that are expected to emerge from this experiment: in a semantically incongruent configuration, where for example a radio sound would be played at a phone's position, it can be expected that the subject will locate the

source at one of the two neighboring objects (one of the two radios aside the phone) - Figure 11. It seems unlikely that the subject will locate the sound source at a position further than these two possibilities. Consequently, the *magnitude of the expected semantic incongruence effect* [2] cannot be evaluated in an adequate way with this configuration.

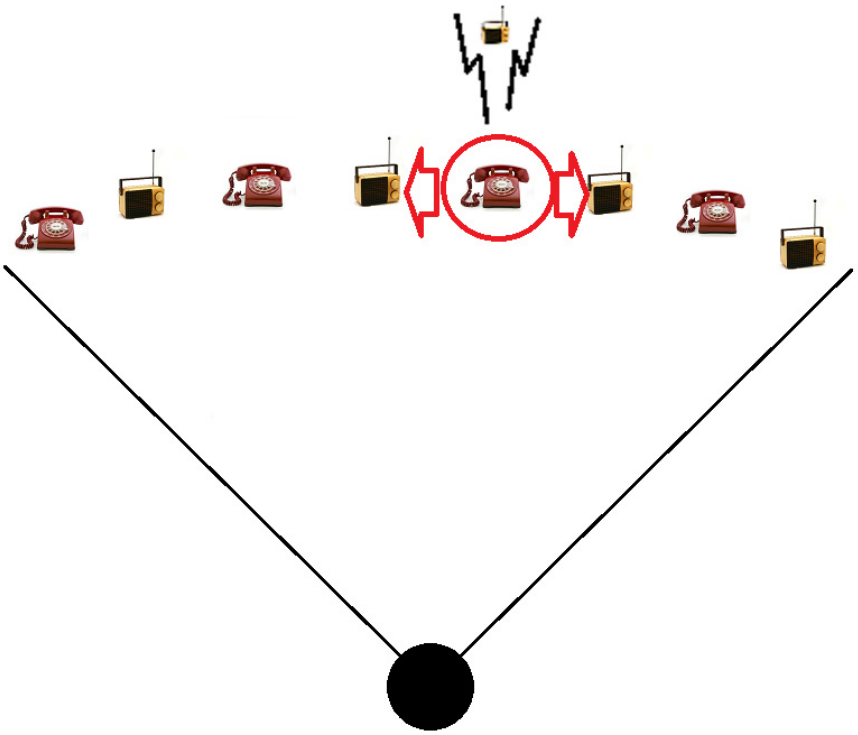


Figure 11: Assumption concerning the results of a localization trial in semantically incongruent position. In this regular one-by-one alternation between phones and radio, it can be assumed that a sound played at the position of an incongruent object will be located in the limits of the two neighbouring objects.

On the other hand, this configuration seems adequate to measure the influence of semantic congruence on the localization task: one of the hypotheses is that semantic congruence will improve such a task in comparison with a situation with abstract stimuli. In this configuration, as every object is surrounded by two other objects of a different nature, semantic congruence can possibly play this expecting role by discriminating the surrounding objects that are not congruent with the audio.

2 In this research, *magnitude of the semantic incongruence effect* will define a measurement of the distance between the position of the sound source and the closest semantically relative object, when such an effect occurs.

With a regular two-by-two alternation of phones and radios (Figure 12), the same limitation as previously mentioned can be expected for the incongruent situation; in this case, though, only one neighbor of each object would be compatible with an incongruent sound. This configuration also seems less adequate for the observation of the expected effect of semantic congruence on the improvement of the localization task, as for each object, its duplicate placed just on its side can induce confusion. As a consequence, this configuration will be discarded.

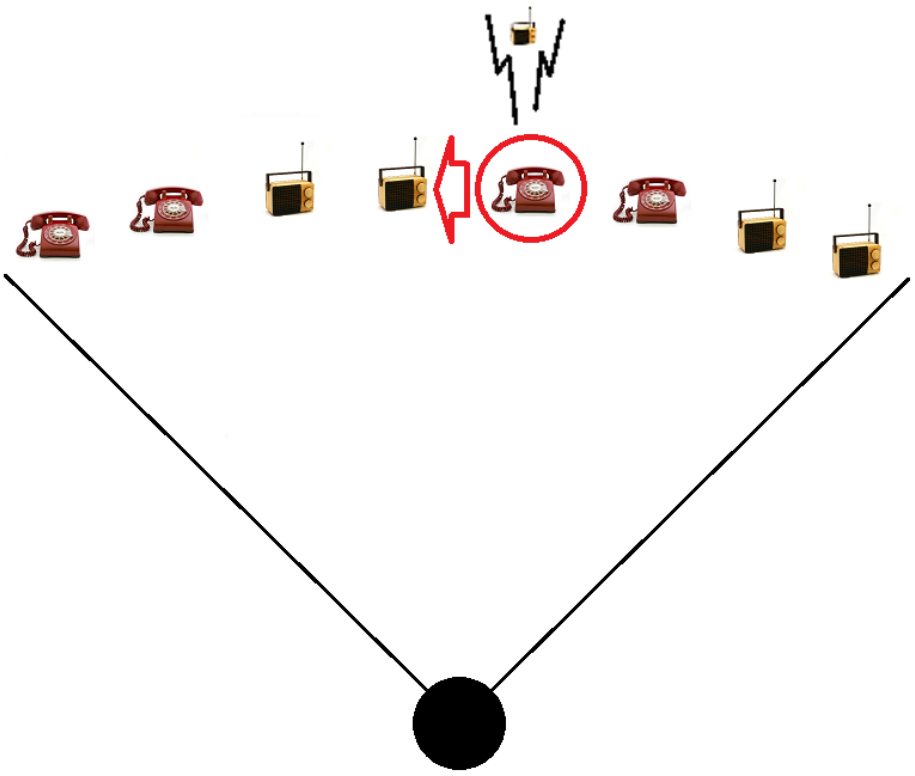


Figure 12: Assumption concerning the results of a localization trial in semantically incongruent position. In this regular two-by-two alternation between phones and radio, it can be assumed that a sound played at the position of an incongruent object will be located in the limits of the one neighbouring object congruent with the sound.

It seems consequently to be more interesting to place all the phones consecutively on one side, and all the radios of the other. In this way, the magnitude of the expected semantic incongruence effect could be evaluated.

In a configuration where all the phones are consecutively placed on one side and all the radios on the other, the effect of semantic incongruence can be expected to occur. Moreover, this configuration allows to variate the distance between the sound source and the position of the next congruent object (previously referred as the magnitude of semantic incongruence).

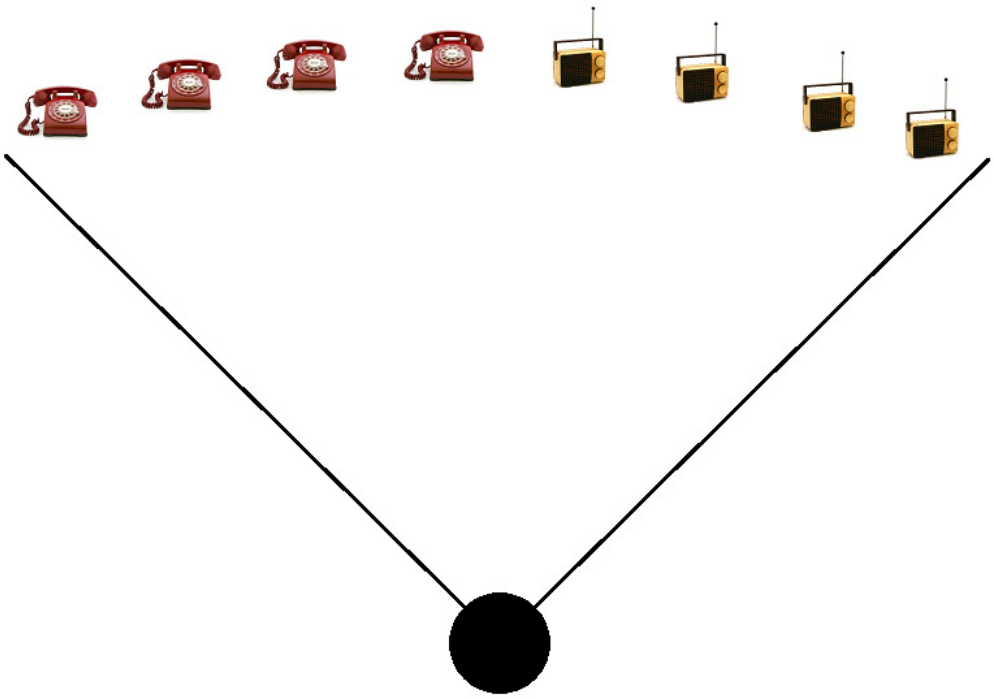


Figure 13: A four-by-four alternation positioning of telephones and radios

An extreme case would be for example to play a radio sound on the phone placed completely on the left. A more subtle incongruence would be to play a radio sound on the phone placed on the most right position, consecutive to a radio. In this way, the magnitude of the spatial difference can be variated, in order to measure until what point semantic incongruence can overtake the perception of location of the sound source.

As a consequence, it will be chosen to use two different configurations for the non-abstract conditions: first, a test with telephones and radios placed in a one-by-one alternation, with which the hypothesis about the effect of semantic congruence on spatial localization will be tested. Secondly, a placement of telephones and radios in a four by four alternation will be used to test the hypothesis on the effects of semantic incongruence.

2.3 - EXPERIMENTAL PROCEDURE

In the configuration involving abstract stimuli, the subjects will have to perform sixteen localizations, with two trials for each of the eight possible positions, played at a random order.

In the configuration involving non-abstract stimuli, sixteen localizations will also be performed by the subjects, two for each possible position. In the semantically congruent condition, the order in which these trials will be conducted will also be random, as in the abstract condition. However, in the test of the semantic incongruence effect, the order of the trials will be fixed for the spatial difference between the sound source and the next congruent object to increase over time. This choice is to avoid having first trials with too big spatial incongruences that would lead the subjects to understand that phone sounds can come from radios and vice-versa. The order of the trials will consequently be made for the spatial differences to increase step by step.

Sounds semantically congruent with the object at their source will also be added to the playlist, also to avoid subjects realizing that semantic incongruences might take place. The whole playlist will contain eight semantically incongruent sounds (for each object's position) and eight semantically congruent sounds.

There will be a regular alternation between congruent and incongruent sounds. Every odd trial (trials 1, 3, 5, etc.) will involve sounds congruent with the object positioned at their sources. On the other hand, every even trial (trials 2, 4, 6, etc) will use a sound incongruent with the object placed at its source position.

The 2^{nd} and 4^{th} trials (Figure 14.b and Figure 14.d) will have the smallest magnitude of incongruence (the closest object semantically incongruent with the sound is the next one to the source). The 6^{th} and 8^{th} trials (Figure 14.f) will then go one step further in the incongruence: the closest semantically congruent object to the sound played will be placed two positions from the source. In the 10^{th} and 12^{th} trials, the closest object congruent with the audio stimulus will be three objects further from the sound source, and so on for the following trials (Appendix A contains a complete graphical explanation of the order for the sixteen trials of this test).

The congruent sounds in the odd trials will be arranged so to avoid the same speaker to play twice consecutively. It will also be avoided to have a case where an odd trial (involving congruent stimuli) will play a sound from the object’s position that will be the closest congruent object to the sound in the following trial. Such an occurrence would also seemingly make the presence of semantic incongruence more obvious to the subject.

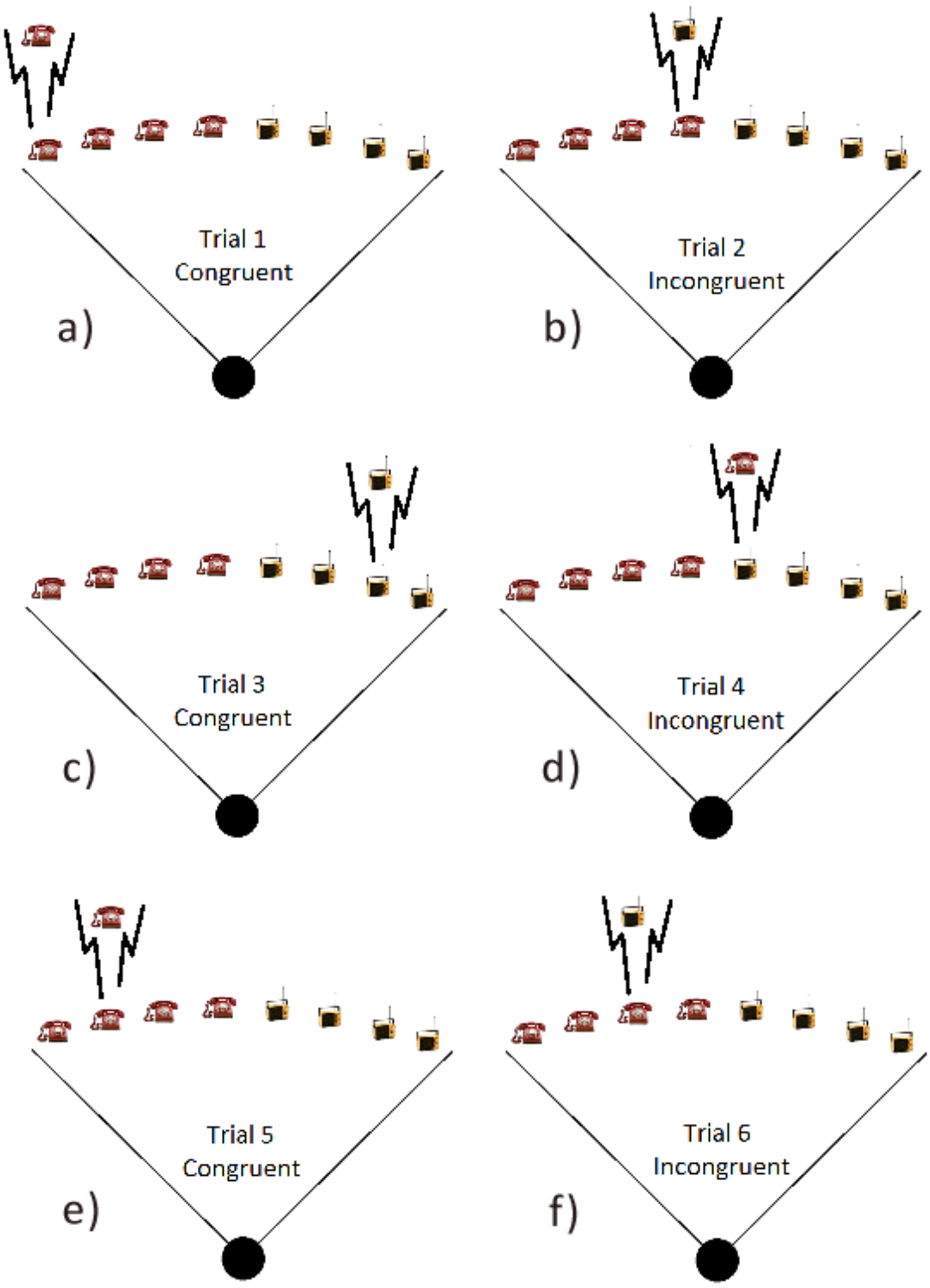


Figure 14: The six first trials of the test on the effects of semantic incongruence on a localization task.

During all the experiments, an audio recorder will be on during the whole experimental procedure. Rough reaction times will be extracted from it,

corresponding to the time difference between the beginnings of a localization trial (the moment when the sound is played) and when the subject locates a target.

PART III –DATA ANALYSIS

CHAPTER 3 - METHOD

3.1 - TEST MATERIAL

The experimental setup required eight audio speakers (ESI nEar 04 Speakerset), plugged to the eight analogic outputs of a fireware soundcard (RME Fireface 800). The RME Fireface 800 soundcard was plugged to a controlling station running Microsoft Windows 7. The audio outputs were controlled through the mixer software provided with the drivers of the soundcard. For each trial, the channel corresponding to the sound position was selected through the software mixer. The sound files were played using the Winamp media player (ver. 5.57).

Prior to conducting the tests, a few preparations were necessary in order to run the test as intended. Each test was conducted in one room location with movable panels separating the subject from the experimenter controlling station.

For the abstract configuration test, four pyramids and four cubes differentiated by different colours (red, blue, yellow and green) where placed in front of the subject in an alternate order as shown in Figure 15.

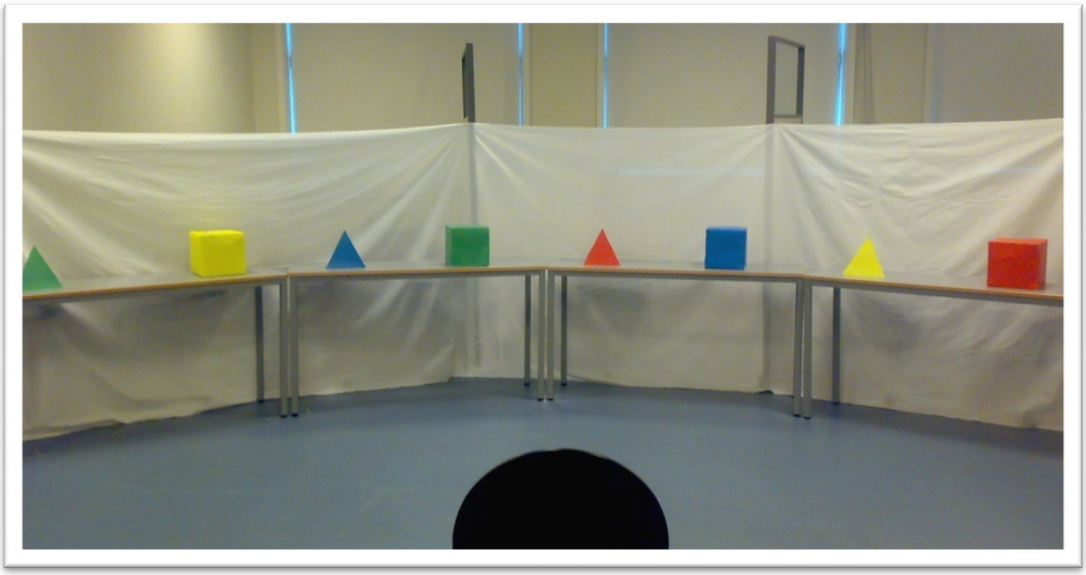


Figure 15: Experimental setup for the abstract condition test

In the semantically congruent and incongruent setups, four identical phones and radios were used as visual references. Coloured tapes were used to make them identifiable by the subjects according to the tape colour. As mentioned in paragraph 2.2 - phones and radios were placed in an alternate order in the semantically congruent condition (Figure 16) whereas phones and radios were grouped

respectively on the left and on the right for the semantically incongruent condition (Figure 17).

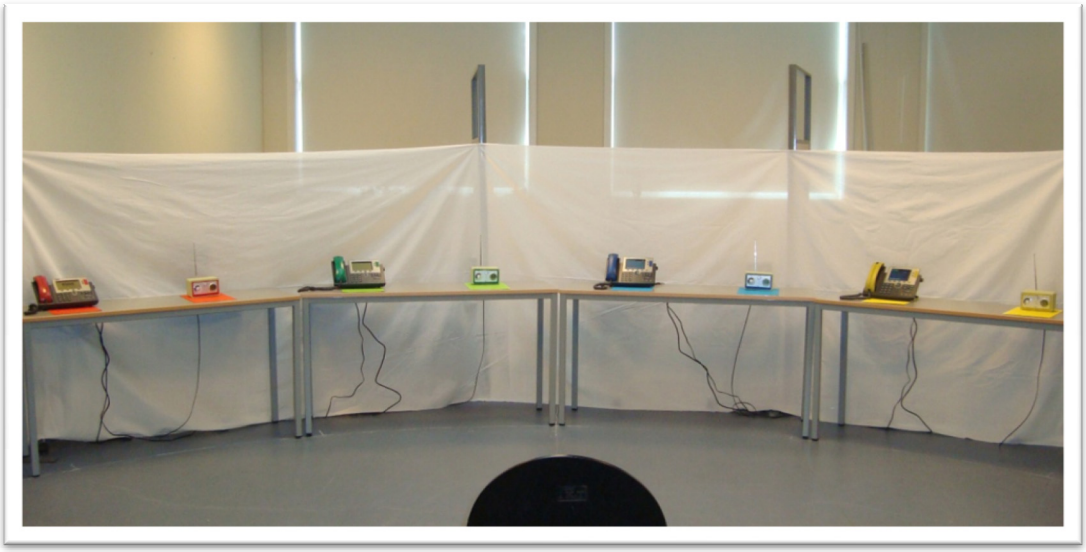


Figure 16: Alternate placement of phones and radio for the semantically congruent test

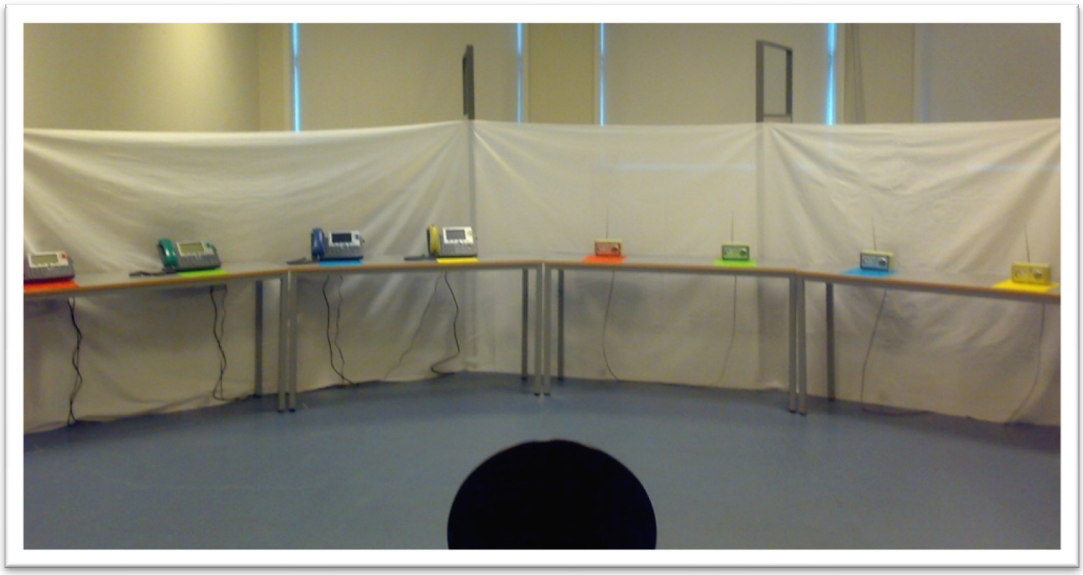


Figure 17: Grouped placement of phones and radio for the semantically incongruent condition

3.2 - TEST PROCEDURE AND PARTICIPANTS

The three experimental tests were taking place on three different days. The abstract, semantically congruent and semantically incongruent tests were conducted respectively May 04th, 09th and 12th 2010 in one of the Medialogy sound laboratory at the campus of Aalborg University Copenhagen.

Forty-five subjects participated to the experiment, fifteen for each condition. The participants were all students met in the university building (between 22 and 31 years old) without any gender preference. The nature of the tests required independency of subjects for each test, hence who took part in one test condition could not take part to any of the other tests.

The tasks performed by the subjects were the same in all the test conditions, hence the instructions given to the subjects were the same. The instructions were written on paper and read to the subjects to make sure that they were explained in the same way to all the participants (see Appendix B). All of them were welcomed and thanked for taking part of the test, subsequently the placement of the objects was discussed (in order to detect any case of colour blindness) and the task they had to perform was explained to them. Finally they were informed there was only a microphone recording their responses and there was not any video recording.

Besides this information, participants were unaware of what was tested for the purpose of avoiding any bias. At the end of the test they were offered snacks and soft drinks as a token of our gratitude.

Moreover, to minimize any experimenter influences, the order of sounds played was fixed in three playlists relative to the test condition and the sound sources were controlled over a computer interface. Figure 18 summarises the three conditions in the test design.

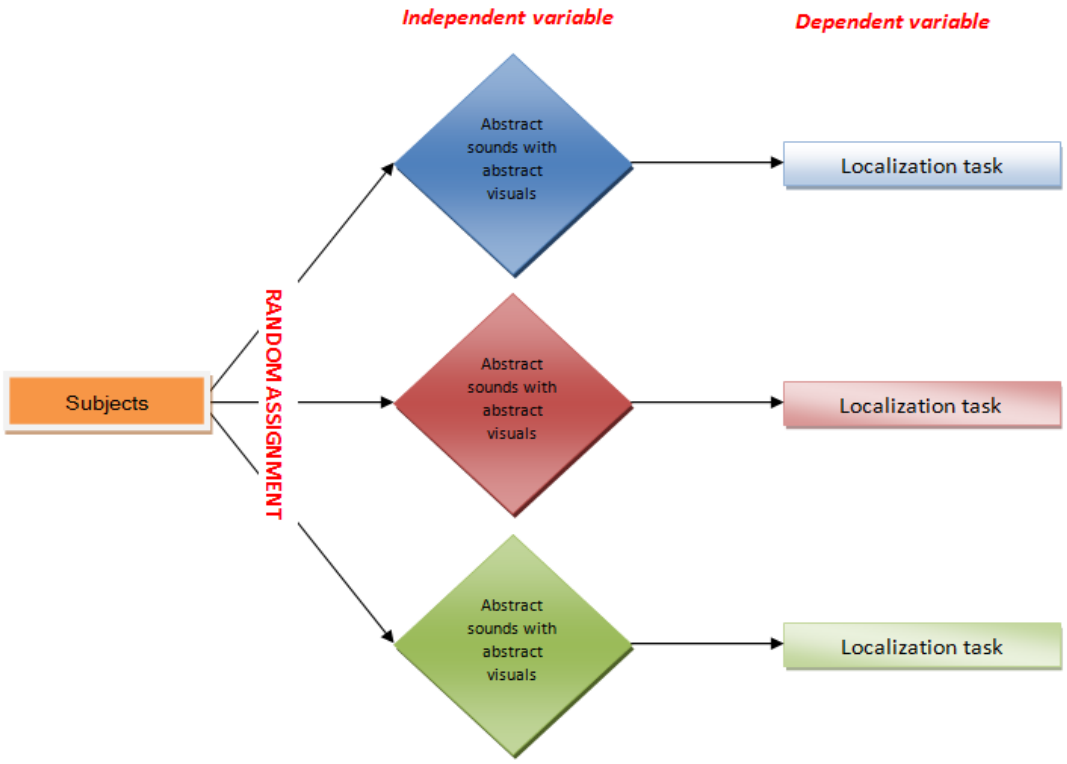


Figure 18: Test design schema

CHAPTER 4 - RESULTS

From the experimental design it becomes clear what would be the most appropriate statistical procedure in order to evaluate the significance of the data.

In statistics data are usually divided in two main categories: *measurement data* (also referred as quantitative data) and *categorical data* (also called frequency data) (Cohen and Lea, 2004, p. 455). The first group involves observations that represent a score along a continuum whereas in the second one the data consist in frequencies of observations that fall into two or more categories.

Although mean and variation are the most common way to describe the distribution of measurement data, the experimental design is use in this research involves the gathering of categorical data, from which the relative frequencies between different populations must be extracted. The Chi Square analysis method is adequate in this situation, and will be the main statistical tool in use in this research.

More specifically, the data from each test configuration will provide simply frequencies as the number of trials (or stimuli) perceived as coming from a specific object, such as a yellow phone or a green radio. Therefore, in each test condition

each object corresponds to a category and each subject response falls in one of them. In other words, in this case data do not involve any nominal or ordinal values but rather the perceived source for each trial.

4.1 - TEST RESULTS

4.1.1 - FIRST HYPOTHESIS VERIFICATION

The first task in the analysis of results is to verify the validity of the first hypothesis of this study.

> **Hypothesis 1:** *Subjects would be more precise in localizing audio stimuli which is semantically congruent with the visual object located at the sound source, than localizing abstract audio stimuli that presents none or little semantic relations with the object located at the sound source.*

To verify this hypothesis, the amount of successful localizations in the experiment with abstract stimuli will be compared with the amount of successful localizations in the experiment with semantically congruent concrete stimuli. A *Chi-Square test* will be used to check if there is a difference in the distribution of successful and failed localizations between the two different conditions (abstract and semantically congruent). If a difference is observed, and if the amount of correct localizations in the semantically congruent test is higher than in the abstract one, then the first hypothesis of this study will be verified.

4.1.1.a ABSTRACT CONDITION RESULTS ANALYSIS

In the abstract condition, on an overall amount of 240 trials (16 trials for each 15 subjects), 218 successful and 22 failed localizations were observed (91% of successful trials, as shown in figure 19).

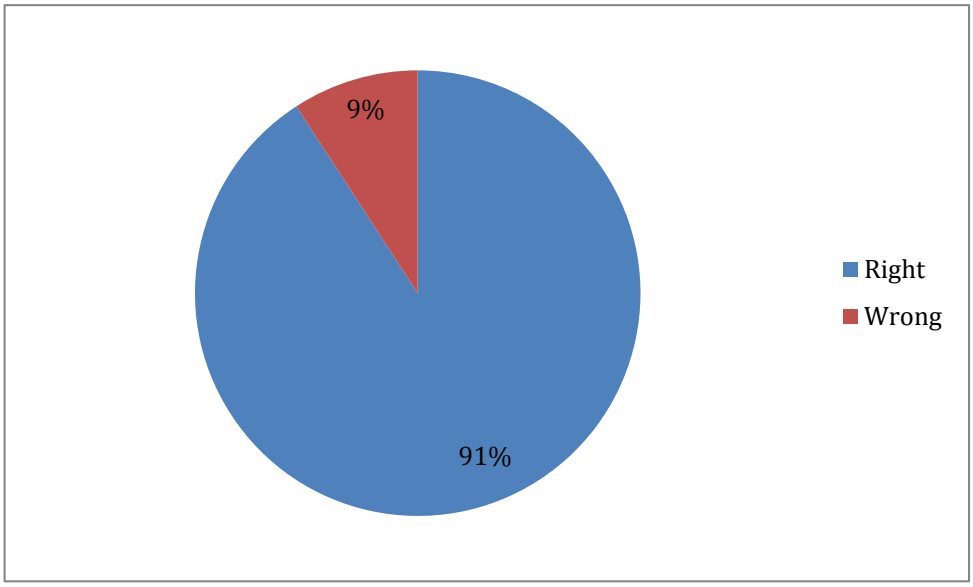


Figure 19: Percentage of correct and incorrect localization trials obtained in the experiment with abstract conditions

Among all subjects, an average amount of 15 correct localizations out of 16 with a standard deviation of 1.12 (*2SD*=2.25) is consequently reached (as shown in figure 21). This relatively low standard deviation permits to say that the results are relatively homogeneous among all subjects (see figure 20 for the amount of correct localizations per subject).

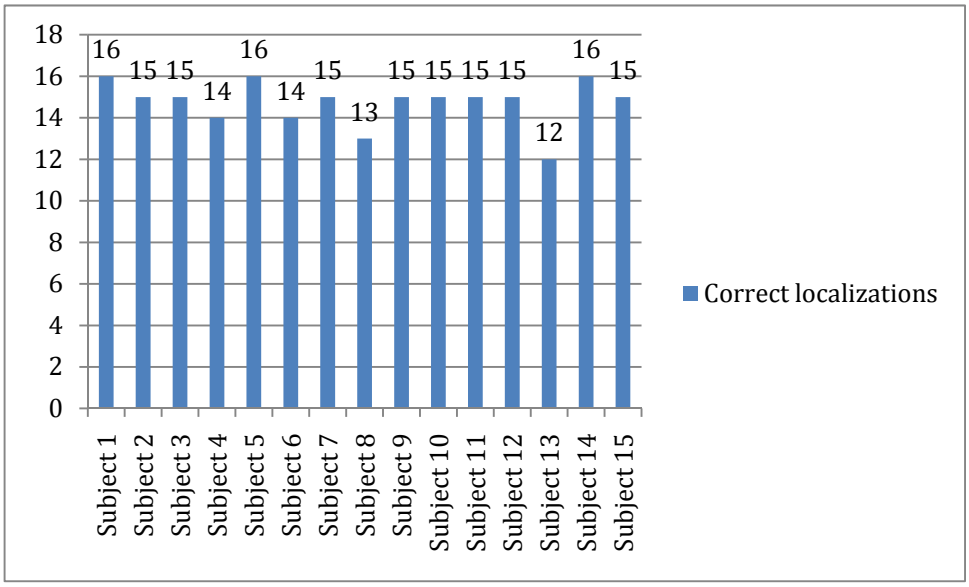


Figure 20: Amount of correct localizations per subject in the experiment with abstract condition; the values are in a range between 12 and 16.

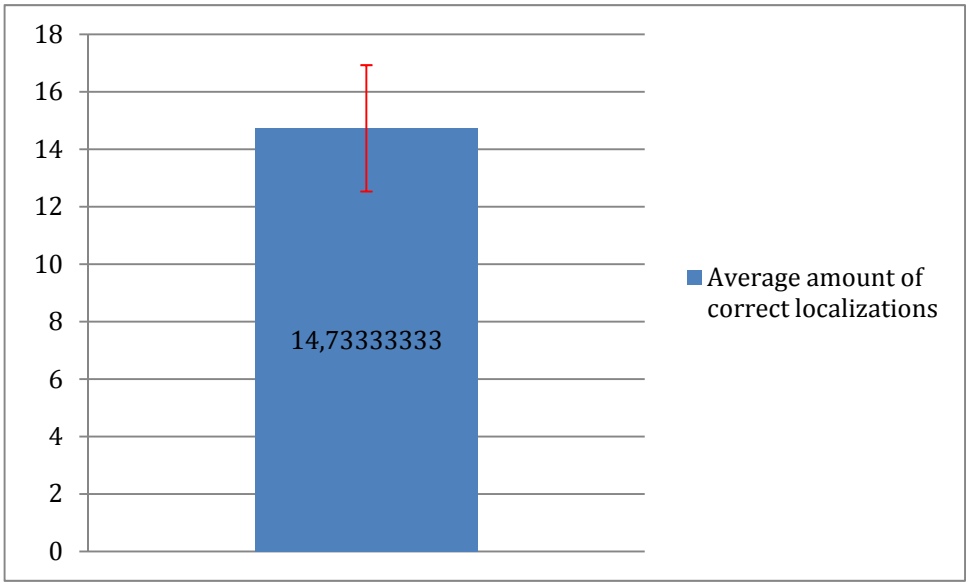


Figure 21: Average amount of correct localizations across subjects. The error bar of ±2.25 represents two standards deviations

The total amount of correct localizations in the experiment with abstract conditions will be used as the Null Hypothesis in the following Chi-Square test on the results obtained in semantically congruent condition (this null hypothesis has to be violated in order to verify the first hypothesis of this study, by proving that localization was improved by the introduction of congruent semantics).

4.1.1.b SEMANTICALLY CONGRUENT CONDITION RESULTS ANALYSIS

In the semantically congruent condition, on a total of 240 trials, 237 successful and 3 failed localizations are observed (98.7% of successful trials), with a standard deviation among subjects of 0.41 (*2SD*=0.82). The three incorrect localizations relate each to a different subject and a different source position. Consequently, it can be concluded that the results are homogeneous across subjects and sound source positions.

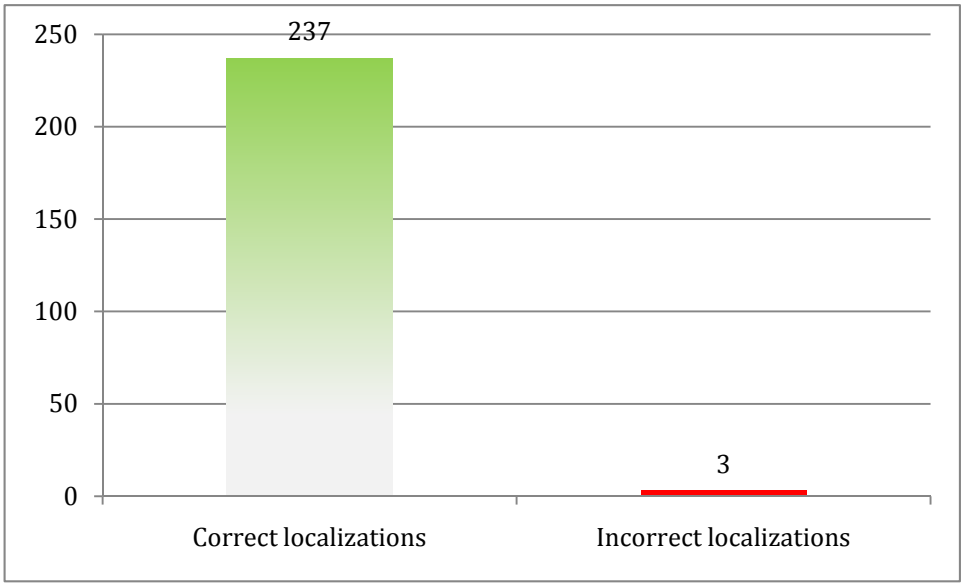


Figure 22: Total amount of correct and incorrect localizations in the semantically congruent condition

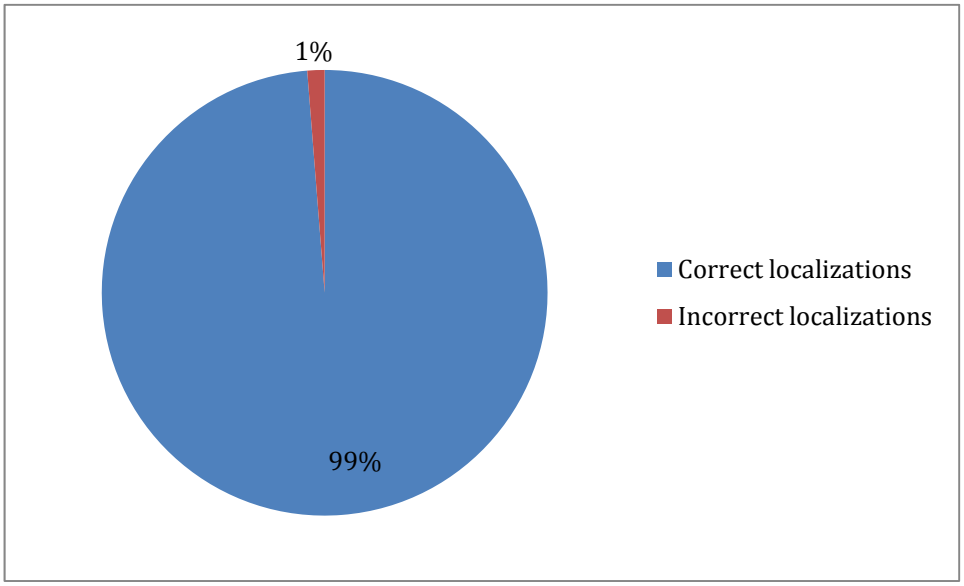


Figure 23: Percentage of correct and incorrect localization trials obtained in the experiment with semantically congruent conditions

4.1.1.c COMPARISON BETWEEN ABSTRACT CONDITION RESULTS AND SEMANTICALLY CONGRUENT CONDITION RESULTS

The frequencies observed in the semantically congruent condition, computed with a Null Hypothesis based on the results obtained in the abstract condition, show a significant difference between these two sets of data (χ^2=18.1, *df*=1, $p \leq 0.001$).

Nevertheless, a closer analysis of the results gathered in the abstract condition permits to observe an unusual amount of failed localizations on the first trials of the subjects, with 8 subjects over 15 failing in their first trials (53.3% of failure on first trials, compared to an overall rate of 9.2% of failures).

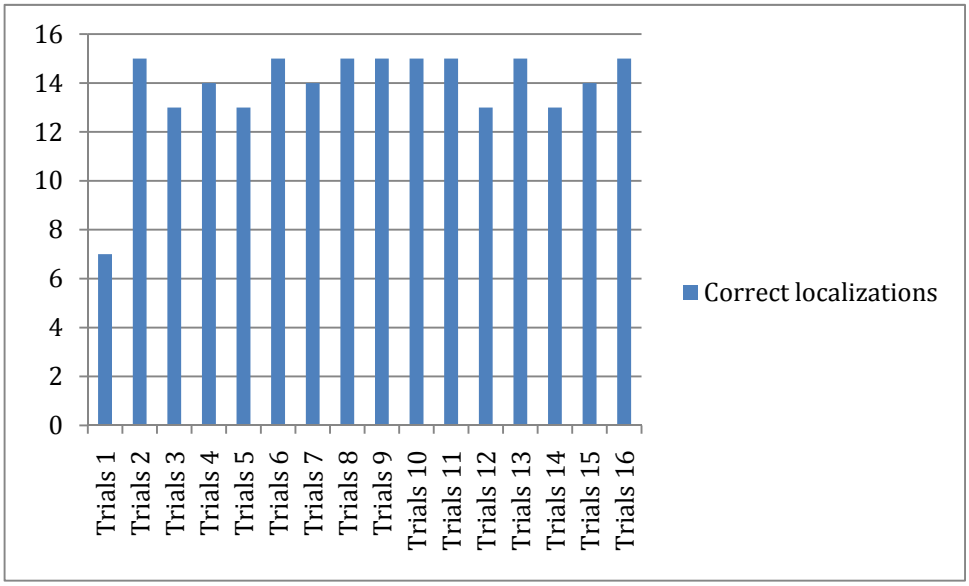


Figure 24: Amount of correct localizations in function of the number of the trial. The first trial is the source of a particularly high amount of mistakes, which permits to suppose the existence of a fast learning curve from the subjects at the beginning of the experiment.

It should be underlined that in the semantically congruent situation, this effect did not appear. The hypothesis can be made that in the abstract condition, the absence of training (the subjects were not exposed to the auditory stimuli before the beginning of the experiment), plus the assumption that no semantic link existed between these sounds and the coloured shapes, lead to the appearance of a learning curve at the beginning of the experiment. This learning curve is obviously very short as the results are relatively homogeneous from the trials 2 to 16. In the semantically congruent condition, the fact that the sounds and visual objects were already semantically linked (due to the past experience of subjects on telephones and radios) possibly permitted to skip this very short process of adaptation to the task.

A new Chi-Square test can be conducted after removing the first trial of every subject in both abstract and semantically congruent situations. This changes the overall amount of trials to 225 in each situation. The rate of successful trials is now brought to 93.8% in abstract condition and 98.7% in semantically congruent condition.

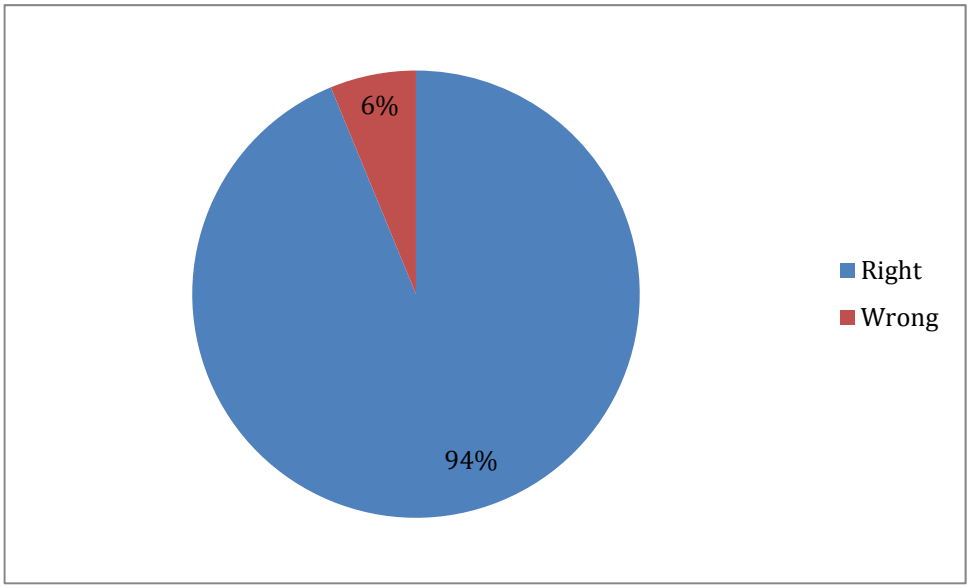


Figure 25: Percentage of correct and incorrect localization trials obtained in the experiment with abstract conditions, after the removal of the first trials

The statistical test still shows that there is a significant difference between the two sets of data (χ^2=9.22, *df*=1, $p \leq 0.01$).

Moreover, the higher rate of correct localizations in semantically congruent conditions (98.7%) compared to the one in abstract conditions (93.8%) shows that semantic congruence between auditory and visual stimuli enhanced the performances of subjects. The first hypothesis leading this study is consequently confirmed.

4.1.2 - SEMANTICALLY INCONGRUENT CONDITION RESULTS

In the condition involving auditory stimuli semantically incongruent with the visual objects at their source, the spatial difference between the sound source and the closest congruent object was varied over four steps. The localization performances will be analysed separately depending on this variable. Half of the trials in this condition were semantically incongruent, and the other half was semantically congruent, in an attempt to avoid the subjects realizing early in the experiment that semantic incongruence could occur. In this result analysis though, only the trials involving semantic incongruence will be considered.

Among these semantically incongruent trials (120 trials), there is a rate of 77% of correct localization(Figure 26: Percentage of correct and incorrect localization trials obtained in the experiment with semantically congruent conditions. This rate is

substancially lower than in the experiment with abstract stimuli (90.8% of successful localizations).

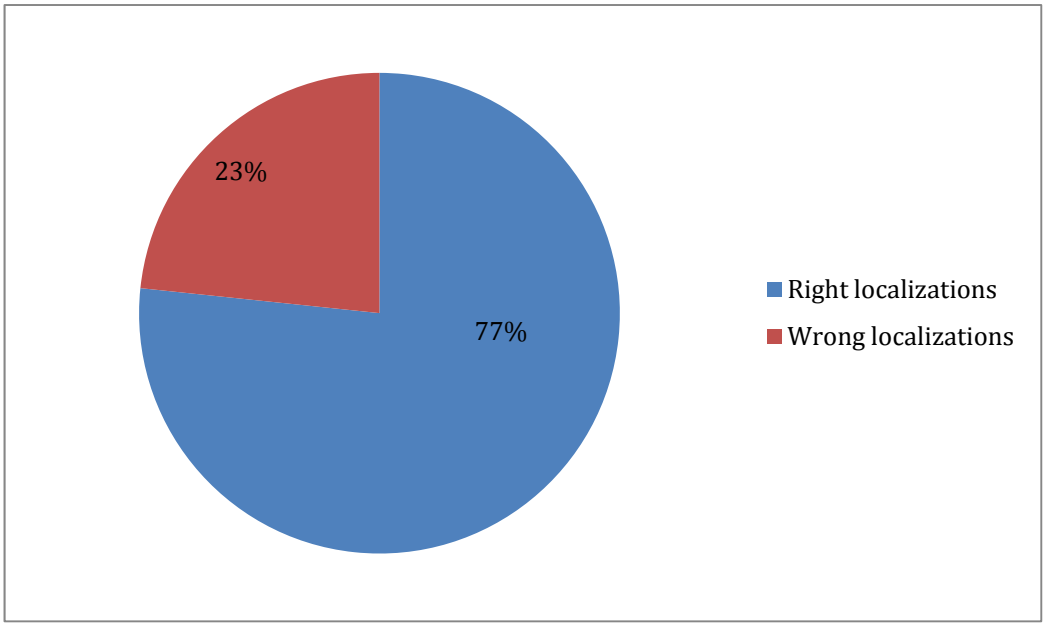


Figure 26: Percentage of correct and incorrect localization trials obtained in the experiment with semantically congruent conditions

The average amount of correct localizations per subject is of 6.1, with a standard deviation of 1.6 (*2SD*=3.2) as shown in Figure 27 and Figure 28.

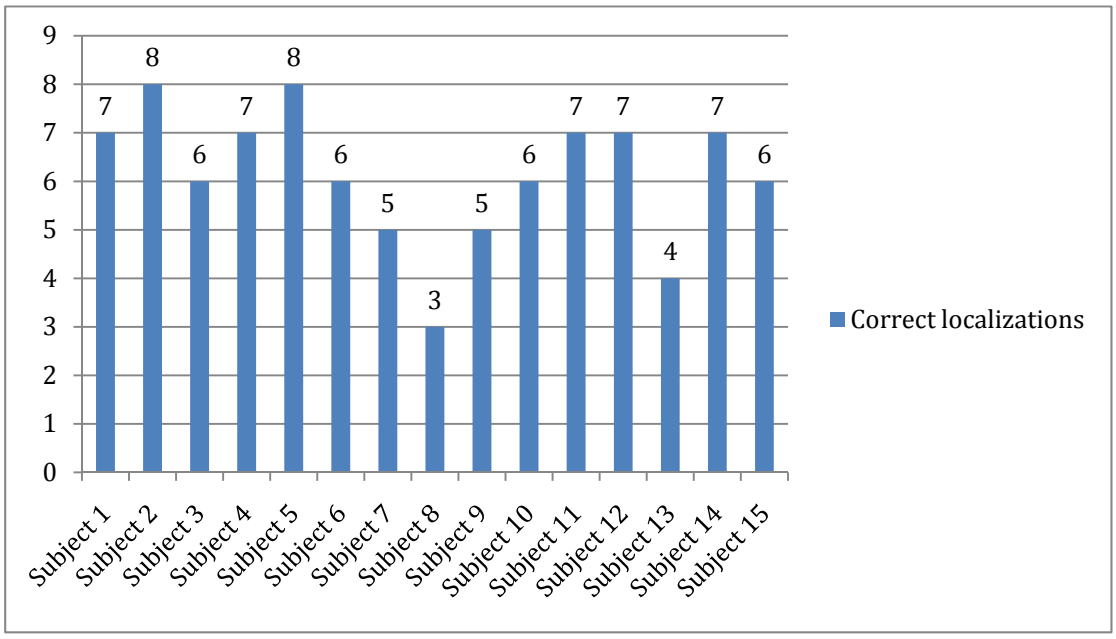


Figure 27: Amount of correct localizations per subject in the experiment with semantically incongruent condition

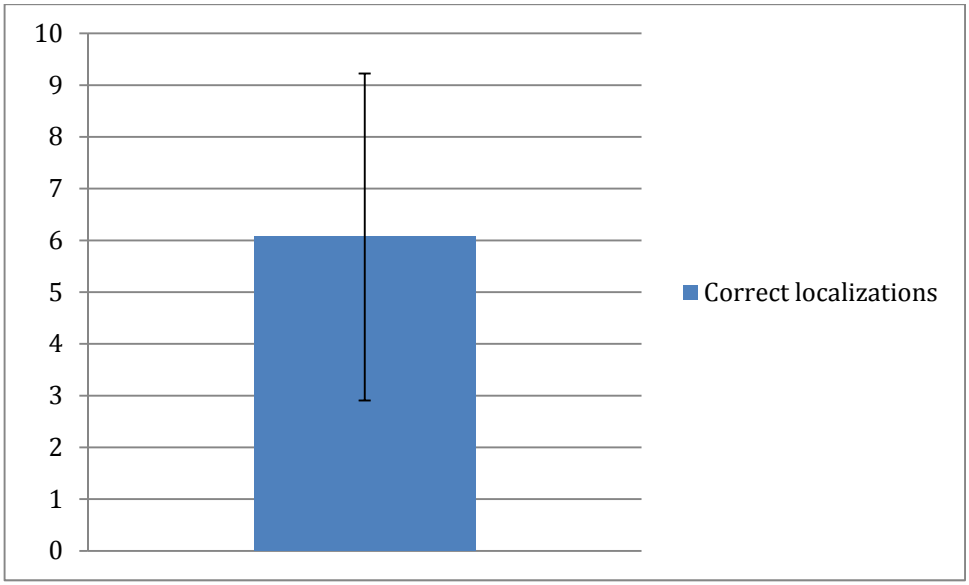


Figure 28: Average amount of correct localizations across subjects during the semantically incongruent condition. The error bar of ±3.2 represents two standards deviations

Trial #	Distance to the closest semantically congruent object	Occurrences of the expected semantic incongruence effect
1	1	1
2	1	3
3	2	0
4	2	0
5	3	0
6	3	0
7	4	0
8	4	0

Table 1: Mismatch between sound and semantically relative object for each trial and the number of the semantically congruent replies

It is interesting to observe that only few subject had a behaviour following the second hypothesis laid in this study (implying that subjects would choose the closest congruent visual object as the source of the sound). This behaviour could only be observed in the specific case where the distance between the sound source and the

next semantically congruent object was minimal (a gap of one object). In this specific case, only four trials over 30 permitted to observe this behaviour (Table 1).

In the other cases, where the distance between the sound source and the next semantically congruent object was over the minimum value (a gap of more than one object), there are no trials reporting a relation between the sound source and the closest semantically congruent object.

PART VI – DISCUSSION & CONCLUSION

CHAPTER 5 - EVALUATION AND DISCUSSION

Of the two hypotheses tested in this study, only the first one was verified: the semantic congruence between an auditory stimulus and the object at its source lead to more precise localization performances than with abstract stimuli. The second one, proposing that when an auditory stimulus is semantically incongruent with the object located at its source the subjects would perceive the sound source as coming from the closest semantically congruent object, failed to be verified.

The localization performances of subjects in the semantically congruent condition were significantly better than in the abstract condition, which permits to say that the semantic congruence between the auditory stimulus and the object at its source helped the subjects to localize the source more accurately. In other words, the semantic link between the two types of objects used in the experiment and their specific sound, learned through past experience by interacting with such objects, did influence the subjects in the perceptual task performed.

This result proves that during a localization task of an auditory stimulus, the structural factors of the auditory signal are not the only ones to influence the final decision of the subjects: it appeared that semantic congruence, a top-down factor emerging from past experience, also has an influence on this decision. Consequently, the results of this experiment support Helmholtz' theory of unconscious inference, according to which perception is a deductive process, based on knowledge gathered in an inductive way through past experience. It could have been interesting to have an analysis of the subjects' reaction times, to see if semantic congruence also made the subjects faster in their localization performances (as it could be expected from the results obtained by Suied et al., 2008; Gallace and Spence, 2006; Yuval-Greenberg and Denouell, 2008). Such an observation could have allowed us to have a better idea of to which extent semantic congruence was decisive in this location decision (was semantic congruence a factor merely refining a decision mainly based on structural factors? Or in other words: to which degree was semantic congruence an important factor in such a task?). Nevertheless, this question is a complex one from which it is only possible to speculate at the moment.

The fact that a sort of learning process occurred in the abstract condition but not in the semantically congruent condition allow us to advance the hypothesis that this learning process was needed for the users to build a link between the visual objects (three dimensional shapes) and the abstract sounds. This link did not have to be built in the semantically congruent condition, as the subjects had already learned by experience the semantic link between a phone and its ringing tone. To verify this

hypothesis, it could be interesting to conduct again the experiment in the abstract condition, but this time with the subjects being introduced to the stimuli before the beginning of the localization trials.

The hypothesis according to which the subjects had to first experience the stimuli and build meaning from it before being able to locate the auditory event in an accurate way, can also lead to interesting speculations on human perception, and how meaning is built from the multimodal stimuli surrounding a person. Here, it will be taken as a proposition going in the direction of the argument that perceptual performances depend highly on the semantic links between events through experience.

On the other hand, the results acquired in the semantically incongruent condition did not give satisfactory results. Consequently, they did not allow the verification of the hypothesis stating that when an auditory stimulus is semantically incongruent with the object located at its source, subjects would perceive the source at the place of the closest semantically congruent object.

The experimental conditions used in this study should be modified for further research, in order to assure whether this hypothesis should be rejected or kept as a plausible one. For example, it could be that the distance between each object was too important (in this case too large), and consequently allowed most of the subjects to know very early in the experiment that semantic incongruence could occur. This argument can be supported by the observation that in the abstract condition, the localization task was considered as very easy by the subjects (91% of accurate localization trials occurred in this condition). To avoid this problem, it could have been a solution to conduct several pilot tests with different distance gaps between objects, and analyze the localization performances of the subjects in each configuration. A gap value, for which localization would be less easy while still keeping a relative accuracy, could be selected and the experiment could be re-run in this condition.

Nevertheless, it is interesting to observe that in this experiment with semantically incongruent trials, the amount of successful localizations was much more subject-dependant than in both abstract and semantically congruent conditions (figure 28). An explanation for this effect can be that the presence of both semantically congruent and semantically incongruent trials in the same experimental procedure may have confused some subjects more than others. Some of them may have "learned" early in the experiment to ignore semantic congruence and incongruence and perform the localization task independently of this factor (this would concern the most accurate subjects). Others may have been distracted by this alternation

from semantic congruence to semantic incongruence; the performances of these subjects would correspond to the less accurate ones observed in the results. In further experiments, it would be better to avoid such a regular alternation from semantic congruence to semantic incongruence. Rather, it could be a better idea to use less semantically incongruent trials per experiment, in such a way that they would be "hidden" in a larger amount of semantically congruent trials.

CHAPTER 6 - CONCLUSION

Semantic congruence as a perceptual factor still has to be thoroughly studied to have a beginning of insight on how much perception may depend on this factor, and on past-experience in general. The theory of unconscious inference has much to gain from experiments on semantic congruence, as this concept permits to observe some influences of past experience on perception.

The experimental work described in this report permitted to obtain some interesting results, which shed light on some effects that semantic congruence between stimuli has on human perception. Mainly, it was observed that the localization of an auditory stimulus is generally executed in a more accurate way when such stimulus is semantically congruent with the object placed at its source, than when both auditory stimulus and visual object are abstract and therefore not incorporate in the subject's history. This result permits to underline that the localization of an auditory event (as perceptual process) does not solely depend on the structural factors of the auditory stimulus, but is also dependent on top-down factors such as semantic congruence emerging from past-experience.

An interesting an unexpected result permitted also to observe that in such abstract conditions a learning process, though relatively short (one localization trial), is generally occurring whereas in semantically congruent condition this process was not observed. Even though no definitive conclusion could be made concerning the reasons of this observed process, it is possible to advance the hypothesis that subjects had to first "learn" a semantic link between the abstract auditory stimuli and the objects placed at their source before being able to efficiently perform the localization task.

It also appeared that the experimental conditions in the semantically incongruent test used in this study were not adequate for the observation of an effect of semantic incongruence. Nevertheless, semantic incongruence should be at the basis of further experiments before a decision can be taken on whether the hypothesis is invalid or not (hypothesis stating that when an auditory stimulus is semantically incongruent with the object at its source, subjects would perceive the source at the

place of the closest semantically congruent object). The discussion on the reasons why the experimental conditions in use here failed to allow the observation of effects of semantic incongruence can be used as a starting point for the design of such further experiments.

Finally, the whole experiment allowed us to investigate the use of some spatial audio techniques in the frame of cognitive and perceptual experiments. The overall design process followed in this research gives a deeper insight on how multimedia technologies can enhance the possibilities available to cognitive scientists. For instance, an interesting further development in the investigation of the cognitive effects of semantic incongruence could involve the use of Brain Computer Interfaces. This could give deeper insight on brain-activities while processing semantically incongruent stimuli.

BIBLIOGRAPHY

R.J.Audley, C.P.Wallis, (1964) “Response instructions and the speed of relative judgments. I. Some experiments on brightness discrimination”, *British Journal of Psychology*, 55, p.59-73

P.Bertelson, J.Vroomen, B.de Gelder, J.Driver, (2000) “The ventriloquist effect does not depend on the direction of deliberate visual attention”, *Perception & Psychophysics*, 62, p.321-332

M. Bréal, (1897) ”Essai de sémantique: science des significations”

C. Spence (2007),”Audiovisual multisensory integration”, *Acoust. Sci. & Tech*, 28, p.61-70

C. P. Brown and R. O. Duda, "An efficient HRTF model for 3-D sound," in WASPAA '97 (1997 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics)

P.J. Burke, S.L. Franzoi, (1988) “Studying situations and identities using experiential sampling methodology”, *American Sociological Review,* 53(4), p.559-568

B. H. Cohen, R. B. Lea (2004) “Essentials of statistics for the social and behavioral sciences”, Wiley, New York

J.F.Dashiell, (1937) “Affective value distance as a determinant of esthetic judgment times”, *The American Journal of Psychology*, 50 (1), p.1887-1937

O. Doehrmann, M.J. Naumer (2008) “Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration”, *Brain Research,* 1242, p.136-150

A.Gallace, C.Spence, (2006) “Multisensory synesthetic interactions in the speeded classification of visual size”, *Perception & Psychophysics, 68,* p.1191-1203

C.V. Jackson, (1953) “Visual factors in auditory localization”, *Q. J. Exp. Psychol.*, 5, p.52-65

C. Koppen, A. Alsius, C. Spence (2008) “Semantic congruency and the Colavita visual dominance effect”, *Exp. Brain Res.*, 184, p.533-546

P.J.Laurienti, M.T.Wallace, J.A.Maldjian, C.M.Susi, B.E.Stein, J.H.Burdette, (2003) “Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices”, *Human Brain Mapping*, 19, p.213-223

S.Lehmann, M.M.Murray, (2005) “The role of multisensory memories in unisensory object discrimination”, *Cognitive Brain Research*, 24, p.326-334

M.M.Murray, C.M.Michel, R.Grave de Peralta, S.Ortigue, D.Brunet, S.Gonzalez Andino, A.Schnider, (2004) “Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging”, *NeuroImage*, 21, p.125-135

C.E.Osgood, P.H.Tannenbaum, (1955) “The principle of congruity in the prediction of attitude change”, *Psychological Review*, 62(1), p.42-45

H.R.Pollio, (1964) “Some semantic relations among word-associates”, *The American Journal of Psychology*, 77(2), p.249-256

E.L.Smith, M.Grabowecky, S.Suzuki, (2007) “Auditory-visual crossmodal integration in perception of face gender”, *Current Biology*, 17, p.1680-1685

W.C.Shipley, J.L.Coffin, K.C.Hadsell, (1945) “Reaction time in judgment of color preference”, *Journal of Experimental Psychology*, 35, p.206-215

C. Suied, N. Bonneel, I. Viaud-Delmon (2008), “Integration of auditory and visual information in the recognition of realistic objects”, *Experimental Brain Research*, 194, p.91-102

A. Vatakis, C. Spence, (2007) “Crossmodal binding: evaluating the ‘unity assumption’ using audiovisual speech stimuli”, *Percept. Psychophys.*, 69, p.744-756

J. Vroomen, (1999) “Ventriloquism and the nature of the unity decision: commentary on Welch”, in *Cognitive Contributions to the Perception of Spatial and Temporal Events*, G.Aschersleben, T.Bachmann, J.Müsseler (Eds.), Elsevier, Amsterdam

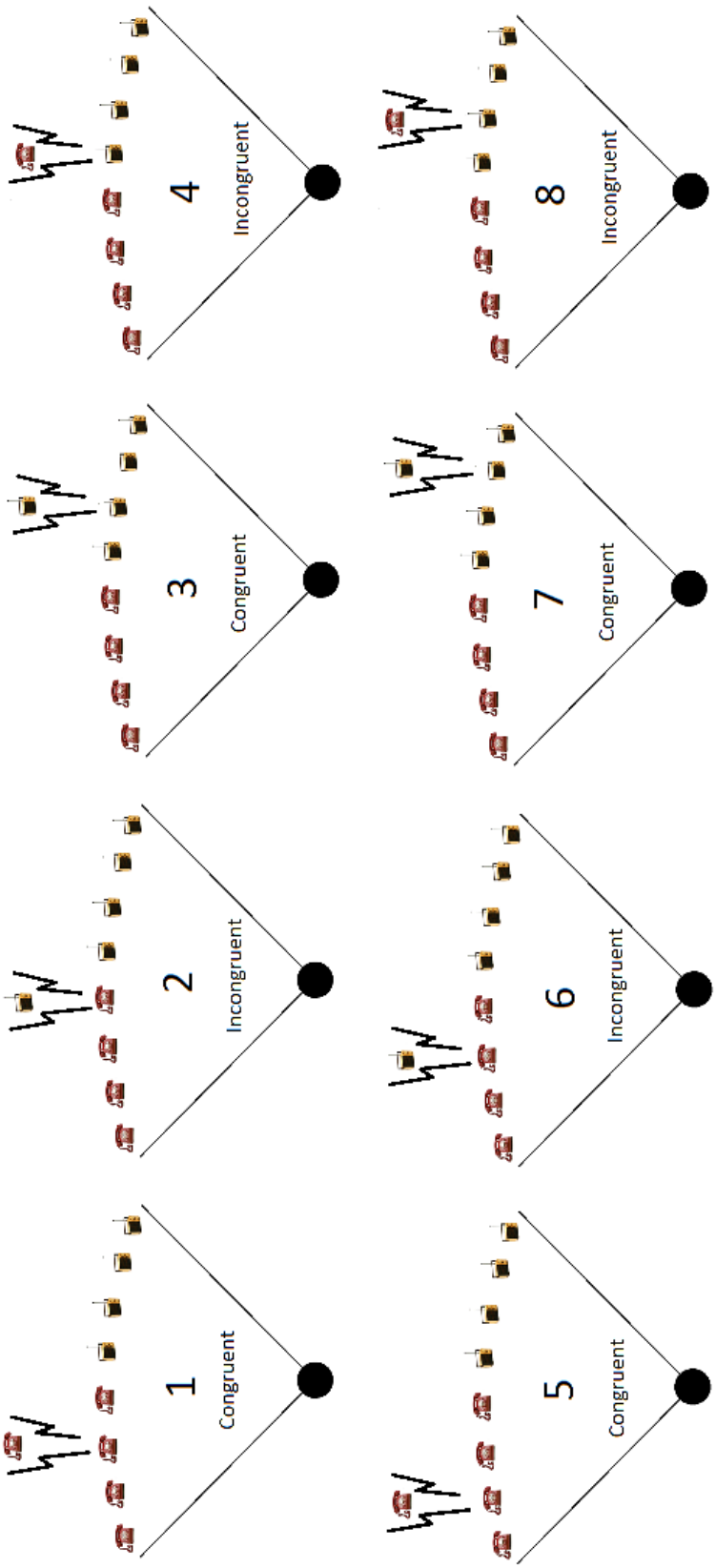
D.H.Warren, R.B.Welch, T.J.McCarthy, (1981) “The role of visual-auditory “compellingness” in the ventriloquism effect: Implications for transitivity among the spatial senses”, *Perception & Psychophysics*, 30, p.557-564

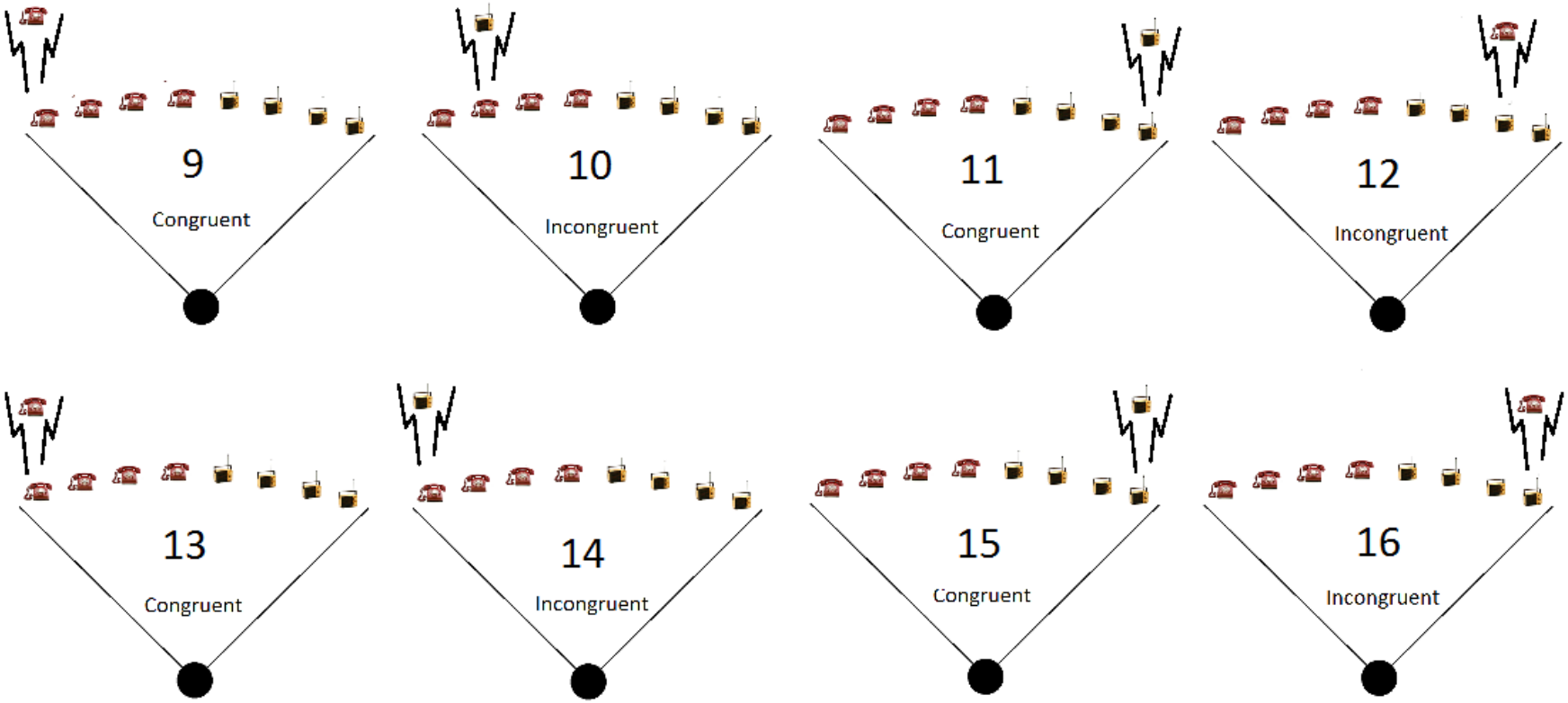
R.B.Welch, D.H.Warren, (1980) “Immediate perceptual response to intersensory discrepancy”, *Psychological Bulletin*, 88, p.638-667

S.Yuval-Greenberg, L.Y.Deouell, (2008)”The dog’s meow: asymmetrical interaction in cross-modal object recognition”, *Experimental Brain Research*, 193, p.603-614

APPENDIX

Appendix A. Order of the sixteen trials for the semantically incongruent condition




9
Congruent
10
Incongruent
11
Congruent
12
Incongruent
13
Congruent
14
Incongruent
15
Congruent
16
Incongruent

Appendix B. Test instructions for the semantically (congruent and incongruent) conditions

Hello and welcome to our test. Thank you for taking part of it. The test is pretty simple, you have in front of you four telephones and four radios identified by colours. Can you clearly see the colours?

We will make them play in a random order. For each sound you have to tell us from which telephone or radio the sound is coming from, using the colours to identify it.

Each trial will be unique and won't be repeated, so please listen carefully.

Your answers will be recorded by a microphone so we can analyse them afterwards.

Appendix C. Binaural Sound

In order to test the effects of semantic congruence on multimodal perception, a controlled audiovisual environment will be created. The idea is to be able to control the occurrence of audio and visual events in time and space, in order to be able to place subjects in front of congruent and incongruent situations.

The environment in question consists in a room, containing a certain amount of objects that are known by the subjects as sound-generating objects (phones or radios for example). The experimenter will trigger audio events at different positions in the room to create the situations of congruence or incongruence. To create this controlled three-dimensional audio environment, binaural sound files containing the different audio events virtually situated at several positions in the room will be created and played to the subjects through headphones. This setup will involve the subjects to have their head kept in the same direction through the whole experiment.

Binaural sound consists in audio with filtering corresponding to the natural filtering of shoulders, pinna and skull that occur in a natural listening condition. This filtering is variable depending on the position of the sound source (azimuth, elevation and distance from the listener).

Consequently, audio binaurally processed is made to be listened through headphones, to avoid such a filtering to occur twice. Binaural sound can be implemented through two different processes: it is possible to record sounds binaurally with the help of a mannequin with microphone placed in the ears. The materials used to manufacture such mannequin must reproduce the filtering characteristics of the corresponding parts of the human body.

Figure 29: Example of mannequin for binaural recording

It is also possible to compute binaural audio algorithmically. The above-described mannequins can be used to obtain a Head Related Impulse Response (or HRIR), representing the frequency filtering caused by the mannequin's body before the audio reaches the microphones, depending on the position of the sound source. This impulse response can be used to model a Head Related Transfer Function (HRTF), which can then be implemented in a filtering algorithm to process mono sounds.

Such a technique has the advantage of being flexible, as the algorithm can be changed according to the situations in which it has to be used. A restriction, in the context of the experiment to be designed in this research, is that the subjects cannot move their head during the experimental procedure (the possibility of a dynamic binaural algorithm adapting the filtering in real time according to the subject's position and head angle is rejected, because of technical limitations)

This part will describe an attempt of experimental design based on the binaural audio algorithm from Brown and Duda (1997).

FIRST EXPERIMENTAL SETUP

An important point in the experimental setup is that the subjects must feel that the audio events they will have to locate are taking place in the actual room and are coming from the objects placed in this room. Consequently, these sounds must be merged with the ambient audio background specific to the room. To achieve that, a stereo microphone will be placed over the subject. The binaural sounds will be mixed with the output of the microphone and fed into the headphones; a main task is now to make the quality of the binaural audio files homogeneous with the quality

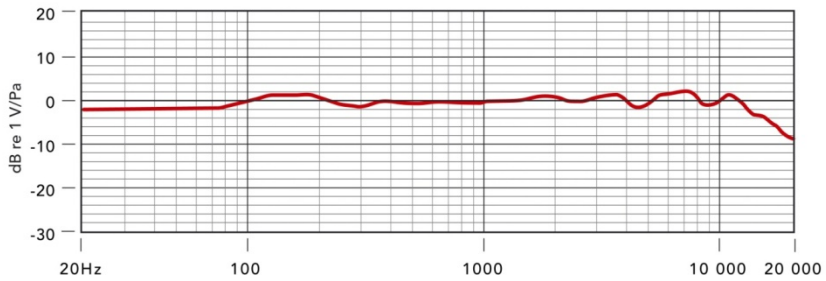
of the audio fed from the microphone. Factors such as the reverberation inherent to the experiment's room and the microphone's frequency response will have to be analyzed and applied on the binaural sounds to reach this mixed-reality state.

BINAURAL AUDIO

The creation of the experimental setup started with the creation of a program applying the algorithms from Brown and Duda (1997). This program, written in C++ and with the help of the Synthesis Toolkit (Stk) library for sound synthesis, takes as arguments a mono-channel audio file, the azimuth of the audio event, and two reverberation variables that are the T60 value (time for the reverberant waves to be dampened by 60dB) and the direct-to-reverberant ratio. A sound file in the wav format is generated from these values.

SOUND CAPTURE SETUP

The sound capture setup was then built in the experiment's room. The stereo microphone used is a Røde NT4, which has a relatively flat frequency response from 20Hz to 10 kHz, with a linear fall of 10dB between 10 kHz and 20 kHz.



Røde NT4 frequency response

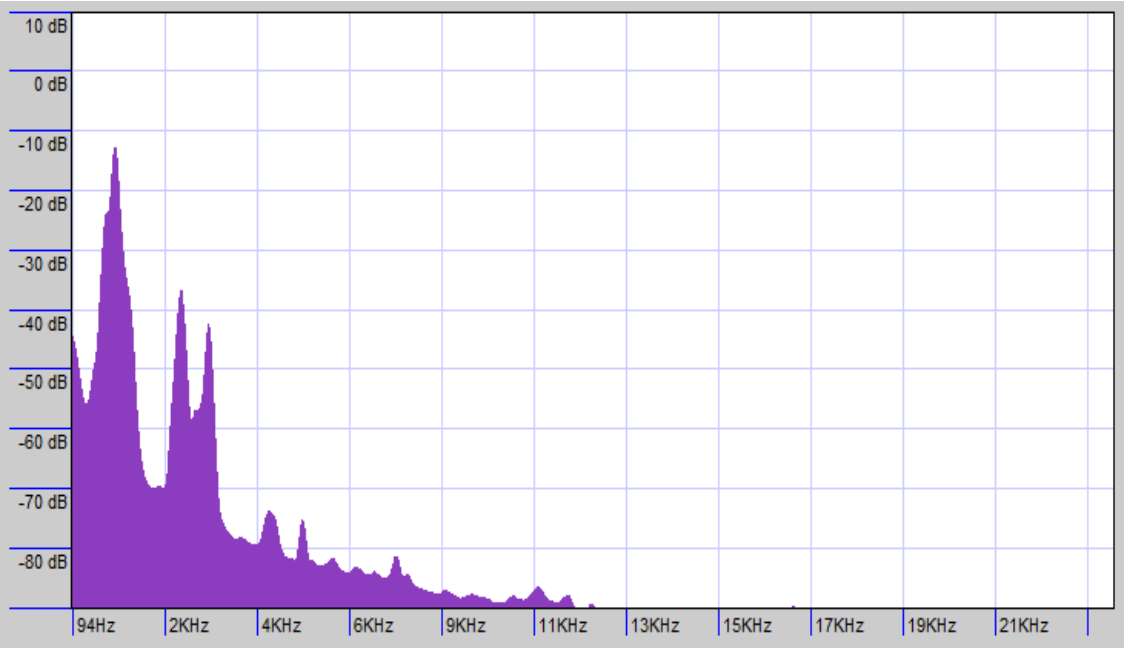
The room used for the experiment has a relatively important reverberation. Due to time limitations, it has been chosen to avoid doing precise room reverberation measurements. As an alternative, sound files with different reverberation values will be generated, and the best fit (in terms of subjective perception of the room reverberation) will be saved and applied to the rest of the audio to be played.

A choice had to be made concerning the reverberation algorithm to be used. The Stk library contains three of them, each having specificities. The choice was made to use the NRev reverberation algorithm. This choice was again based on subjective comparison with the room's natural reverberation.

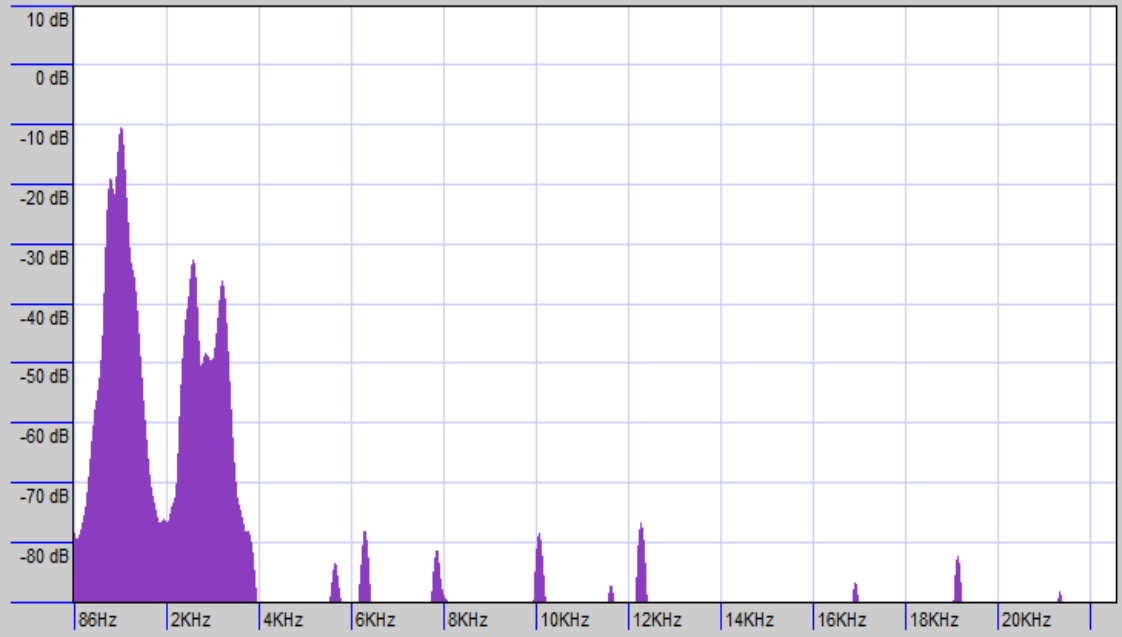
MIXED REALITY: ANALYSIS OF RECORDINGS AND FILTERING

Tests on the mixed-reality setup: the everyday objects used for this pilot test are electronic phones, a radio and noisy toys (laughing stuffed animals). The first analyses are done with the phone sound.

A spectral comparison was done between the phone sound computed by the binaural program with the reverberation corresponding to the experiment's room and an azimuth of 26 degrees and the same sound recorded by the experiment's microphone at the same azimuth and a distance of 2 meters. The sound obtained through the binaural program was then reduced by 18dB so the amplitude of the highest harmonic (at 1078Hz) would correspond to the one on the recorded sound.



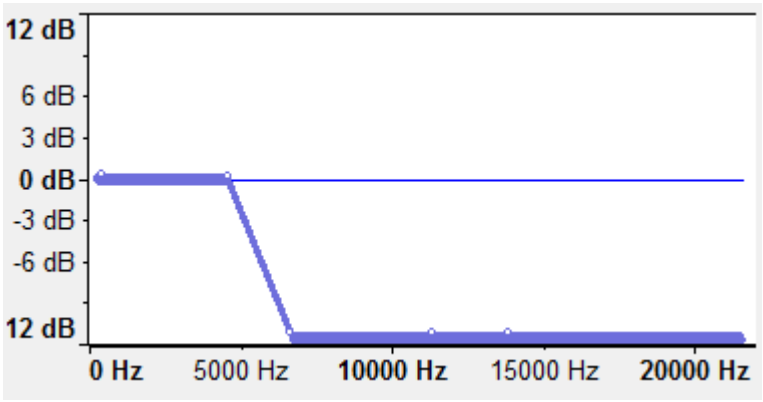
Phone sound spectrums recorded in the experiment's room



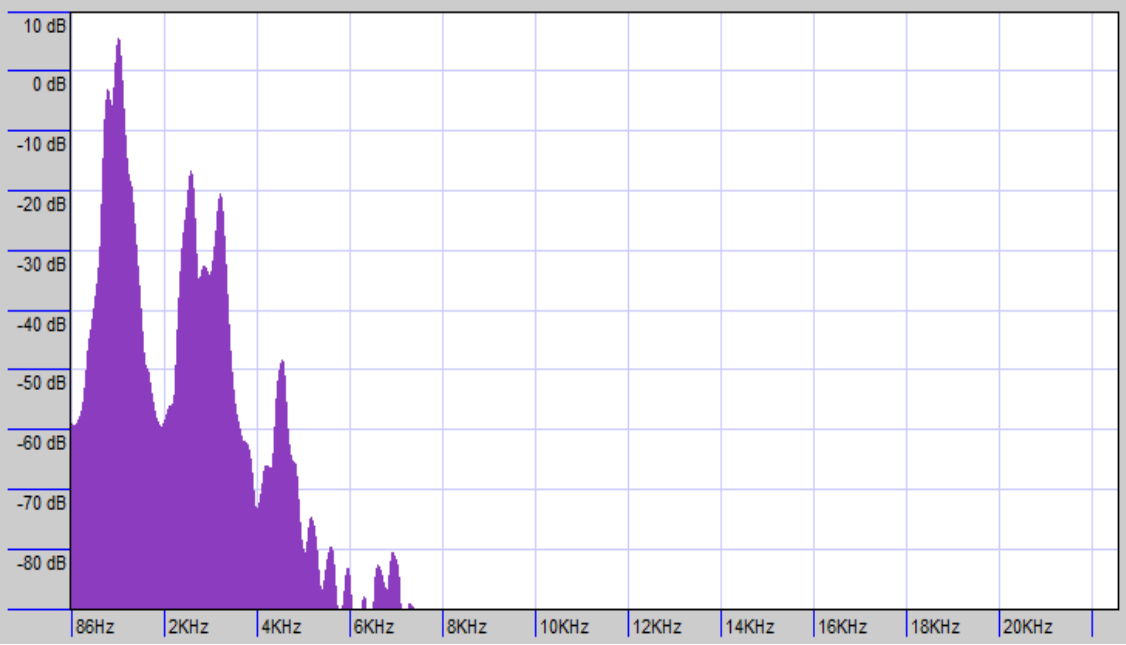
Phone sound spectrums generated by the binaural program

It appears that in the sound recorded by microphone in the room, the high frequencies are filtered, starting over 5 kHz (the harmonic at 5582Hz goes from -82dB on the binaural sound file to -72dB on the recorded sound). Over 5kHz, it appears that the phone harmonics are sunk in the ambient noise (with the exception of one harmonic at 7725Hz that is just dampened by 3dB; it will nevertheless be ignored).

To reach a sound quality close to the one of the recorded audio, the binaural sound is thus passed through a low pass filter (Audacity's FFT filter, with 0dB amplification from 0Hz to 5kHz, a logarithmic decrease of the amplification curve from 5kHz to 7kHz until reaching -12dB over 7kHz; this filter is applied four times to the binaural sound to reach the wanted sound quality).



FFT filter frequency response



Spectrum of the binaural sound file after being passed four times through the FFT filter

LOCALIZATION AND REALISM TEST

This test has two objectives: first, to verify if the subjects can still locate the binaural sounds after the application of reverberation; then, to see if these sounds are merged enough in the ambient sounds outputted by the microphone, so the subjects would consider them as audio events occurring in the experiment's room.

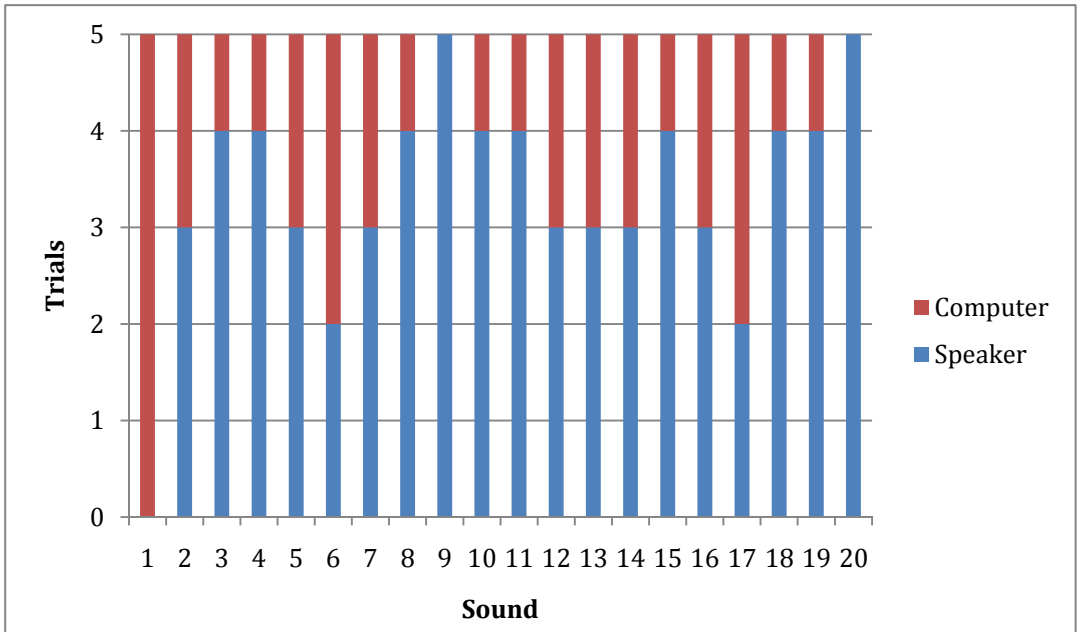
The brief localization test lead by Brown and Duda (1997) to test their model was, in the words of the authors, not providing a good measure of the accuracy of the localization effects from their model. Moreover, their experiment was purely auditory (no visual cues) and was based on a comparison between binaural sounds recorded with a dummy head and sounds obtained through their algorithm. In the case of the present work, the stimuli of interest are multimodal (auditory-visual), and a task of localization of an auditory stimulus should be made in function of several visual references. These visual references have presumably an influence of the localization tasks, which makes the experimental setup from Brown and Duda (1997) non-adapted for the present study.

The recording setup (room and microphone position) stays the same as for the previously described sound-analysis sessions. The ambient soundscape outputted by the Røde NT4 stereo microphone, placed over the subjects, is mixed with a playlist of binaural sounds to which reverberation has been applied. Three meters in front of the subject (on a circular perimeter), five speakers are placed, at azimuths -46, -23, 0

23 and 46 degrees (a range where a subject can see all of the speakers without having to turn his or her head). Each speaker is assigned a number for the further localization task. In this experiment, speakers are used to avoid as much as possible any semantic effect on the subject's answer. This is made under the assumption that subjects would consider that speakers can be a source for any type of sound. During the test, the subject is asked to keep his or her head straight, in direction of azimuth zero (one experimenter makes sure that this instruction is respected during the test).

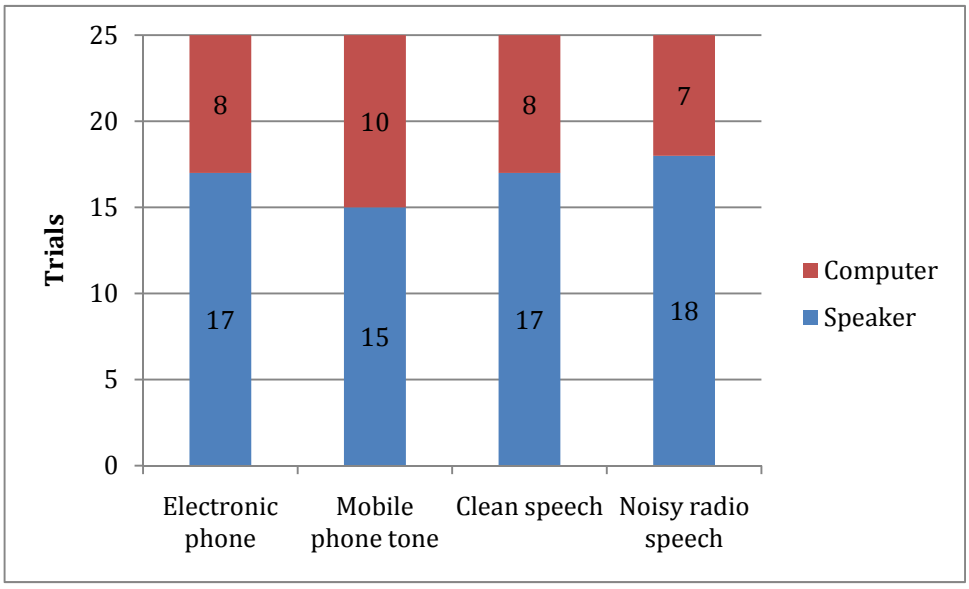
Four sounds were selected for this test: two relatively short ones being an electronic phone sound and a widely known mobile phone ringing tone (the "classical" Nokia tune). The two others are longer speech sounds (eleven and twenty seconds), one being relatively clean and the other one noisier (an old AM radio show). Each sound is passed through the binaural algorithm, with a direct-to-reverberant ratio of 2 and a T60 of one second. Five binaural sounds, corresponding to the five different azimuths of the speakers, are processed. The final experimental playlist contains thus twenty sounds arranged in random order.

During the test sessions, the subjects are introduced in the experiment's room and asked to sit on a chair three meters in front of the five speakers. They are told about the technical setup used for the experiment: the microphone will feed the headphones with the ambient sounds; some sounds will be played and they either can come from the speakers (caught by the microphone), or from a computer plugged directly in the mixer. The subjects are instructed that for each sound played, they are asked to tell if they perceived it as coming from the speakers or as added over the ambient mix. Independently of their first answer, they are then asked to tell form which direction they perceived that the sound was coming from, using the numbers assigned to each speaker. The subjects were told that if they perceived that the sound was coming from neither of the speakers, they could give a location in their own words. At the end of the test, the subjects were asked about their experience, the difficulty of the task and what was making them "know" that some sounds were played by the speaker and some were not.
Mixed-reality perception results:



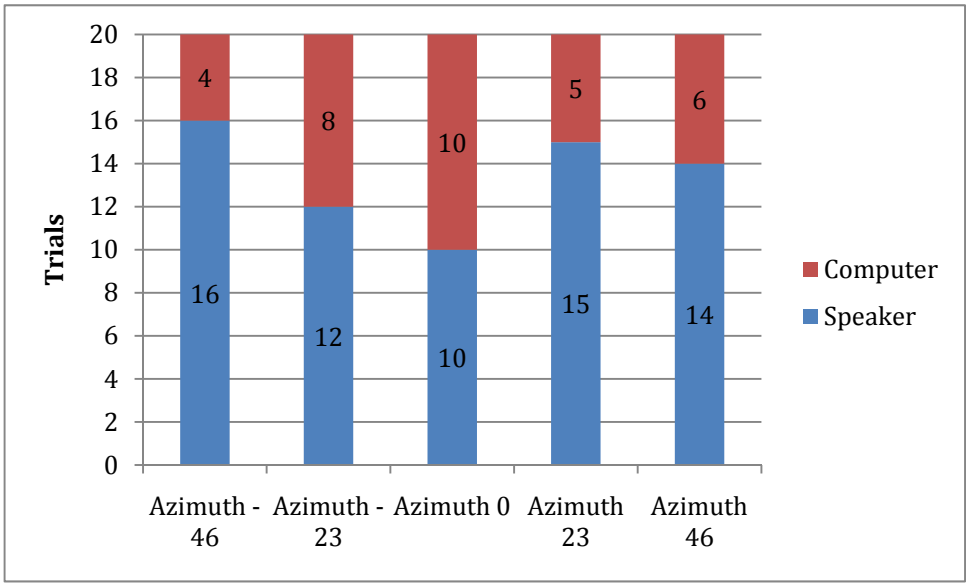
Perception of the 20 sounds as coming the speakers or not. A mean of 67% sounds (Standard deviation: 23) were perceived as coming from the speakers.

For each sound of the playlist, its perception as coming from the speaker or as having a virtual source is variating among subjects, as shown by the mean of 67% sounds perceived as speaker ones with a relatively important standard deviation of 23%. An exception comes with the first sound of the playlist that has been unanimously considered as having a virtual source. It should be added that concerning this sound, all the subjects perceived it as coming either from behind them or from above them.



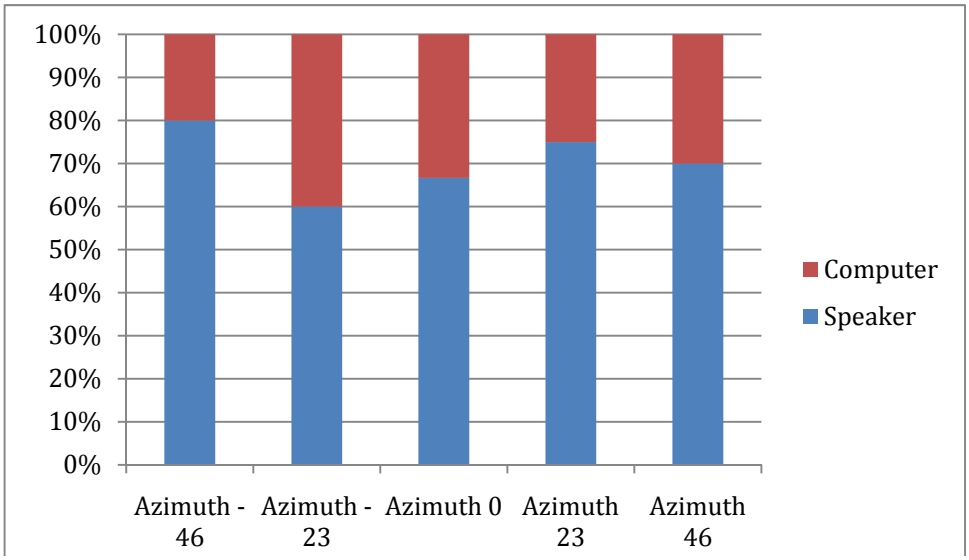
Perception of each of the five sounds as coming from the speakers or not
(independently of the azimuth of the virtual source)

An analysis of the results for each sound separately shows no major variation. This will be taken as a proof that the sounds' audio qualities were relatively uniform. It also confirms the assumption that using speakers permitted to get rid of extra semantic effects across different types of sounds.

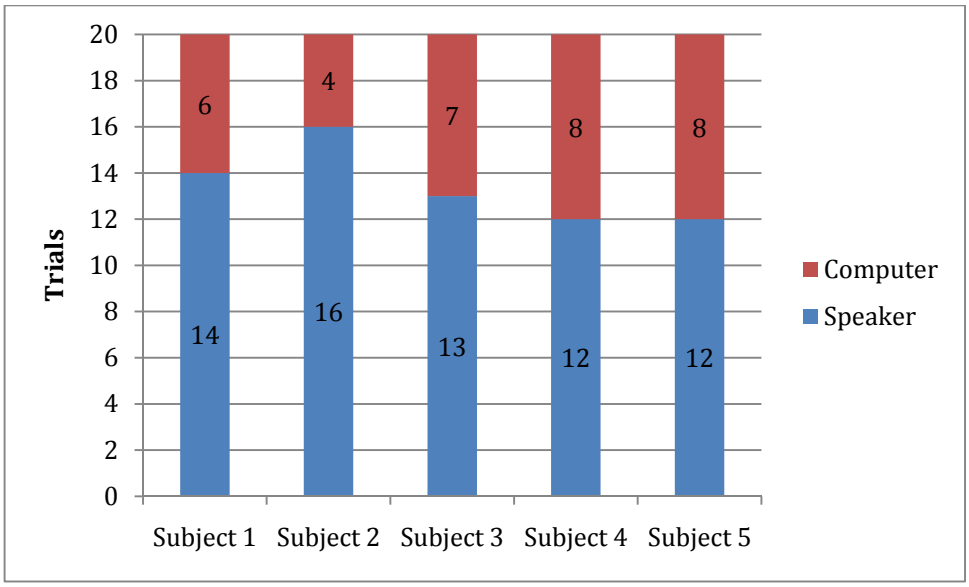


Perception of the sounds for each azimuth as coming from the speakers or not

An analysis of the perception of sounds as having a real (the speakers) or a virtual source permits to note a difference between the sounds with a virtual source at azimuth zero, and the ones with azimuths different than zero. The sounds computed with a virtual source at azimuth zero are indeed less believed as coming from the speakers. Nevertheless, it should be noted that at this azimuth, the mobile phone tone, which is the only one unanimously considered as not coming from the speakers, might have an important weight on the results. Analyzing the values a second time with ignoring the results on this sound and scaling the results in percentages shows no meaningful difference of perception according to azimuths.



Perception of the sounds for each azimuth as coming from the speakers or not (the mobile phone tune at azimuth zero is here ignored, and the results scaled in percentages)



Perception of sounds as coming from the speakers or not, in function of the subjects (*M* = 13.4; *SD* = 1.7)

An analysis of the perception of the sounds sources in function of the subjects does not show any major difference. The mean of 13,4 sounds considered as originating from the speaker, with a relatively low standard deviation of 1,7, permits to conclude that the skepticism of the subjects on whenever the sounds were coming from a real or a virtual source was relatively uniform.

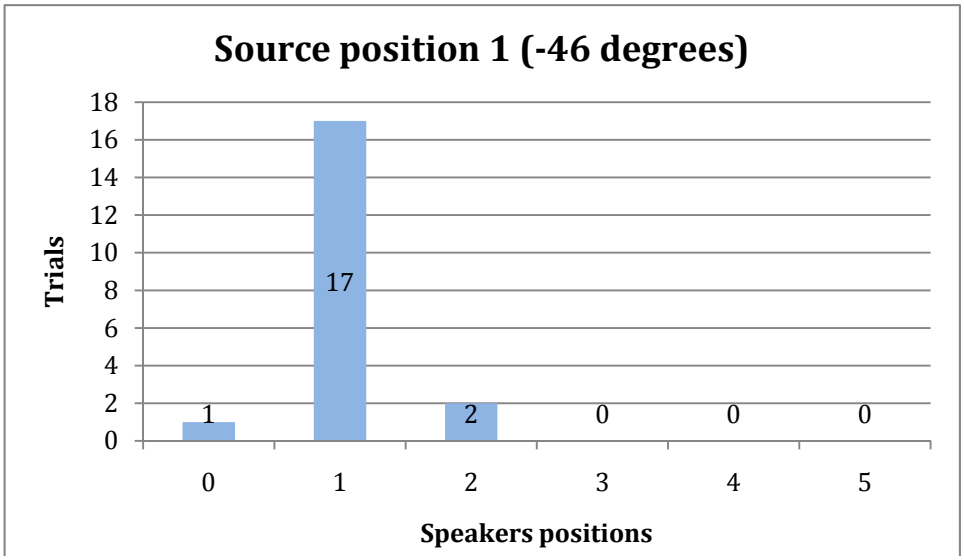
To conclude on the analysis of the results on the mixed-reality perception in this test, it seems that the perception of the played sounds as coming from a real source

(one of the speakers) or from a virtual source is highly variating. The amount of sounds considered as having a virtual source is relatively uniform across the subjects; however, there was no general agreement on which specific sounds were having virtual sources (with the notable difference of the first sound on the playlist, unanimously considered as having a virtual source).

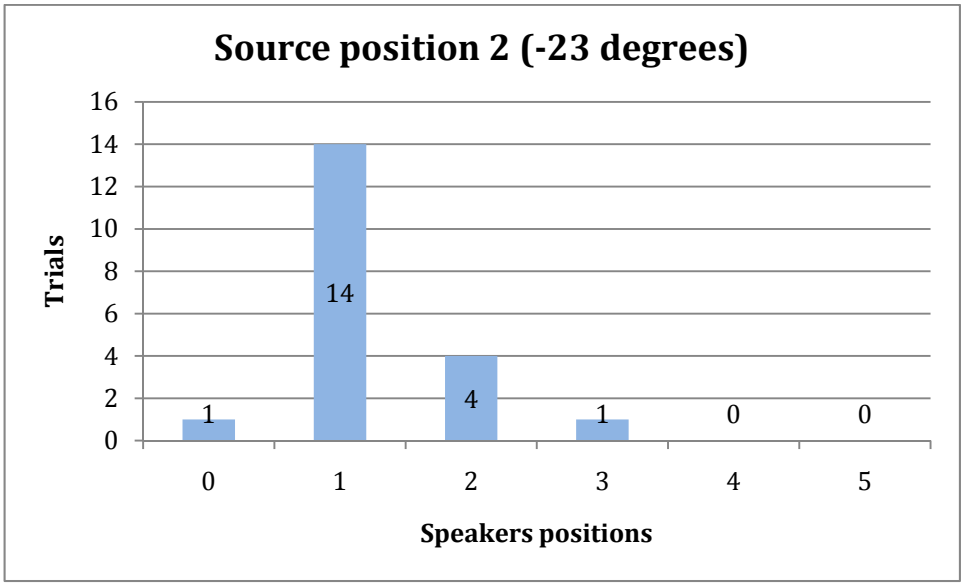
It should be underlined that all the subjects considered that discriminating speaker sounds and mixed-reality sounds was a difficult task. Some said that to decide on their answer, they were focusing on details such as the reverberation “quality” and comparing it with sounds that they were sure had a source in the room, such as the experimenters’ voices. It is then possible to assume that this reliance on small details in the sounds, and the hesitations the subjects showed while choosing their answers, are partly responsible of the lack of patterns observed in the quantitative results of the test. To gather more accurate results concerning whether the binaural sounds are merged in the sound background or not, two new test procedures could be valuable: first, the task of discriminating real and virtual sources could be replaced by another measurement method, or by a qualitative interview occurring at the end of the experiment. Alternatively, the same test as the one that has been lead could be redone, with a change consisting in giving an “extreme” reference of what would be a virtual source sound to the subjects, such as raw mono sounds directly fed in the headphones. These references, compared with the sounds that are known by the subjects as having a real source, such as the experimenters’ voices, could possibly “recalibrate” the subjects’ perception to avoid the reliance on small details in the sounds, and consequently could permit us to gather answers closer to the first impression of the subjects when they are exposed to the playlist sounds.

LOCALIZATION TEST RESULTS

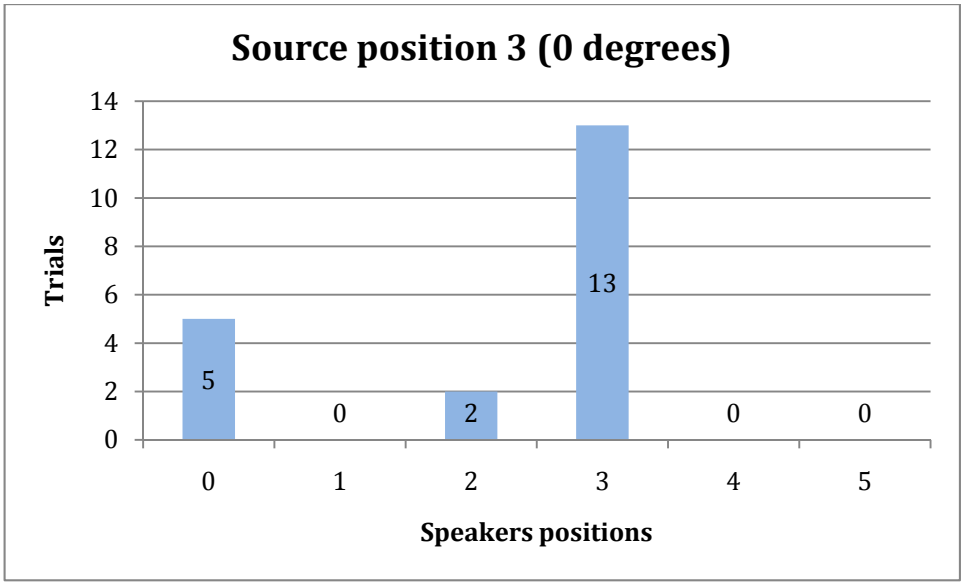
The part of the test concerning localization of sources gave twenty localization trials for each of the five speakers positions (five subjects times four different sounds). In the following graphs, the speaker position 0 corresponds to source localizations out of the scope of the five speakers (perception of the sources as coming from behind, above, etc.).



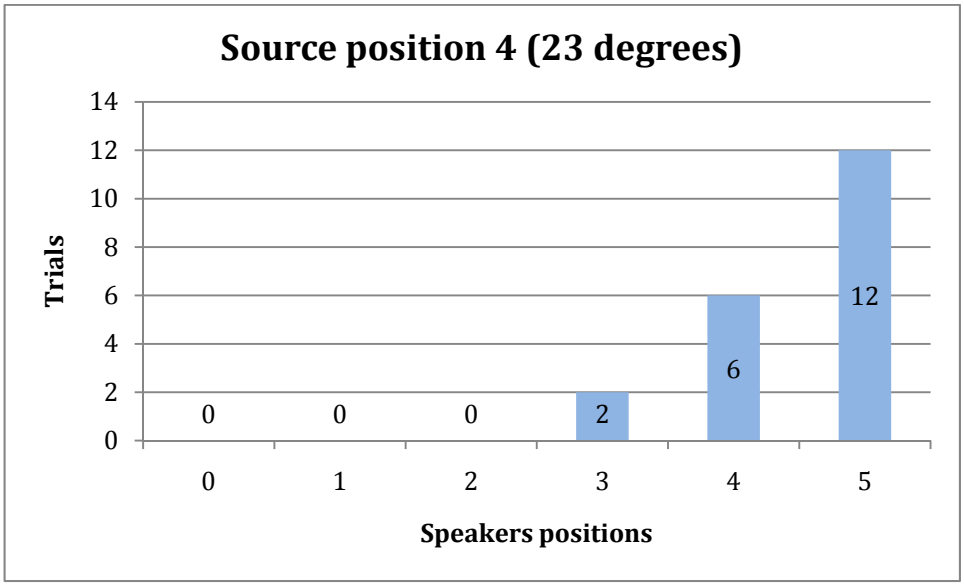
Localization of sources at azimuth -46 degrees (85% of right judgments)



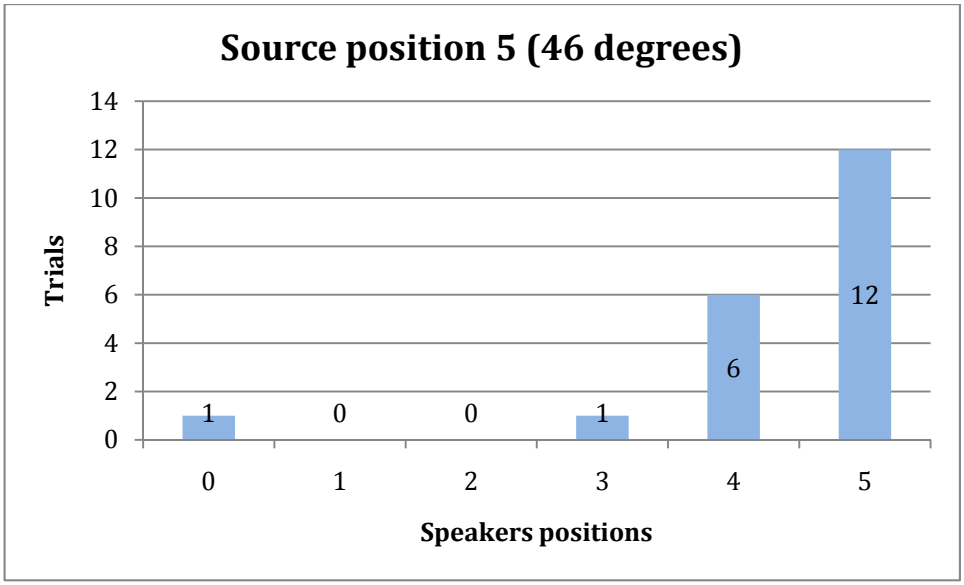
Localization of sources at azimuth -23 degrees (20% of accurate judgments)



Localization of sources at azimuth 0 degrees (65% of accurate judgments)



Localization of sources at azimuth 23 degrees (30% of accurate judgments)



Localization of sources at azimuth 46 degrees (60% of accurate judgments)

The results of the localization test are showing a difficulty for the subjects to accurately localize sources placed at intermediate positions: indeed, when the virtual sources were at azimuths 23 and -23 degrees, intermediate positions between the extreme left and right speakers (azimuths -46 and 46 degrees) and the center one, most of the subjects had a tendency to locate them at azimuth 46 and -46 degrees. Hence, it can be assumed that the subjects had a tendency to categorize the sources as “left”, “right” and “center”, and had difficulties to make intermediate judgments.

TEST CONCLUSIONS

On the topic of the mixed-reality impression emerging from the experimental setup, the results of this experiment give a beginning of insight on how the subjects are reacting during such an experience, and how it would be possible to gather reliable results, by means of new experimental procedures. But the most important point comes from the localization performances measured during this experiment: it is obvious that the binaural algorithm used do not allow an optimal accuracy in localization. Consequent to this finding, the choice has been made to compare the results gathered with Brown and Duda’s (1997) algorithm with results that will be obtained with a different method of creating spatial audio. The decision was taken to build a setup based on playing sound through speakers placed each position corresponding to the visual references to be used.