Toward sequential segregation of speech sounds based on spatial cues

Marion David

THÈSE

Pour l'obtention du grade de

DOCTEUR DE L'ÉCOLE NATIONALE DES TRAVAUX PUBLICS DE L'ÉTAT

Université de Lyon

ÉCOLE DOCTORALE MEGA - Mécanique, Énergique, Génie Civil et Acoustique

SPÉCIALITÉ : Acoustique

Présentée par

Marion David

Towards sequential segregation of speech sounds based on spatial cues

Soutenue le 13 Novembre 2014
Devant la Commission d'Examen

JURY

M. Daniel Pressnitzer	Directeur de Recherche, ENS Paris	Rapporteur
M. Steven van de Par	Pr., Université d'Oldenburg	Rapporteur
M. Andrew J. Oxenham	Pr., Université du Minnesota, Minneapolis	Examinateur
M. Jean-Christophe Valière	Pr., Université de Poitiers	Examinateur
Mme. Catherine Marquis-Favre	Chargée de Recherche HdR, ENTPE Lyon	Examinatrice
M. Nicolas Grimault	Chargé de Recherche HdR, CRNL Lyon	Directeur de thèse
M. Mathieu Lavandier	Chargé de Recherche à l'ENTPE, Lyon	Directeur de thèse

Acknowledgements - Remerciements

Avant tout autre chose, je souhaiterais remercier celui qui m'a mise sur la "voix" de la psychoacoustique. Après quelques années d'errance intellectuelle, et à deux doigts d'être happée par une turbine aérodynamique, Jeremy Marozeau m'a ouvert les yeux sur les oreilles. Voilà le début de l'histoire, Miles Davis en fond musical, un projet sur les implants cochléaires, l'Australie en hiver et une thèse à la clé.

Par ailleurs, cette thèse n'aurait pu voir le jour sans le soutien financier du LabEX CeLyA. Je les remercie d'avoir cru en notre projet qui arrive maintenant à son terme. CeLyA m'a également soutenue dans mon échange avec l'Université du Minnesota, une expérience unique et profondément enrichissante.

Je tiens également à remercier chaleureusement mes directeurs de thèse, Mathieu Lavandier et Nicolas Grimault. Merci pour leur confiance, leur patience et leur ténacité. Au delà de leurs qualités scientifiques reconnues, ce sont des personnes ouvertes, humaines et passionnées. Nos discussions, toujours ponctuées de digressions plus ou moins longues, nous ont conduit à ce travail de thèse. J'ai énormément appris auprès d'eux, c'est pourquoi je leur adresse aujourd'hui ces remerciements. Je remercie également les deux équipes avec lesquelles j'ai travaillées, le LGCB et CAP.

Let's switch in English to thank now my colleagues and friends from Minnesota. First, I would like to thank Andrew Oxenham for his invitation in his lab. He trusted me since the very beginning that gave me self confidence to carry on my research project there. This project has grown from experiments to experiments and I am really glad to continue this collaboration in the future. Beyond this scientific collaboration, I would like to acknowledge Andrew's human qualities and those of his team who helped me a lot during my first winter in the cold Minnesota. Thanks to Dorea, Kyle, Gene, Jordan, Shaye, Nyinguan, Emily and Magda. I am particularly

grateful to Kelly and Jackson who accepted to give me their voices for one of my studies. I would also like to thank all the lovely Minnesotan people that I met and who became real friends: Dan and all the Andersen family, Mary, Vina, Aloida and Jesse, Jessus, Helen and Jim, Alex, Shin, Chris, Travis, and Eric. It was a pleasure to spend time with all of them, I can't wait to see them at the end of the falls...

Un dernier et grand merci à "tous les autres", c'est à dire à tous ceux qui m'ont soutenue, supportée dans mes hauts et mes bas, qui m'ont remise dans le droit chemin malgré tout. Je les remercie sincèrement, car sans eux, cette thèse n'aurait probablement jamais aboutie, c'est pourquoi je tiens à leur dédier ces quelques lignes. J'ai la chance d'avoir de vrais amis, de ceux que je connais depuis toujours à ceux rencontrés plus récemment, chacun d'entre eux m'a apportée un petit (ou grand) quelque chose qui m'a permis de me construire jusqu'à aujourd'hui. Alors merci à vous mes collègues et amis, Thierry, Achim, Thibaud, Arnaud, Riccardo, Guillaume, Tatiana et Maïté. Merci à vous les "anciens", Maxime, Pierre, Greg, Seby, Jo et Fayçal. Merci à vous les "nouveaux", Marie, Chouchou, You, Aurel, Melody et Laurent. Sans tomber dans un discours guimauve et inutilement émouvant, je tiens à remercier les membres de ma famille, ceux qui ont toujours cru en moi et su me changer les idées avec maestria dans n'importe quelle circonstance. Et enfin, un merci particulier pour une personne particulière, Laurence Tregnier pour son soutien depuis ces x dernières années.

Contents

List of Figures

List of Tables

Introduction

In most of everyday acoustic environments, the sounds reaching the ears result from multiple sources. As a parallel with a visual scene, an auditory scene consists of these sound events perceived by the auditory system. Albert Bregman, one of the pioneers of the auditory scene analysis, used a visual analogy to explain the aim of solving an auditory scene. This analogy considers a lake, which corresponds to the acoustic environment. Boats on the lake produce waves as sources emit sound waves. Two channels are dug at the water's edge and a piece of cloth is placed over these channels following the motions of the waves. The channels symbolize the auditory canals, and the cloth the ear drums. Based on this analogy, solving an auditory scene would consist of figuring out what is happening on the lake by looking at the motion of the cloth.

An auditory scene appears to be a theoretically complex problem to solve. However, we are exposed to these types of situation almost every days, and we are not so bad at dealing with multiple sound source environments. For example, let us consider a situation - namely a cocktail-party - where the listener tries to understand the message delivered by a target speaker while several other speakers talk at the same time. Even though the listener will be impaired by the noisy background, he/she will be able - with more or less success - to separate the target from the competing voices and understand the message.

When a sequence of sounds reaches the listener's ears, he/she can perceptually organize this auditory stimulation in one or several streams, usually corresponding to one or several sound sources. Back to the cocktail-party situation, the sounds coming from the target speaker and from the competing voices have to be separated into distinct streams for the target to be intelligible. The major question raised here is how? How our auditory system can organize

the coming sounds? Which acoustical cues are relevant to perceptually separate the different sound sources?

The formation of distinct streams depends on several acoustic parameters which differ between the perceived sounds. Many studies so far have investigated parameters such as spectral differences, temporal differences or differences in localization. The main limit that can be underlined in these studies would be the artificial nature of the situations tested. The present work proposes a more realistic approach of the auditory streaming field.

Reverberation in rooms and the human head induces temporal and spectral distortions of sounds. Several parameters are covered by this general phenomenon: monaural spectral differences, binaural differences and differences in lateralization. The aim of this thesis was to investigate, by mean of behavioural listening tests, to what extent these distortions could influence the perceptual organization of sequential speech sounds. Two distinct parts are developed in the present thesis. The first part investigated the influence of head and reverberation on the segregation of broadband noises (as a first approximation of speech), and the second part investigated how the results of the first part could be applied to real speech. Since running speech present a high degree of acoustical variability, preliminary experiments had been conducted. First, the robustness of stream segregation based on a frequency difference to variability on the same acoustic dimension (i.e., frequency) was assessed using pure tones, and second, the fundamental frequency difference required to segregate speech items was evaluated.

Dans la plupart des environnements sonores quotidiens, les sons qui parviennent aux oreilles résultent de plusieurs sources. Pour faire un parallèle avec une scène visuelle, une scène auditive est constituée de tous les évènements sonores perçus par le système auditif. Albert Bregman, qui est l'un des pionniers dans le domaine de l'analyse de scènes auditives, avait l'habitude d'utiliser une analogie visuelle pour décrire l'objectif de résoudre une scène auditive. Dans cette analogie, un lac symbolise l'environnement acoustique. Des bateaux qui naviguent sur le lac produisent des vagues de la même manière que des sources émettent des ondes acoustiques. Deux canaux sont creusés au bord de l'eau et un bout de tissu est déposé sur ces canaux, leurs déplacements suivant le mouvement des vagues. Ces deux canaux représentent les canaux auditifs, et le tissu les tympans. Résoudre une scène auditive consisterait, suivant cette analogie, à comprendre ce qu'il se passe sur le lac simplement en analysant le mouvement des tissus.

Une scène auditive semble être un problème théorique difficile à résoudre. Cependant, nous sommes exposés à ce type de situations presque tous les jours, et nous arrivons plutôt bien à comprendre un environnement constitué de plusieurs sources sonores. Par exemple, dans un contexte de cocktail-party, un auditeur tache de comprendre le message délivré par un locuteur donné alors que d'autre personnes parlent en même temps. Même si l'auditeur sera gêné par l'environnement bruyant, il sera capable - avec plus ou moins de réussite - de séparer la voix cible des voix concurrentes et de comprendre le message.

Lorsqu'une séquence de sons atteint les oreilles de l'auditeur, celui-ci peut organiser perceptivement cette stimulation auditive en un ou plusieurs flux, généralement correspondant à une ou plusieurs sources sonores. Dans un contexte de cocktail-party, les sons provenant du locuteur cible et les voix concurrentes doivent être séparées de manière à ce que la cible soit intelligible. La question principale est comment ? Comment notre système auditif est-il capable d'organiser ces sons ? Quels sont les paramètres acoustiques pertinents pour séparer les sources concurrentes ?

La formation de flux sonores distincts dépend de nombreux paramètres acoustiques qui diffèrent entre les sons perçus. Plusieurs études ont montré l'influence de certains paramètres comme les différences spectrales, les différences temporelles ou encore les différences de positions dans l'espace. La principale limitation de ces études est leur caractère artificiel, notamment concernant les configurations testées. Le travail présenté ici propose une approche plus réaliste du domaine de l'analyse des scènes auditives.

La réverbération dans les salles, ainsi que la tête de l'auditeur induisent des distorsions temporelle et spectrales des sons. Ce phénomène général regroupe plusieurs paramètres: des

différences spectrales monaurales, des différences binaurales et des différences de latéralisation. L'objectif de cette thèse a été d'étudier, au moyen d'expériences comportementales, dans quelle mesure ces distorsions peuvent influencer l'organisation perceptive de séquences de signaux de parole. Deux parties distinctes sont développées dans cette dissertation. D'abord, nous avons étudié l'influence de la coloration induite par la tête et la salle sur la ségrégation de bruits large-bandes (pris comme première approximation de la parole). Dans un second temps, nous avons étudié comment ces premiers résultats pouvaient être appliqués à des signaux de parole. Puisque la parole présente une forte variabilité de ses paramètres acoustiques, des expériences préliminaires ont dû être menées. Tout d'abord, la robustesse de la ségrégation de sons purs basée sur une différence de fréquence à une variabilité sur la même dimension acoustique (i.e., la fréquence) a été testée. Puis la différence de fréquence fondamentale requise pour séparer des signaux de parole a été évaluée.

Auditory scene analysis: from the Gestalt Theory towards the perceptual organization of realistic auditory stimulations

1 The Gestalt theory

1.1 Origins and properties

The Gestalt theory, or shape theory, is based on the idea that the human brain considers the sensitive external stimulations as organized and structured entities instead of a sum of constituents. The theory was formulated by the psychologists of the Berlin school in the mid 20-30s. The principles were introduced by Wertheimer and developed later by Köhler, Koffka and Metzger. Applied first to the visual field, the Gestalt principles were then used in auditory perception.

The Gestalt theory counts four main properties which characterize perception. First, the **emergence property** is the process by which a shape can appear from separated unidentified elements. The shape is only recognized when the constituents are perceived as a whole. In the vision field, a common example is the picture of a dalmatian dog sniffing the ground (see Figure 1, panel a.). Second, the **invariance property** allows recognizing simple shapes even if they are rotated, translated or scaled. For example, the objects depicted in panel b. of Figure 1 are immediately recognized as the same shape. Third, the **multi-stability property** illustrates the tendency for the perception to switch back and forth between two interpretations. Note that the two alternatives can not be perceived at once. Two examples are shown in panel

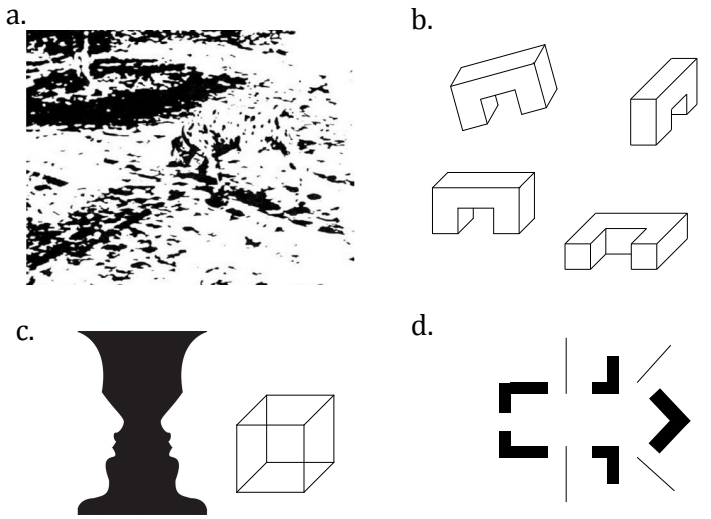

Figure 1 – The four properties of the Gestalt theory: emergence (the famous Gestalt dalmatian (expired copyright) a.), invariance (b.), multi-stability (the illusion of the vase from Edgar Ruben and the Necker's cube, c.) and reification (an arrow inspired by the Kanizsa's fictional figures, d.).

c. of Figure 1: the Necker cube which can be perceived either outgoing or ingoing, and the Rubin's illusion which can be perceived either as a vase or as two facing faces. And fourth, the **reification property** that characterizes the generative aspect of perception. For instance, a shape can be perceived even if it is not presented, based on the context. Panel d. of Figure 1 shows this property: an arrow is recognized even though it is not drawn.

1.2 Laws of form perception

Figure 2 illustrates the six main Gestalt's laws, based on which a visual scene is organized.

According to the **law of proximity**, the elements which are close in space tend to be grouped in a same object. For example, the three lines of panel a. consist of the same six elements, but they can be organized differently if their positions differ. The elements of the first line are all equidistant and thus are all grouped in the same coherent object. In the second line, three pairs of two elements are recognized, so three distinct objects appear. In the third line, the first and the last elements appear as two distinct objects, and two other objects consisting of two elements are recognized in the middle.

The **law of similarity** states that elements are grouped together if they are similar to each other. The similarity can concern any element's characteristic, such as colour, shape... For example, the elements of lines 1, 3 and 5 in panel b. tend to be grouped together as they share

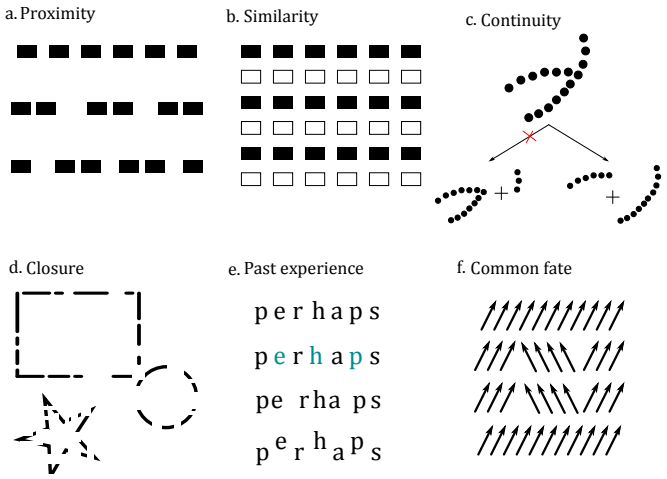

Figure 2 – Illustration of the six main laws of form perception of the Gestalt theory. a. the law of proximity states that elements which are close in space tend to be integrated. b. the law of similarity states that elements tend to be integrated if they are similar to each other. c. the law of continuity states that elements tend to be integrated if they are aligned together. d. the law of closure states that elements can be perceived as a whole even if they are incomplete. e. the law of past experience states that elements can be grouped together according to the past experience of the observer. f. the law of common fate states that elements tend to be integrated if they move in the same direction.

a similar colour. In the same way, the elements of lines 2, 4 and 6 form a second object.

In panel c., one would expect the elements on the top to be organized as shown in the bottom left, however, they are grouped as shown in the bottom right. This organization follow the **law of continuity**. Indeed, elements tend to be integrated into perceptual objects if they are aligned with each other.

Based on the **law of closure**, objects are perceived as a whole even if they are not complete. For example, on panel d., the three objects are clearly recognized (i.e., a rectangle, a circle and a star) even though they are incomplete.

Panel e. illustrates the **law of past experience**. In some cases, the elements of a visual scene are integrated according to the past experience of the observer. For example a pattern of letters, like the one presented in panel e., tends to be grouped as a meaningful word even if other laws are violated.

The **law of common fate** states that visual elements tend to be perceived as a whole if they have the same motion. For example, the eight arrows presented in the middle of panel f. are integrated as a coherent object because they moved in the same direction and all the other arrows form a second distinct object since they moved in another direction.

1.3 From visual perception to auditory perception

The same issues of organization of a visual scene (grouping, separation) were considered in the auditory field. Thus, the principles of the Gestalt theory were transposed, and a parallel was made between a visual object and an auditory stream. The next part of this dissertation focuses on the perceptual organization of auditory streams and details some major studies investigating the analysis of auditory scenes.

2 Auditory scene analysis

2.1 Segregation, integration and bi-stability of auditory streams

In an everyday sound environment, multiple sound sources produce sounds in every directions, this is what is commonly named an *Auditory scene*. Bregman (1990) clearly explained the aim of solving such an auditory scene: "The job of perception is to take the sensory input and to derive a useful representation of reality from it". To do so, the auditory system has to perceptually organize the sounds by grouping some sound events together and segregating others. Thus, the notion of auditory stream refers to a percept which corresponds to a coherent event of sounds. An auditory stream corresponds to the notion of object in the vision field. Besides, the perceptual organization is made by determining which sounds events *come together* and which do not. On one hand, the notion of integration (also named fusion or coherence) refers to situations where the perceived sounds are grouped together, as if they came from a single source. On the other hand, the notion of segregation (also named fission) refers to situations where the perceived sounds are separated, as if they came from different sources. The term *streaming* corresponds to the processes that determine whether one or more streams are heard.

In musical pieces, the percepts of integration and segregation are widely used. Indeed, a melody can be defined as a coherent and logical succession of complex tones. In this particular context, the coherence (i.e., integration) can rely on the timing (rhythm) or on the pitches (melo). The percept of segregation is also commonly used suggesting several melodic voices. In the context of Western musical tradition, these types of percepts are exploited, especially in polyphonies to create strong melodic coherence or to give more tension. Johann Sebastian Bach was one of the most famous Baroque composers and he used this notion of integration / segregation in many of his pieces. Figure 3 illustrates the notion of several streams in a prelude of Bach.

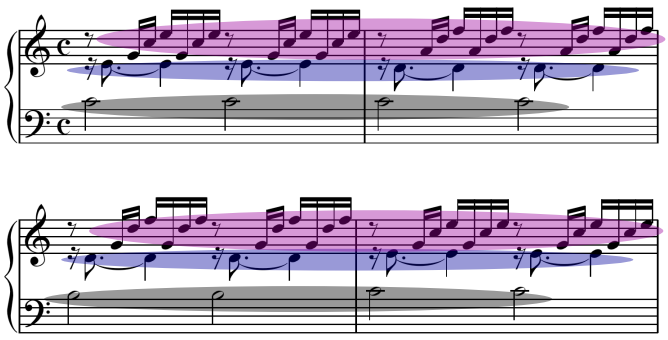

Figure 3 – Excerpt of a piece of Bach: Prelude I, Book I, Well-Tempered Clavier. The three different colours were added to illustrate the three different voices present in the melody.

Auditory streaming depends on many properties that will be reviewed later in this dissertation. The main idea, evoked by the Gestalt principle of similarity, is that when two stimuli share many of their acoustic features, they are more likely to come from a same source and thus they tend to be grouped together. Conversely, when two stimuli share just a few or no acoustic feature, they are more likely to come from different sources and thus tend to be segregated. For example, when two stimuli present a large difference along any acoustic dimension, for example a frequency difference between two tones, the sequence made of the repetition of these two tones [A-B-A-B...] will clearly split into two separate and coherent streams. However, when the difference is small, the sequence will be heard as a single event of coherent sounds. In between this two extreme percepts (i.e., clearly segregated or clearly integrated), for intermediate values of the observed parameter, the percept can flip from segregation to integration. This is called the ambiguity region, or the bi-stability percept (Pressnitzer and Hupé, 2006; Schwartz *et al.*, 2012; van Noorden, 1975). Figure 4 illustrates these three different percepts (integration, segregation and bi-stability) across time depending on a given acoustic property.

van Noorden (1975) introduced two important notions, characterizing the percepts of integration and segregation. First, the **temporal coherence boundary (TCB)** represents the critical value of the considered acoustical parameter above which listeners are no longer able to hear coherence. Second, the **fission boundary (FB)** is the critical value under which listeners are no longer able to hear segregation.

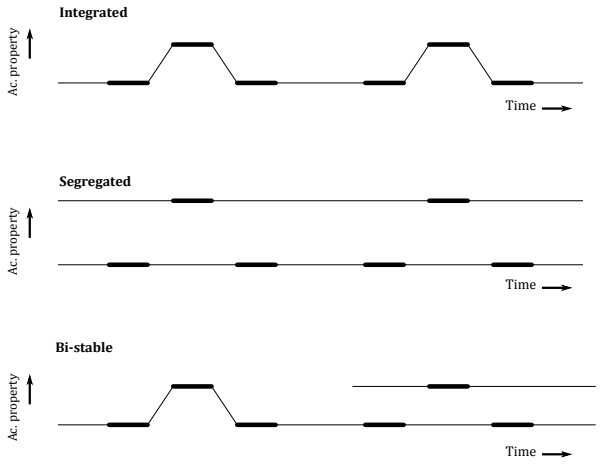

Figure 4 – Illustration of three different levels of segregation. The top panel represents a sequence of sounds heard in a single coherent stream. The middle panel represents the same sequence split into two separate streams. The bottom panel represents the same sequence, integrated at the beginning and then split into two segregated streams, so a bi-stable percept.

2.2 Sequential vs simultaneous mechanisms

When two or more sounds occur at once, some parts of these signals overlap in time and other parts alternate in time. The auditory system has to deal with these two types of mechanisms, simultaneous and sequential, in order to make sense of the mixture of sounds (Summers and Leek, 1998). Figure 5, from Étienne Gaudrain, illustrates the two types of mechanisms. On a waveform of two voices (1 female, 1 male), some time-windows contain the two voices which energetically overlap (dark blue parts) and some other time-windows contain only a single voice (magenta or light blue). This figure shows that both mechanisms occur in real environments.

The scope of this dissertation was limited to sequential mechanisms.

2.3 Auditory data processing

2.3.1 Stimuli driven vs schema driven auditory scene analysis

It is worth wondering how the stimuli are processed by the auditory system to perceptually organize the sounds. This question is still an open issue on its own. Bregman (1990) has proposed two different schemes to explain the listener's ability to solve an auditory scene. The first scheme is ascending, meaning that the auditory system is able to extract relevant information from the acoustical nature of the stimuli. This type of process is driven by the

Figure 5 – Waveform of two voices, one male (light blue) and one female (magenta). On some time-windows the two voices overlap (dark blue parts) and on some other time-windows the two voices alternate without overlapping. This figure shows that solving this auditory scene implies to deal with both simultaneous and sequential mechanisms.

stimuli itself and is referred to as primitive or bottom-up process. The second scheme is descending, meaning that the stream organization is governed by learnt expectations (i.e., knowledge of familiar sounds or patterns, acquired concepts or grammar...). Figure 6, taken from Plack (2005), illustrates the two schemes used to draw a useful interpretation of the auditory scene.

2.3.2 Obligatory vs voluntary streaming

Auditory scene analysis is governed by both voluntary and obligatory mechanisms, depending on the task the listener will have to complete. Voluntary streaming refers to tasks where the listeners try to hear out a target sound from a mixture; whereas obligatory or irrepressible streaming refers to situations where the listeners are biased towards grouping and fail to group (Bregman, 1990). The different types of task measuring streaming, and thus which mechanism is involved (voluntary or obligatory), will be detailed in the next session. Obligatory stream segregation can be associated with the TCB and voluntary stream segregation with the FB.

Both types of streaming were investigating in this PhD thesis.

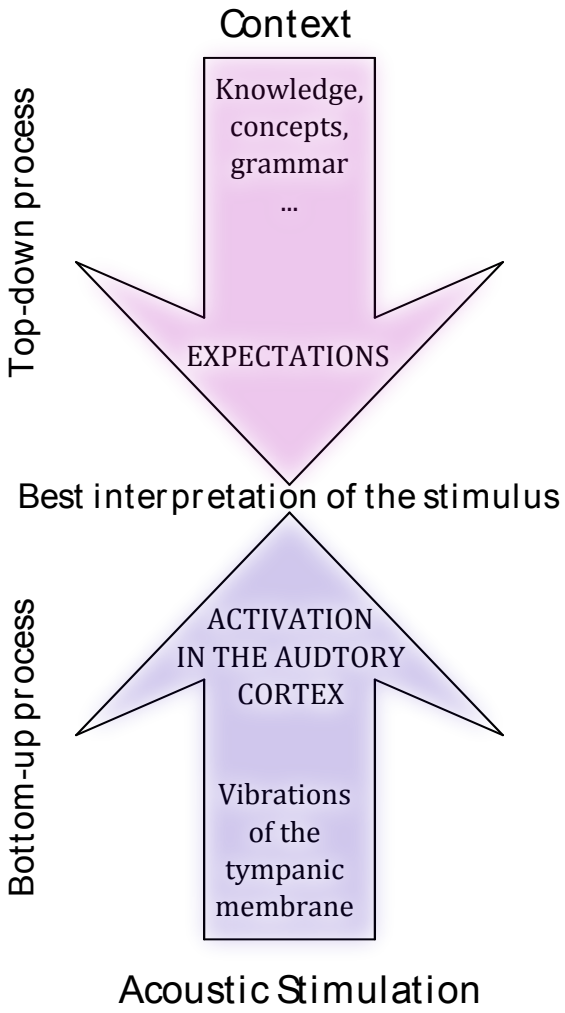

Figure 6 – Illustration of the two schemes used by the auditory system to draw the best interpretation of the coming stimulus, taken from Plack (2005). The ascending path is named bottom-up or primitive process and is based on the acoustical nature of the stimulus itself. The descending path is named top-down process and is based on learnt expectations.

3 Auditory stream formation

3.1 Parameters influencing auditory stream organization

3.1.1 Spectral differences

Role of frequency

Auditory streaming depends on the frequency difference between successive pure or complex tones. Large frequency differences (equal to or greater than 4 semitones) tend to produce segregation while small frequency differences tend to produce integration (van Noorden, 1975). Closer to real stimuli, Gaudrain *et al.* (2007) investigated the influence of a difference in fundamental frequency (F0) on the perceptual segregation of vowel sequences. They used a temporal-order procedure (see next session for more details) where listeners had to report the order of the vowels constituting the sequences. The results, shown in Figure 7, indicate that F0 contributes to segregation. Especially, they found that the listeners managed to integrate the streams when the F0 difference was small (less than 20 Hz), but segregation increased and thus scores decreased as the F0 difference increased.

The second study presented in Chapter 2 focused on the F0 difference required to separate speech items consisting of a fricative consonant and a vowel.

Peripheral channeling

It has been proposed that auditory stream segregation primarily depends on the filtering that occurs in the cochlea. In fact, the peripheral auditory system can be seen as a bank of bandpass filters owing to the tonotopic behaviour of the cochlea. The coming sounds are analyzed by frequencies and thus bandpass filtered (Moore and Glasberg, 1983). Hartmann and Johnson (1991) defined two different peripheral channels: "those based on frequency (tonotopic) and those based on ear presentation (lateral)". They ran a series of experiments to investigate to what extent listeners can form distinct streams from tones that excite different peripheral channels. The results showed a strong influence of peripheral channeling on stream formation. Thus, when the stimuli excite overlapping peripheral channels, tonotopic or lateral, they are mostly heard as integrated.

The excitation pattern is the spectral representation of the auditory nerve response to a stimulation at any time (Moore and Glasberg, 1983). Thus, the excitation pattern represents the magnitude of the output of the auditory filters at any time. The results of Hartmann and

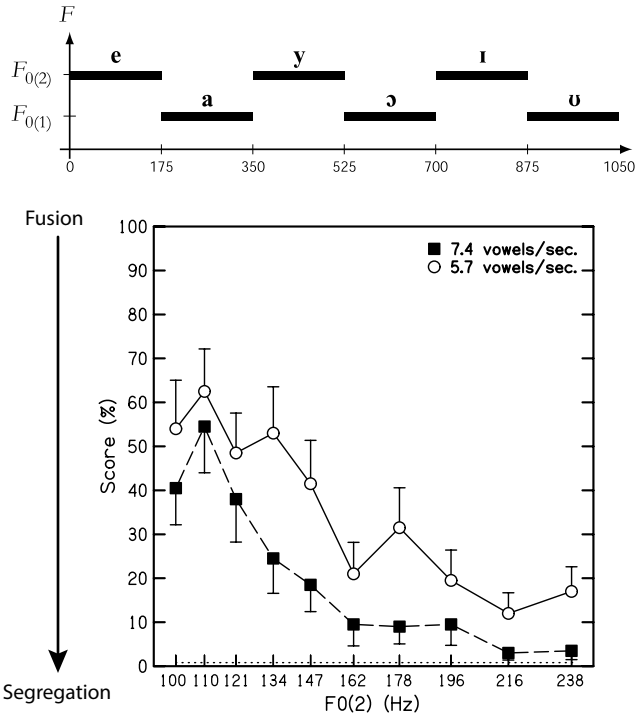

Figure 7 – Results of Experiment 1 of Gaudrain *et al.* (2007). The listeners were presented with sequences of vowels and were asked to report the order of the vowels across the streams (in this example, they had to report e-a-y-o-I-u). The F0 of the first stream ($F0_1$) was held constant to 100 Hz and the F0 of the second stream ($F0_2$) was set x Hz above $F0_1$. The scores are expressed in percent of correct answers as a function of the fundamental frequency difference x (i.e., $F0_1$-$F0_2$). The listeners obtained higher score if they were able to fuse the streams. These results show the influence of F0 on stream segregation.

Johnson (1991) showed that stimuli presenting a large degree of overlap of their excitation patterns tend to produce integration while stimuli that present a small degree of overlap tend to produce segregation. Some computer models of auditory scene analysis are based on this idea that streaming depends on the overlap of the excitation patterns Beauvois and Meddis (1996); McCabe and Denham (1997).

The peripheral channeling introduced by Hartmann and Johnson (1991) traditionally referred to sounds that stimulated different auditory channels, so it corresponded to large spectral differences. The first study detailed in Chapter 1 of this dissertation investigated the influence of slight spectral differences, that occurs within a single auditory channel, on stream segregation.

3.1.2 Temporal differences

Tone repetition time

So far in this manuscript, stream segregation has been defined for sequences of sounds, but the rate of presentation was not mentioned. It is worth noting that if the stimuli are alternated too slowly, no segregation can be heard even with a large difference (of any type) between the streams. In other words, the auditory system is able to support large acoustical variations as long as the tempo is slow enough. This characteristic can be associated with the property of similarity in the Gestalt theory.

Several parameters are used to characterize the rate of a sequence. The tone repetition time (TRT) corresponds to the time duration between the onset of a stimulus and the onset of the following stimulus. The TRT is often named stimulus onset asynchrony (SOA) and can be defined within a single stream or across the streams. The inter stimulus interval (ISI) also characterizes the rate of a sequence. It corresponds to the time duration between the offset of a stimulus and the onset of the following stimulus. Like the SOA (i.e., TRT), the ISI can be defined within a single stream or across the streams. Finally, the last parameter which influence the rate of the sequence is the duration of the stimuli (see Figure 8).

van Noorden (1975) investigated to what extent stream segregation of pure tones is influenced by the rate of the sequences. For a given TRT, the listeners were asked to adjust the frequency difference between two alternate tones until they reached the limits of coherence (TCB) and fission (FB) (method of adjustment, see next session for more details concerning the procedure). The results showed that the TCB is strongly affected by the TRT: an increase of the TRT induces an increase of the TCB, while the FB is only weakly affected (see Fig-

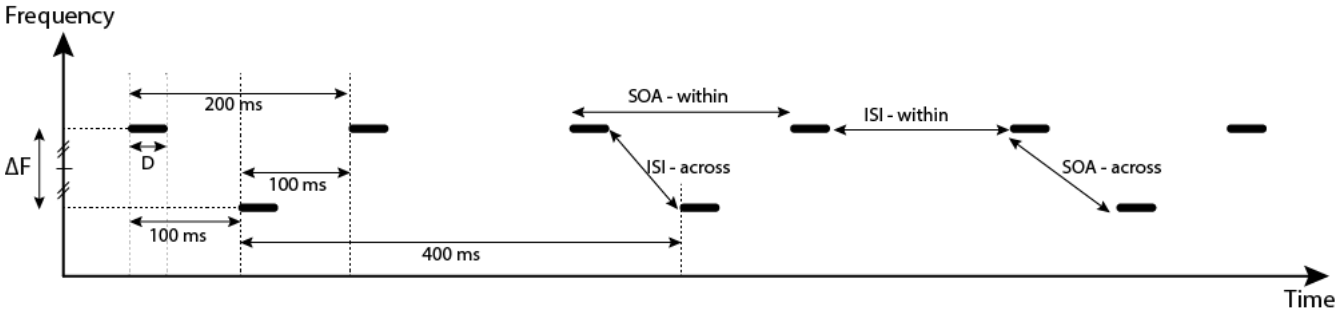

Figure 8 – Sequence of galloping pattern with high- and low-frequency tones, reproduced from Bregman *et al.* (2000). The five types of time interval are illustrated: SOA-within, SOA-across, ISI-within, ISI-across and duration of the stimuli.

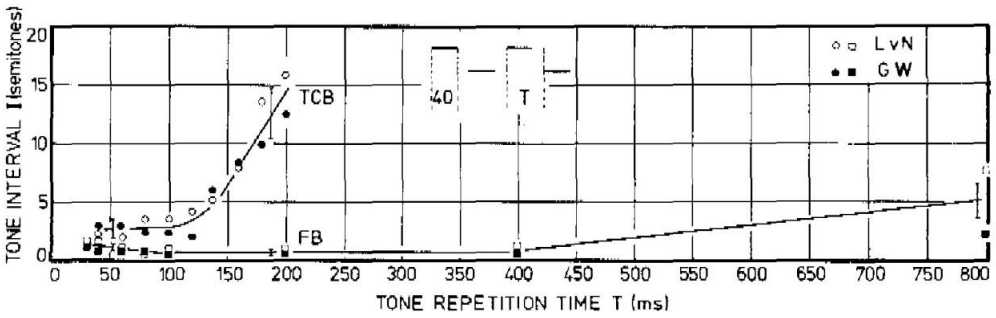

Figure 9 – Temporal coherence boundary and Fission boundary (TCB and FB) measured by van Noorden (1975) as a function of the tone repetition time (TRT). For a given TRT, the listeners were asked to adjust the frequency difference between two tones, presented in sequences, until they reached the limits of coherence (TCB) and fission (FB). These results indicate that the TCB is strongly influenced by the TRT while the FB is only slightly influenced by it.

ure 9). This indicates that a small frequency difference is required to segregate fast sequences of alternate sounds, and conversely, a large frequency difference is required to segregate slow sequences.

The results obtained by van Noorden (1975) have been very useful to design streaming experiments. However, this experiment focused on the TRT and as explained earlier, other factors can influence the rate of a sequence, especially the ISI and the stimulus duration. Thus, Bregman *et al.* (2000) investigated the influence of all temporal parameters of the sequences (i.e., SOA-within, SOA-across, ISI-within, ISI-across and tone duration, see Figure 8) on stream segregation. Listeners were presented with sequences of [ABA-ABA] triplets, asked to hold on to the single-stream percept and to rate the difficulty to do so. This type of procedure is another way to measure the TCB. They found that the ISI within streams accounted for a large proportion of the observed effect of speed increasing stream segregation.

3.1.3 Lateralization differences

The sounds emanating from a source at a given position in space are diffracted by the head and the external ear. This diffraction induces spectral and temporal modification of the signals. Depending on the position of the source related to the listener, the two ears do not receive the exact same stimulation. Indeed, the farthest ear from the source is shadowed by the head, which interrupts the sound path, resulting in a difference in sound pressure level, ILD (i.e., interaural level difference). The shadowing depend on the wavelength of the sound (λ) and on the dimensions of the head. Besides, a difference of time arrival between the ears (expressed in terms of interaural time delays, ITD) results from the longer path for the sound wave to reach the farthest ear. These spectral and temporal interaural differences (ILD and ITD, respectively) provide cues to the auditory system for sound localization (Middlebrooks and Green (1991); Moore (2007)). Figure 10 illustrates the interaural differences.

Above 1000 Hz, the wavelength of the sound λ becomes substantially larger than the dimensions of the head which no longer induces diffraction. Thus, above 1000 Hz, the ILD can be neglected. Moreover, the sensitivity to phase differences decreases with increasing frequency Rayleigh (1907). Thus, spatial information derives from ITD at low frequencies and from ILD at high frequencies. This notion is often named the "duplex theory" of sound localization (Rayleigh, 1907).

The auditory system can rely on differences in lateralization to perceptually organize the sounds. The next paragraphs detail how interaural differences influence stream segregation based on the results of some previous studies.

A series of experiments presented later in this dissertation focuses on this particular point (see Chapter I.2).

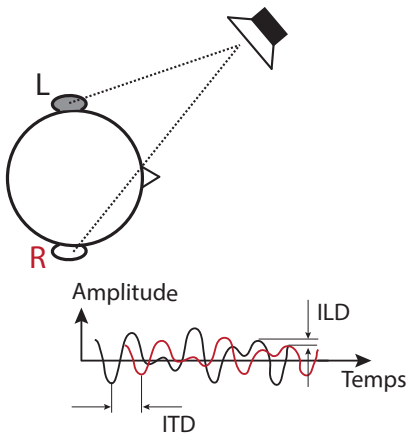

Figure 10 – When a source is located at a given position in space, the sound produced presents binaural differences. A level difference between the two ears, ILD, is introduced because of the shadowing of the head. Here, the sound recorded in the left ear is louder than the sound recorded in the right ear. A difference of time arrival is also observed between the ears, ITD, due to the different distance to reach the ears. Here, the sound leads at the left ear because of the shorter path.

Influence of monaural level differences

Stainsby *et al.* (2004) investigated to what extent differences in intensity influence stream segregation. In their second experiment, the listeners had to detect a rhythmic irregularity within sequences of alternate complex tones [A-B-A-B...] (rhythmic discrimination procedure, see section 3.2). The level of the As was kept constant at 70 dB and the level of the Bs was set 0, 1, 3, 5, 10 or 15 dB below depending on the tested condition. The results suggested that a level difference of 5 dB (or more) across streams can induce stream segregation.

Influence of ILD

It has been shown that ILD is used by the auditory system to segregate auditory streams. For example, Hartmann and Johnson (1991) found that listeners were more accurate to recognize a target melody of pure tones interleaved with a masker when they differed in ILD by 8 dB. These results were confirmed later by the study of Sach and Bailey (2004) who used a task where listeners were asked to identify a target rhythm interleaved with an arrhythmic masker. The target was better identified when the difference in ILD was as small as 4 dB between the target and the masker.

Besides, considering a sequence of two repeating alternate sounds, ILD induces interaural differences but also monaural intensity differences at each ear over time. It has been shown that level differences across streams lead to segregation (see paragraph above).

Experiment 2 of Chapter I investigated the individual contribution of the monaural level differences associated with an ILD and the ILD *per se*.

Influence of ITD

Some studies in the literature showed the influence of ITD on the segregation of sound sources. Hartmann and Johnson (1991); Sach and Bailey (2004) used a pure tone melody recognition task to investigate the influence of a difference in ITD on stream segregation. A target melody was interleaved with a masker, and the aim of the listeners was to recognize the target melody. These two studies showed that the listeners obtained higher performance when the target and interferer differed in ITD. The performance increased as the difference increased from 100 to 600 μs.

Gockel *et al.* (1999) showed that ITD can influence the perceptual organization of complex tones. The task consisted of detecting a change in F0 of a complex tone. They presented a target sound which was either preceded or followed by harmonic complexes (i.e., temporal "fringes"). When the fringes and the target were grouped together, the F0 discrimination was impaired, and when they were separated there was no interference in the F0 discrimination. The results indicated that the impairment induced by the fringes was reduced when target and fringes differed in perceived position (i.e., differences in ITDs). This suggested that stream segregation of complex tones can be observed based on a difference of lateralization.

ITD can also be used by the auditory system to organize speech sounds. Indeed, Kidd *et al.* (2008) presented two sequences of alternate speech words (the odd number words corresponded to the target and the even number words corresponded to the masker), and the listeners were asked to track the target words. The results indicated that the scores of the target-words identification increased as the difference in ITD between target and masker increased from 150 to 700 μs.

However, other studies using different tasks, showed only a weak effect of ITD on auditory stream segregation. Boehnke and Phillips (2005) presented two sequences of alternate broadband noises [ABA-ABA] to the listeners. In one sequence, the silent interval between the stimuli was kept constant throughout the sequence and in the other sequence the interval [AB] was longer than the interval [BA]. The listeners had to detect the sequence with the temporal asymmetry. The smallest asymmetry detected (i.e., the threshold) by the listeners rely on the perceptual organization of the streams since the task was facilitated by integration. Thus, a large threshold indicated a greater tendency to hear the streams segregated. The results showed no significant improvement in segregation when the stimuli differed in ITD by 500μs.

Stainsby *et al.* (2011) investigated the influence of ITD on the segregation of complex tones. They used a rhythmic discrimination task where the performance was favoured by the integration of the stimuli (see section 3.2). They tested five differences of ITD between the streams: 0, 250, 500, 1000 and 2000 μs. Note that the ITDs in the physiological range are comprised between 0 and 700 μs (Feddersen *et al.*, 1957). The results showed that a difference in ITD can influence stream segregation when ITD was larger than 1 ms. This result was much weaker than the influence of a difference in passband. Füllgrabe and Moore (2012) replicated the experiment of Stainsby and colleagues with pure tones and ITD within the physiological range (i.e. from 0 to 500 μs). The results indicated only a weak effect of ITD.

Conclusion

The previous studies detailed above suggest that ILD and ITD are relevant cues for stream segregation for situations where voluntary streaming is measured. In the situations where obligatory streaming is observed, ILD was a useful cue but the influence of ITD seemed to depend on the type of stimuli. In fact, the literature showed no influence of ITD with pure tones, and only a weak effect with complex tones.

An hypothesis might be that pure and complex tones do not provide enough ITD information compared to a broadband noise. The second study presented in Chapter I of this dissertation investigated this question.

3.1.4 Cumulative effect of segregation and influence of attention

Stream segregation is not an instantaneous percept, instead, it is cumulative. When listening to an auditive stimulation, the auditory system begins to hear a single stream, thus, integration is the default percept. Seconds after, evidences can lead to a split of the streams. Bregman (1978b) demonstrated the effect of build-up of segregation. In their first experiment, sequences of pure tones were presented to the listeners. The sequences consisted of several repeats of tones packages. The packages were of different lengths, containing 4, 8 or 16 tones while the frequencies of the tones were fixed. The tones were separated by a silence gap such as the sequences lasted 4 s. An "infinite" condition consisted of presenting tones without silence interval between them. In each condition, the listeners were asked to adapt the speed of the sequence by varying the onset-to-onset time, until the point of splitting was obtained. The results are shown in Figure 11. The y-axis represents the onset-to-onset time at the splitting point (high values corresponds to slow sequences, so a greater tendency to segregate the streams

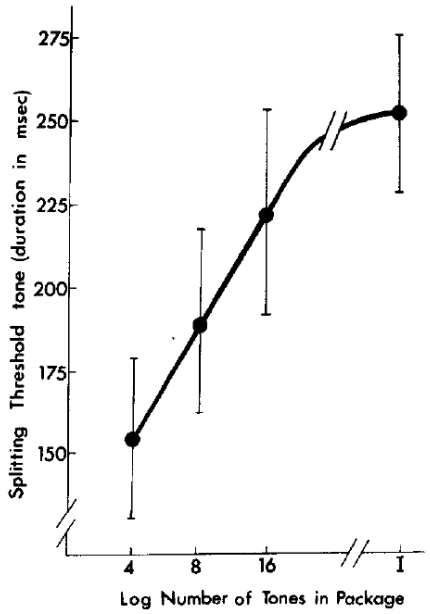

Figure 11 – Results extracted from Bregman (1978b). The y-axis represents the speed thresholds as a function of the number of tones in the sequence (4, 8, 16 or infinite repetition). A high value of threshold means a greater tendency to segregate the streams independently of the speed of the sequence.

independently of speed), and the x-axis represents the number of tones per packages in the sequences. This experiment showed that the speed required to perceive segregation decreased as the sequence length increased, meaning that the percept of segregation built-up across time.

Once the percept of segregation is installed, the perceptual organization of the sounds can be reset by abrupt changes in their parameters such as location, loudness, ear of entry and tone frequency difference (Anstis and Saida, 1985; Roberts *et al.*, 2008; Rogers and Bregman, 1998).

Carlyon *et al.* (2001) investigated the influence of attention on the build-up of auditory stream segregation. They presented sequences of [ABA-ABA] tone triplets. The sequences were presented to one ear for 21 s. In the baseline condition, no sounds were presented in the contra-lateral ear and the listeners were asked to report whether they heard one stream or two streams (i.e., a streaming judgment task). In the "two tasks" condition, noise bursts were presented in the contra-lateral ear. These bursts were either approaching (i.e., their level increased) or departing (i.e., their level decreased). For the first 10 s of the sequence, the listeners were asked to ignore the tones presented in the first ear and report whether the noise bursts were approaching or departing. After 10 s, they were told to switch their attention to the other ear and perform the judgment task. Finally, in the "one-task with distractor"

condition, the same noise bursts were presented to the contra-lateral ear but listeners were asked to ignore them and perform the streaming judgment task. The results showed that the tendency to report two streams increased as the sequence progressed in time. The build-up of segregation was strongly reduced when the listeners focused on the contra-lateral ear (i.e., without attention). However, the competing task did not completely prevent the build-up. These results provided evidence that the build-up of stream segregation depends on attention.

Thompson *et al.* (2011) assessed the build-up of stream segregation and whether it was influenced by attention using an objective task (i.e., not directly based on the listeners' judgment). In their first experiment, they presented sequences of 25 tone triplets [ABA-ABA]. The task consisted of detecting a delay of a B-tone within one triplet, all the other triplets remaining regular. The task was favoured by integration (see the next session for further explanations). The delay occurred either early or late in the sequence. The results showed that the delay was harder to detect when it appeared late in the sequence indicating that segregation build-up over time. In their second experiment, the tone triplets were presented in one ear, and noise bursts were presented in the other ear. Listeners either had to focus on the first ear throughout the sequence and perform the delay-detection task, or focus for 10 s to the contra-lateral ear (i.e., the noise bursts) and then switch their attention to the delay-detection task. In this switch-attention condition, the detection of a late delay was better than in the constant-attention condition indicating a reset in the build-up of segregation. Thus, according to the findings of Carlyon *et al.* (2001); Thompson *et al.* (2011), the build-up of segregation can be reduced by a switch of attention or by the absence of attention, or by a combination of both.

The results of these studies are relevant to design new streaming experiments. In each experiment presented in this dissertation, a particular attention was paid to the build-up and resetting of the segregation.

3.1.5 Influence of regularities in stream organization

The last parameter influencing auditory stream organization that will be discussed in this dissertation is the pattern regularity. In the studies presented so far, the sequences consisted of either alternate sounds [A-B-A-B...] or sound triplets [ABA-ABA...]. The stimuli used were constant, meaning that the As (Bs, respectively) were all the same within a sequence. Even if real sound sources often present regular characteristic patterns, such as accentuations on some tones or frequency patterns induced by a fixed location in space, they are unfrozen. Indeed, considering speech sounds, speakers do not repeat the same stimuli in sentences. Bendixen *et al.*

(2010) investigated whether regularities in these patterns influence the perceptual organization of pure tones. They presented sequences of tone triplets [ABA-ABA] where A and B had either both random frequency and intensity, or presented regularities within the A tones or the B tones or both. The regularities consisted of a regular repeating frequency pattern (i.e., $F_{A_2}F_{A_2}F_{A_1}F_{A_1}$ and $F_{B_1}F_{B_2}F_{B_3}$), a regular intensity pattern (i.e., an accent every four A tones and every three B tones), or both regular frequency and intensity patterns. The listeners had to report their percept (one stream or two streams) at any time. The results, shown in Figure 12, indicated that the proportion of integrated percept tended to increase as the number of regularities increased. Thus, the introduction of temporal regularities is used by the auditory system and tends to stabilize the stream organization once it has been formed. In the same way, Devergie *et al.* (2010) found that the performances in a melody recognition task (see next session) were better when the rhythm of the interferer was regular rather than irregular.

While Bendixen *et al.* (2010) investigated the influence of regularity on stream segregation, the first study presented in Chapter II of this dissertation evaluated the robustness of stream segregation to spectral variability.

3.1.6 Other parameters influencing auditory stream organization

The previous sections were deliberately not exhaustive. The aim was to focus the literature review on the parameters which were used in the different studies of this thesis. However, other factors play a role in auditory stream segregation. This section is dedicated to briefly describe some of these parameters.

Phase spectrum

Stream segregation can occur without peripheral channeling. For instance, Vliegen and Oxenham (1999) investigated the ability to perceptually organize sounds without spectral cues. They presented complex tones with either resolved or unresolved harmonics. The results showed that the listeners could hear segregation more than half of the time once the fundamental frequency difference between the stimuli was greater than 4 semitones in all conditions. This indicates that having resolved harmonics (so peripheral channeling) is not a necessary condition to observe perceptual stream segregation. Roberts *et al.* (2002) used complex tones containing

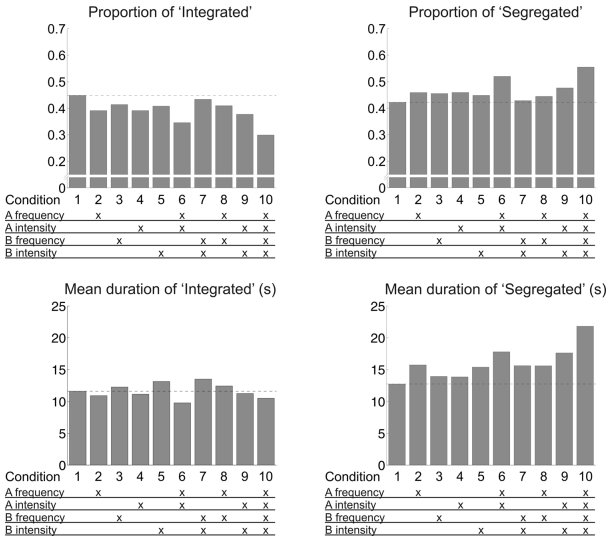

Figure 12 – Results obtained by Bendixen *et al.* (2010). Sequences of tone triplets were presented to the listeners who had to report their perception (one stream or two streams). In some conditions, regular patterns were introduced within one or both streams. The results are expressed in terms of proportion of integrated percept (segregated percept, respectively, top panels) and mean duration time when the percept was integrated (segregated, respectively, bottom panels). The results indicate that the introduction of temporal regularities tend to stabilize the stream organization once it has been formed.

only unresolved harmonics (so with similar excitation patterns). They presented sequences of complex tones that phase relations between the harmonics were manipulated. This way, the pitch and the timbre could be modified without changing the energetic content. They found that dissimilarities in phase between the stimuli favoured segregation. This result indicated that segregation can be induced without differences in peripheral channeling.

Temporal envelopes - Timbre

Some studies showed that temporal differences, in the absence of spectral cues, can affect stream segregation. For example, Dannenbring and Bregman (1976) used sequences where the alternate stimuli could be both pure tone, both narrowband noises or a pure tone and a narrowband noise. Their results indicated that segregation was stronger in the "non-matching" stimuli combinations (i.e., pure tone - noise), rather than in the "matching" stimuli combinations (i.e., pure tone - pure tone or noise - noise). Grimault *et al.* (2002) found that a difference in amplitude-modulation rate between stimuli of 9 semitones or less induce a greater tendency to hear integration; for a difference of 12 semitones or more, segregation was often reported. These results showed that large envelope fluctuations can produce stream segregation.

Such differences in temporal envelope produce differences in timbre. According to the American Standard Association, the timbre is "that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar". The effects of timbre on stream segregation were studied, among others, by Iverson (1995). By presenting sequences of tones produced by different orchestral instruments, they showed that differences in timbre can be used by the auditory system to segregate items.

3.2 Auditory streaming measurements

3.2.1 Behavioural experiments

Behavioural experiments enable one to measure the percept of auditory streaming. In those experiments, the relationship between a physical parameter of a stimulus and the subjective responses of the listeners can be described by a psychometric function. The percentage of correct answers (the term "correct" depends on the experiment[1]) is displayed on the y-axis and the

[1]Let us consider the example of an experiment which aims to investigate the influence of a parameter α on the perceptual segregation. To do so, the experimental design might consist of presenting sequences of alternate stimuli to the listeners and asking them to rate their perception (one stream or two streams) for each value of α. The experimenter could decide to display the results in terms of percentage of 2 streams. In this example, the "correct" answer would be "2 streams". Thus, the value on the y-axis would be 0 when the listeners reply "1 stream" and 1 when they reply "2 streams".

value of the parameter is displayed on the x-axis. The proportion of correct answers when the parameter is null corresponds to the chance level and is determined by the number of possible responses (simple forced choice named yes-no tasks, two-alternative forced choice (2AFC) and n-alternative forced choice). The psychometric curves present an inflection point. This point is generally taken as the sensory threshold because it gives the value of the parameter which leads to a bi-stable percept (i.e., the midpoint between the chance level and 100% of correct answers). Besides, the slope of a psychometric curve is related to the sensitivity to the parameter. Indeed, when the slope is steep, a slight variation of the parameter leads to a large variation in the subjective response; and when the slope is flat, a large variation of the parameter is required to observe a slight variation in the subjective response.

In those behavioural experiments, the value of the parameter can be held constant across the presented sequences (method of constant stimuli) or be adapted depending on of the listeners' responses (adaptive methods). In the methods of constant stimuli, the value of the parameter varies randomly from trial to trial. This way, the evolution of the percept as a function of the parameter can be measured and the whole psychometric curve can be evaluated. Besides, the adaptive procedures enable to determine one particular point on the psychometric curve, so the value of the parameter which corresponds to x percent of correct answers on the psychometric curve. Note that x depends on the adaptive rule used in the experiment.

Levitt (1971) formalized up-down adaptive procedures. The rule of the adaptive procedure can be expressed as n-up, m-down meaning that the parameter level is decreased after m correct responses (leading to a more difficult task), and it is increased after n wrong responses (leading to an easier task). At the convergence point, the probability of giving a correct response (i.e., $P[X]$) should be equal to the probability of giving a wrong answer (i.e., $1 - P[X]$), and depends on the values of n and m. For example, if n and m are equal to 1, at convergence, $1 - P[X] = P[X]$ so $P[X] = 0.5$. Thus, a simple up-down procedure estimates the 50% correct point on a psychometric curve. Let us take another example that will be used in different experiments described later in this dissertation, a 3-down 1-up adaptive procedure (i.e., $n = 1$ and $m = 3$). In this case, the parameter is increased after one wrong response (i.e., $1 - P[X]$) or after one correct followed by one wrong responses (i.e., $P[X](1 - P[X])$) or after two correct followed by one wrong responses (i.e., $P[X]^2(1 - P[X])$). So, the probability of going up is $1 - P[X] + P[X](1 - P[X]) + P[X]^2(1 - P[X])$. In the same way, the parameter is decreased after three correct responses, so the probability of going down is $P[X]^3$. At convergence, the probability of going up is equal to the probability of going down, so $1 - P[X] + P[X](1 - P[X]) + P[X]^2(1 - P[X]) = P[X]^3$, then $P[X]^3 = 0.5$ and so $P[X] = 0.794$. Thus, a 3-down 1-up adaptive procedure estimates the 79.4% correct point on a psychometric curve.

Besides those considerations on the parameter variation (constant or adaptive stimuli), a distinction has to be made between the objective and subjective psychological measures. An objective measure consists of a task for which performance depends on the percept while a subjective measure consists of situations where the listeners are directly asked to evaluated their perception. The following paragraphs review the subjective and objectives methods used to measure auditory streaming.

3.2.2 Subjective methods

Method of adjustment

In the method of adjustment, the listeners are asked to adapt one parameter of the stimulus (for example its level) to perceptually match another fixed stimulus. A measure of sensitivity is then made taking the mean difference between the tested and the fixed stimulus on several numbers of repetitions. Another declination of this measure consists of letting the listeners adapt the parameter (for example a frequency modulation rate) until it is just barely detectable. An example of this method of adjustment was used by van Noorden (1975) to measure the TCB and FB. Indeed, the listeners were asked to adapt the speed of the sequences until they reached the limit of coherence and the limit of fission (TCB and FB, respectively). Other examples of this type of method can be found in the studies of Bregman (1978b); Dannenbring and Bregman (1976); Heise and Miller (1951); Miller and Heise (1950); van Noorden (1977).

Method of limits

In the method of limits, the stimulus parameter is controlled by the experimenter. This method was used in the studies of Rose and Moore (1997); van Noorden (1975); Vliegen and Oxenham (1999) For example in an experiment testing the minimum amplitude difference between two streams for them to be heard separately, one starts with a null or very small difference so that the streams are integrated and this difference is gradually increased until the listener reports that he/she hears segregation. Then, to avoid any anticipation or habituation bias, the procedure is adapted to the listener's responses. Thus, with the same example, once he/she reports segregation, the difference between the streams decreases gradually until segregation is no longer heard and then increases again. This adaptive procedure enables to evaluate a threshold associated with the limit of segregation. This procedure can also be descending instead of ascending, starting with a high value of the parameter.

Proportion of time integrated / segregated

In this type of psychophysical measure, the listeners hear a sequence of sounds and have to report at any time if the streams are integrated or segregated by holding a corresponding button. The stimulus' parameter considered is gradually modified and each change in percept is reported. This way, the mean duration of one or the other percept can be evaluated depending on the parameter's characteristics. For example, Bendixen *et al.* (2010) used this measure to determinate the proportion of time when the streams are integrated (segregated, respectively) and the mean duration of each percept.

Streaming rate with constant stimuli

In this method, instead of having a parameter which gradually increases or decreases during the stimulus presentation, the parameter's value is fixed within a trial and each value is randomly presented from one trial to the next one. At the end of each trial, the listeners have to report which percept was predominant during the presentation. No indication is given to the listeners to avoid biasing them towards one percept or the other. Besides, the randomization of the parameter's value prevents the prediction of the following stimulus and this way the habituation and expectation errors are reduced. Another declination of this method consists of asking the listeners to focus on a given percept and rating the difficulty to do so. The parameters can vary gradually and they have to rate the difficulty at any time using a cursor for example. Otherwise, the parameters can vary randomly from trial to trial and the difficulty is rated at the end of each presentation.

This method was used in different experiments detailed later in this dissertation (see Chapter I, section 1.2 and Chapter II, section 1).

3.2.3 Objective methods

The subjective methods described above directly access the perception, however it can present experimental bias since the experimenter does not control how the listener understands and interprets the task, neither which strategy is used to do the task. Conversely, objective methods are designed to be independent from subjective interpretations. The listeners are asked to do a task in which performance depends on the segregation state (i.e., integrated or segregated). As they are not directly asked to report what they hear, objective methods give an indirect measure of streaming.

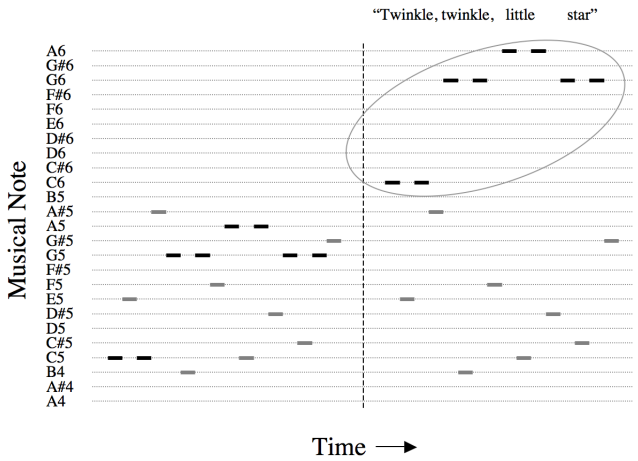

Figure 13 – Design of a pattern recognition experiment, taken from Plack (2005). On the left panel, the target and the masker melodies are interleaved. The task consists of recognizing the target melody which is the well-known "Twinkle, twinkle little star". To succeed the task, the listeners have to segregate the streams and extract the target melody, which is shown in the right panel.

Pattern recognition

In this type of measurement, the listeners are presented with a sequence of sounds which consists of alternate notes from a known melody and notes from a masker. The task is to recognize the known melody within the mixture of sounds (Deutsch, 1975; Dowling, 1973; Gregory, 1994; Hartmann and Johnson, 1991; Iverson, 1995; Rogers and Bregman, 1993; Singh, 1987; Vliegen *et al.*, 1999). Since the two melodies (i.e., target and masker) are interleaved, the listener has to segregate the streams to extract the target and recognize the melody. Thus, this type of tasks is favoured by segregation. The target melody can be made with well-known melodies like "Twinkle, twinkle little star" as presented in Figure 13 taken from Plack (2005), or an unknown melody which is learnt in a first part of the experiment. Different variations of the pattern recognition method can be found in the literature, some replacing the melody recognition by a rhythm recognition (see for example Middlebrooks and Onsan (2012)).

Rhythmic changes

The rhythmic discrimination task was introduced by Roberts *et al.* (2002), and consists of presenting two sequences of alternate sounds [A-B-A-B...]. In one sequence (the reference interval), the B sounds are placed at the exact temporal midpoint between two consecutive A sounds. In the other sequence (the target interval), the B sounds are progressively delayed. Thus, the silent gap [A-B] becomes progressively longer than the silent gap [B-A], leading (when

perceptible) to an irregular rhythm. The listener's task is to identify the interval containing the irregular sequence. The delay applied to the B sounds varies adaptively in order to measure a detection threshold. The top part of Figure 14 illustrates the two intervals (i.e., regular on the left and irregular on the right). The perceived rhythm of the target interval depends on whether the listener hears a single stream or two segregated streams. Indeed, when the percept is integrated, the listener can follow across time the alternation of stimuli and the delay applied to the Bs is detectable since the two intervals [A-B] and [B-A] can easily be compared. Conversely, when the percept is segregated, the listener hears the A- and B-streams separately, and the delay applied to the Bs becomes barely detectable since he/she can only compare the [B-B] intervals. Thus, the irregularity is more easily detected when the percept is integrated, so the task is favoured by integration. In other words, the listener tries to fuse the streams to perceive the irregularity and the delay is adapted to find the threshold below which the listener is no longer able to hear fusion. Thus, it is a measure of obligatory streaming.

This procedure will be detailed later in this manuscript because it has been used in several experiments in this PhD work.

Counting / Order of the elements

The two following procedures rely on the ability to perceive the order of sounds in rapid sequences. This type of procedures was used in several studies, such as Bregman (1978a); Bregman and Rudnicky (1975). Bregman and Campbell (1971) showed that for a given frequency difference between the stimuli, the streams tend to be more segregated at high presentation rates than at low ones. Besides, they also showed that when the streams are segregated, the listeners can only report or count the elements presented in one stream and no longer all the elements of the sequence. Gaudrain *et al.* (2007) used this type of procedure to measure auditory streaming. They presented sequences of vowels, and asked to repeat the stimuli heard in the right order. The frequency difference between the streams varied from trial to trial. If the percept was integrated, the listeners could report in the right order all the vowels. However, when the percept was segregated, they could only report in the right order the vowels within one of the streams. Figure 15 illustrates the procedure used by Gaudrain and colleagues. Depending on what is measured (i.e., the correct order across streams or the correct order within streams), it is possible to investigate auditory stream integration or segregation.

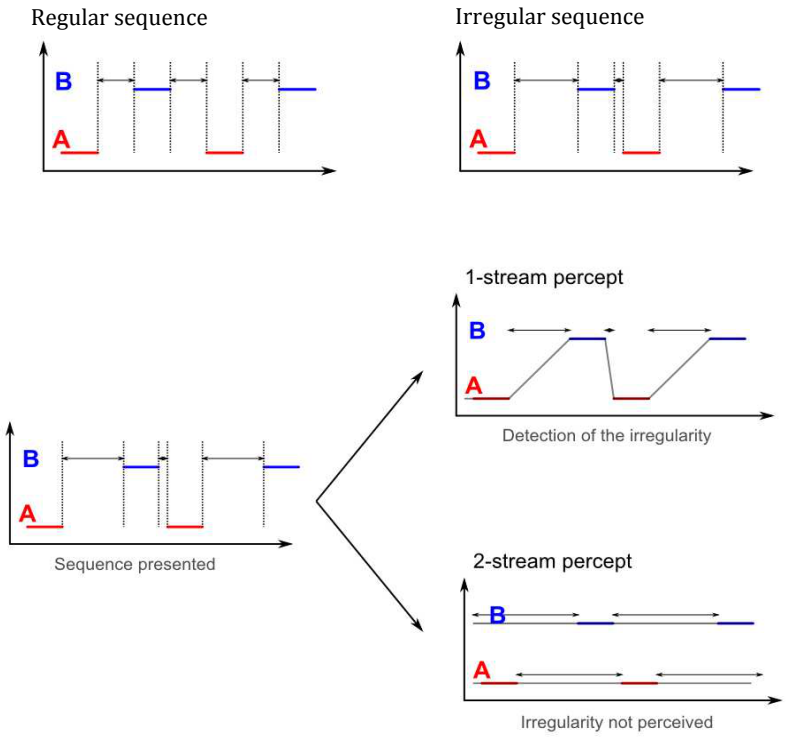

Figure 14 – Detection of a rhythmic change. Two sequences of alternate stimuli [A-B-A-B...] are presented to the listener (see top panel). In one sequence, the stimuli are regularly spaced (left sequence) and in the other sequence, stimuli B are progressively delayed, leading to a longer interval [A-B] and a shorter interval [B-A]. The task consists of reporting which sequence has an irregular rhythm. The perceived rhythm of the target sequence depends on the percept (bottom panel). When one stream is heard, the listener is able to follow the alternation across time and detect the irregular rhythm. However, when two streams are heard, the listener is no longer able to detect the irregular rhythm since he/she hears two separated regular streams.

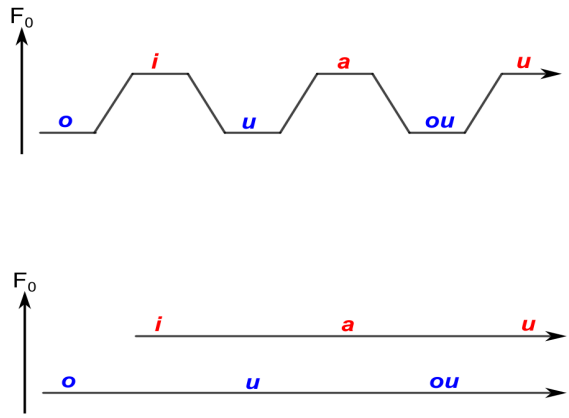

Figure 15 – "Order of the elements" task used by Gaudrain *et al.* (2007). Sequences consisting of several loops of six vowels were presented to the listeners. The task was to identify the vowels. When the percept is integrated (top panel), the listeners are able to name all the vowels of the sequences (o-i-u-a-ou-u). Otherwise, when the percept is segregated (bottom panel) they can only name the vowels within one of the streams (either i-a-u or o-u-ou).

Detection of repetitions

The main idea of the detection of repetition method is to present sequences containing only unique stimuli (i.e., presented once) and sequences containing unique stimuli but one repeat (i.e., one repetition of a single sound just after its first presentation). All the sequences are presented in random order and the listeners have to indicate after each sequence if there was a repeat or not. Figure 16 illustrates this procedure, the black symbols correspond to a repeated stimulus while the other stimuli are different. Then, any acoustical parameter can vary between the A and B-sounds. The repeat can occur across streams (see the top panel of Figure 16). In this case, the task is favoured by integration, so the listeners are biased toward grouping. The procedure estimates the parameter's value above which integration is no longer heard, so this is a measure of obligatory streaming. Conversely, the repeat can occur within streams (see the bottom panel of Figure 16). In this case, the task is favoured by segregation, so the listeners are biased toward segregation. The procedure estimates the parameter's value below which segregation is no longer heard, so this is a measure of voluntary streaming.

This method was used the experiment presented in Chapter II, section 2.

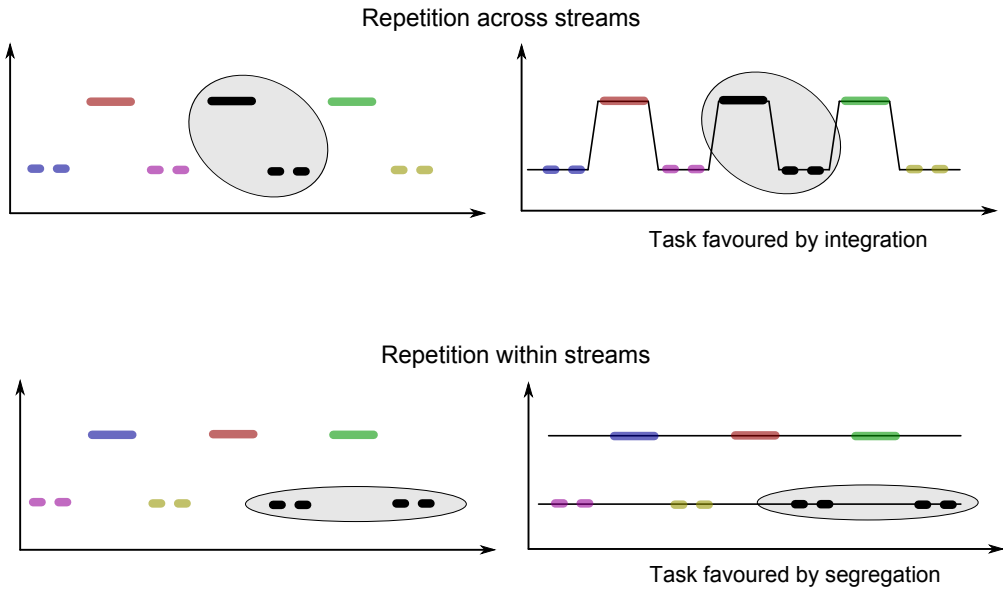

Figure 16 – Sequences of stimuli containing a repeat introduced across the streams (top left panel) and within one stream (bottom left panel). The circled black symbol correspond to a repeated stimulus. All the other sounds are different. When the repeat is introduced across streams, the task is favoured by integration, while when the repeat is introduced within a streams, the task is favoured by segregation.

3.2.4 Conclusion

Both types of measurement (subjective and objective) give a different estimation of auditory streaming. Both have pros (direct measure of streaming for the subjective tasks, independent from listener's interpretation for the objective ones) and cons (bias associated with the listener's comprehension and strategy for the subjective tasks, indirect measure of streaming for the objective ones), but they can be complementary. For instance, Micheyl and Oxenham (2010) compared the results obtained with a rhythmic discrimination task (i.e., an objective method) and those obtained measuring the proportion of time the percept is integrated or segregated (i.e., a subjective method). They found a strong correlation between the two different measures, meaning that high thresholds in the rhythmic task can be interpreted as a strong tendency for the streams to be segregated.

4 Towards auditory stream segregation in realistic situations: aims and organization of the PhD dissertation

Most of the studies presented so far investigated auditory perceptual organization *inside the laboratory*, meaning that the stimuli used were synthesized and processed inside the laboratory. Those manipulations bestowed an artificial (i.e., unrealistic) nature on the stimuli. In fact, sources in real environments often produce sounds that present regular temporal and/or spectral structure, such as harmonic tones produced inside the laboratory. However, in most cases, the sounds produced by real sources are broadband noises instead of pure or complex tones. Thus, auditory stream segregation in realistic situations has to consider ecological sounds such as broadband noises or speech material.

In real environments, the question of the propagation field of the sounds between the source and the listener has to be considered. Presenting stimuli recorded (or synthesized) in anechoic conditions enable to simulate the extreme cases of non-reverberant rooms, near-field sources or outdoors environments. Since these situations represent only a few percentages of the real possible configurations, it is worth considering propagation in finite environments such as rooms. So, considering auditory scene analysis in a real situation, the room and its associated effects (especially colouration due to sound reflections) have to be taken into account.

The main purpose of this work was to try to investigate auditory stream segregation in more realistic situations compared to what had been tested before. Thus, two key points have been emphasized: the realism of the configurations considered and the realism of the stimuli used. First, previous studies have shown that spectral differences can induce stream segregation.

These spectral differences were usually large and resulted from *laboratory manipulations*. The current work focused on spectral differences resulting from real differences in spatial locations (see Chapter I of this manuscript). The influence of spatial differences on streaming was investigated using filtered broadband noises as a first approximation of speech. In a second part (see Chapter II), two preliminary experiments were conducted to assess the acoustical variability contained in running speech. First, the robustness of stream segregation based on a frequency difference to variability on the same acoustic dimension (i.e., frequency) was assessed using pure tones, and second, the fundamental frequency difference required to segregate speech items was evaluated. As shown in Figure 17, the two parts of this work will be concatenated to investigated how spatial differences could affect the segregation of speech items.

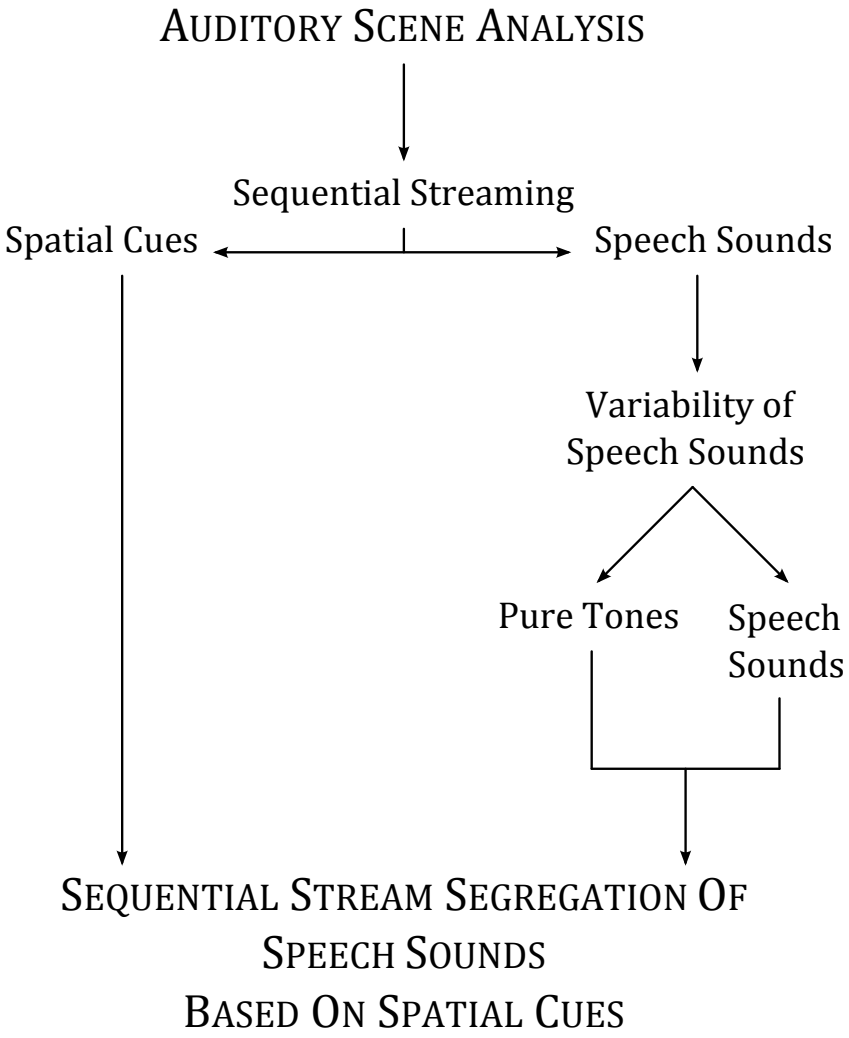

Figure 17 – Outline of the PhD work. The aim was to investigate how sequential stream segregation of speech sounds could be influenced by spatial cues. First, the influence of spatial differences was assessed using broadband noises (see Chapter I). Second, the effect of spectral variability was evaluated using pure tones and real recorded speech material (see Chapter II). Finally prospective are detailed in order to fuse the two parts.

Chapter I

Auditory stream segregation based on spatial cues

1 Room and head coloration can induce obligatory stream segregation

1.1 Abstract - Résumé

Abstract

Sound reflections on room materials induce spectral distortions of sounds. The head and torso of the listener also present reflection and distortion properties which produce slight spectral modifications. This phenomenon referred to as colouration depends on the listener and source positions, and on the room itself.

Generally speaking, room reverberation tends to impair intelligibility (temporal smearing of speech and reduction in spatial release from masking). However, large spectral differences are known to be relevant for the auditory system to segregate competing sources. The aim of our first study was to assess whether the slight spectral differences induced by head and room colouration could help segregation. Three spatial configurations involving different levels of head and room colouration were tested.

The influence of colouration on obligatory stream segregation was evaluated using an objective rhythmic discrimination task with diotic broadband stimuli. In each configuration, thresholds for detecting anisochrony were significantly higher when stimuli differed in spectrum. A subjective experiment confirmed that the streams are often segregated when they come from sources located at different positions. Despite the generally deleterious effect of reverberation on speech intelligibility, these results suggest that colouration can favour auditory stream segregation.

Résumé

Dans les salles, la réverbération entraîne des distorsions spectrales des sons émis par des sources sonores. De la même manière, la tête et le torse de l'auditeur présentent des propriétés de réflexion et de distorsion qui engendrent de fines modifications spectrales. Ce phénomène de coloration, induit par la salle et par la tête, dépend des positions respectives de la source et de l'auditeur ainsi que des propriétés de la salle en elle-même.

D'une manière générale, les réflexions induites par une salle tendent à détériorer l'intelligibilité de la parole (étalement temporel de la parole cible et réduction du démasquage spatial dans le bruit). Par ailleurs, il a été montré que de larges différences spectrales peuvent être utilisées par le système auditif pour dissocier des flux sonores. L'objet de cette première étude était alors de déterminer si les fines différences spectrales induites par la coloration pouvaient aider à séparer des sources concurrentes. Trois configurations ont été étudiées, mettant en jeu différents niveaux de coloration induite par la tête et la salle.

Une tâche objective de discrimination rythmique a été mise en place pour évaluer l'influence de la coloration sur la ségrégation obligatoire de stimuli diotiques large-bande. Dans chaque configuration étudiée, les seuils de détection de l'irrégularité rythmique étaient significativement plus élevés lorsque les stimuli présentaient des différences spectrales. Une expérience de ségrégation subjective a confirmé que les flux étaient majoritairement perçus séparément lorsqu'ils provenaient de sources localisées à des positions différentes. Ces résultats suggèrent qu'en dépit de l'effet nuisible de la réverbération sur l'intelligibilité, les fines différences spectrales induites par la coloration peuvent favoriser la ségrégation de flux auditifs.

1.2 J. Acoust. Soc. Am. 136(1), July 2014

Room and head coloration can induce obligatory stream segregation (L)

Marion David[a] and Mathieu Lavandier
Université de Lyon, École Nationale des Travaux Publics de l'État, Laboratoire Génie Civil et Bâtiment, Rue M. Audin, 69518 Vaulx-en-Velin Cedex, France

Nicolas Grimault
Unité Mixte de Recherche au Centre National de la Recherche Scientifique 5292, Centre de Recherche en Neurosciences de Lyon, Université Lyon 1, Cognition Auditive et Psychoacoustique, Avenue Tony Garnier, 69366 Lyon Cedex 07, France

(Received 25 November 2013; revised 29 May 2014; accepted 3 June 2014)

Multiple sound reflections from room materials and a listener's head induce slight spectral modifications of sounds. This coloration depends on the listener and source positions, and on the room itself. This study investigated whether coloration could help segregate competing sources. Obligatory streaming was evaluated for diotic speech-shaped noises using a rhythmic discrimination task. Thresholds for detecting anisochrony were always significantly higher when stimuli differed in spectrum. The tested differences corresponded to three spatial configurations involving different levels of head and room coloration. These results suggest that, despite the generally deleterious effects of reverberation on speech intelligibility, coloration could favor source segregation.
 [http://dx.doi.org/10.1121/1.4883387]

PACS number(s): 43.66.Mk, 43.55.Hy, 43.66.Ba, 43.66.Qp [VB] Pages: 5–8

I. INTRODUCTION

When considering speech intelligibility in rooms in the presence of competing voices, reverberation is associated with multiple effects, generally impairing intelligibility. The sound reflections in rooms directly affect the target speech, which can be temporally smeared at high levels of reverberation, resulting in lower intelligibility (Houtgast *et al.*, 1980). Sound reflections also reduce the advantage observed when separating target and interferer (spatial release from masking) at relatively low levels of reverberation (Lavandier and Culling, 2008). Another result of listening in rooms is that slight spectral modifications are induced at each reflection, depending on the frequency-dependent absorption properties of the encountered material. The sound reaching the ears results from the addition of the direct sound and the multiple filtered reflections, leading to constructive and destructive interferences which amplify or attenuate different frequency components. These modifications of sound spectrum associated with room coloration are dependent on the listener and source positions, and on the room itself.

The extraction of relevant information from a mixture of competing sounds involves their perceptual organization (Bregman, 1990). A sequence of sounds can be perceived as a single coherent stream (integration), or the scene can be broken into two or more streams (segregation), depending on the acoustical characteristics of the sound. Among these characteristics, the present study focused on the role of spectral differences. It investigated whether the spectral differences associated with coloration could help segregate competing sources.

A rhythmic discrimination task (Roberts *et al.*, 2002) was used to evaluate the influence of coloration on obligatory streaming, which reflects situations where segregation occurs even when the listener is trying to fuse streams. In addition to this objective psychological measure, a subjective measure of auditory streaming was done under the same stimulus conditions. Three objective experiments and one subjective experiment are presented in which the spectral characteristics of coloration were conveyed by head- or room-related impulse responses. The three configurations (rooms and positions) were arbitrarily chosen, but they provided a range of examples and were representative of three extreme cases. The spectral differences induced by coloration were small and the stimuli presented a large degree of overlap in their excitation patterns (Moore and Glasberg, 1983). Most previous studies investigating the influence of spectral differences on streaming used stimuli with large differences that excited either different or partially different auditory channels (Hartmann and Johnson, 1991). It remains unclear whether the relatively subtle spectral differences caused by coloration are sufficient to induce obligatory streaming.

One motivation for this study comes from studies that have demonstrated a relationship between auditory streaming and speech intelligibility [for a review, see Grimault and Gaudrain (2006)]. For example, Gaudrain *et al.* (2012) showed a significant correlation between sequential streaming of vowels and speech-in-noise perception performance. In the present study, broadband noises with the same long-term spectrum of speech were used as a first approximation of speech sounds, without the large spectral variations associated with speech. Although a simplification, this approach enables us to focus on the small spectral differences associated with

[a]Author to whom correspondence should be addressed. Also at: UMR CNRS 5292, Centre de Recherche en Neurosciences de Lyon, Université Lyon 1, Cognition Auditive et Psychoacoustique, Avenue Tony Garnier, 69366 Lyon Cedex 07, France. Electronic mail: marion.david@entpe.fr

coloration. If coloration can enhance the segregation of broad-band noises, it is plausible that it might also enhance the segregation of speech and thus favor intelligibility.

II. METHODS

A. Procedures

1. Objective procedure

To measure obligatory streaming, a rhythmic discrimination procedure (Roberts *et al.*, 2002) was used. This procedure involved the presentation of two intervals of twelve AB pairs of sounds. In the reference interval, the B's were placed at the temporal midpoint between two consecutive A's. The interval between consecutive stimuli was 40 ms. In the target interval, the first six AB pairs were regularly spaced by 40 ms, then the B's were progressively delayed by equal steps for the next four pairs. Thus, the silent gap [A-B] became progressively longer than the silent gap [B-A], leading (when perceptible) to an irregular rhythm. The cumulative delay was kept constant for the last two AB pairs. The listener's task was to identify the interval containing the irregular sequence, whose delay was varied adaptively to measure a detection threshold. The perceived rhythm of the target interval depended on whether the listener heard a single stream or two segregated streams. When one stream was heard, the delay applied to the B's was easily detectable, even for small delays, because it could be compared to the previous time intervals [A-B]. Conversely, when the streams were segregated, the delay applied to the B bursts was more difficult to detect, even for large delays, because it was compared to the previous time interval [B-B].

Thresholds for detecting anisochrony were estimated with a two-interval two-alternative forced-choice adaptive method (Levitt, 1971). The delay applied to the B bursts was adapted according to a three-down one-up rule and varied on a logarithmic scale. The maximum delay was 40 ms. The initial value of the delay was 28.28 ms. When listeners gave three consecutive correct answers, the delay decreased by a factor 1.414, and it increased by the same factor when listeners gave one wrong answer. Each run was divided into successive blocks of ten trials (corresponding to the ten pairs of stimuli, see Sec. II B). Once at least two reversals were obtained at the end of an entire block, the step factor was reduced to 1.189. Then, the number of reversals was reset and the procedure continued until an even number of reversals, greater than or equal to 4, was obtained at the end of a block. Finally, thresholds were estimated taking the geometric mean of reversals obtained at the end of one (or more) entire block(s). In practice, four, six, eight, or ten reversals were used.

Listeners used a computer mouse to enter their answers on a graphical interface visible on a screen outside the booth. Before each test session, listeners were familiarized with the task by ten trials where the cumulative delay of the target interval took pseudo-random values between 0 and 40 ms. Visual feedback was given during the familiarization session by displaying a green square on the screen after a correct answer and a red square after a wrong answer. No feedback was provided during the test session.

A measurement reached saturation when ten successive incorrect answers were provided with a dT value equal to 40 ms. Saturated measurements were assigned a threshold of 40 ms. If more than 50% of the measurements saturated, the listener's data were not included in the analyses. All data were kept in the following experiments.

2. Subjective procedure

The subjective experiment consisted of presenting sequences of ABA triplets to listeners (van Noorden, 1975). In case of segregation, the listeners would hear two regular streams that differ in tempo. In case of integration, they would hear a single galloping rhythm. Stimuli were identical to those used in the objective experiments but consisted of one interval of eight triplets. At the end of the sequence, listeners had to indicate if they heard a single rhythm (i.e., integration) or two rhythms (i.e., segregation). They were instructed to pay attention to rhythm instead of perceived differences in timbre between stimuli.

B. Stimuli and conditions

Stimuli A and B were bursts of speech-shaped noise (SSN) convolved with impulse responses. SSNs were stationary noises with a spectrum similar to the long-term spectrum of speech, so they were approximately flat from 0 to 1000 Hz and then decreased by 20 dB per octave (ANSI, 1989). Ten different SSNs were used to average out spectral peculiarities associated with the choice of a particular SSN. Ten AB sequences were synthesized, each using one SSN for all A and B stimuli in the sequence. The impulse responses were measured either in an anechoic chamber [head-related impulse response (HRIR)] or in two rooms (room-HRIR). The room-HRIRs were a subset of those used by Lavandier *et al.* (2012). The recording position was kept constant, whereas the source was moved within each room. The HRIRs were measured by Gardner and Martin (1995) in an anechoic room. Only the left channels (arbitrarily chosen) of the room-HRIRs and HRIRs were used. Stimuli were sampled at 44.1 kHz and presented diotically using a LynxTwo-B soundcard through Sennheiser HD650 headphones at 70 dB sound pressure level in a double-walled sound-attenuated booth.

Since the room-HRIRs were measured at different distances, the size of their reverberation tails varied. To consider only stationary signals, ten 700-ms SSNs were convolved with the room-HRIRs. A temporal window with 12.5-ms raised-cosine on- and off-sets was applied in the middle of the convolved SSNs, leading to stationary noises of 150 ms, which were equalized in root-mean square (RMS) power. Since the HRIRs were anechoic the convolved SSNs were stationary. Ten SSNs of 138.4 ms were convolved with the 11.6-ms HRIRs, leading to 150-ms stimuli. The same cosine-window and RMS equalization were used. In the objective experiments, the sequences consisted of 12 AB pairs and lasted for 4.5 s [24×0.15 (stimuli duration) $+ 23 \times 0.04$ (silence duration)]. In the subjective experiment, the sequences consisted of 8 ABA triplets and also lasted for 4.5 s.

In each objective experiment, a "colocated" and a "separated" condition were tested five times. The designation

(colocated/separated) implies that A and B were produced using impulse responses measured either at a single position or at two different positions, respectively. If the spectral differences associated with a difference in position could enhance streaming, higher thresholds were expected in the separated conditions. Since ten AB pairs were used in each experiment, the subjective experiment consisted of 60 stimulus evaluations (2 conditions × 3 experiments × 10 pairs).

C. Listeners

Each objective experiment involved fourteen listeners and was divided in two one-hour sessions per listener. Nine of the participants participated in two different experiments. Fourteen listeners participated in the subjective experiment, which lasted about 15 min. One of them previously participated in the objective experiments. All listeners self-reported normal hearing and were paid an hourly wage for their participation.

D. Experiments

The aim of Experiment 1 was to determine whether obligatory streaming could be induced by spectral differences associated with two positions in a lecture hall. For the separated condition, A and B were synthesized with the impulse responses recorded with a source located at 0.65 m and + 25° (A) and 10 m and 0° (B) from the listener. For the colocated condition, A and B were synthesized with the impulse response recorded with a source placed at 0.65 m and +25° from the listener.

Experiments 2 and 3 tested whether the coloration associated with only the head and mostly the room, respectively, was sufficient to induce obligatory streaming. Experiment 2 used anechoic impulse responses measured in free-field at different azimuths (0° for A and B in the colocated condition, +30° for A and −30° for B in the separated condition). This configuration allowed us to investigate the influence of head coloration in isolation, simulating the extreme cases of non-reverberant rooms, near-field sources, or outdoor environments. Experiment 3 used stimuli synthesized from impulse responses sharing the same azimuth in a meeting room (0.65 m/0° for A and B in the colocated condition, 0.65 m/0° and 1.25 m/0° for A and B in the separated condition). Using positions with the same azimuth kept the head filtering approximately fixed, so that coloration was mostly influenced by the room.

The rooms and positions used in Experiments 1 and 3 were described in detail by Lavandier *et al.* (2012). Figure 1 illustrates the influence of coloration on the mean excitation patterns of the stimuli A and B (filled and dashed lines, respectively) in the separated condition of each experiment. Excitation patterns estimate the power level at the output of the auditory filters, after the cochlea frequency analysis (Moore and Glasberg, 1983). The figure shows a large degree of overlap between excitation patterns, especially in Experiments 2 and 3.

As only two conditions were contrasted in each experiment (A and B identical or A and B different), an additional subjective experiment was designed to investigate whether

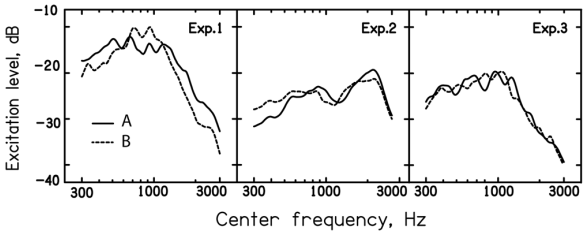

FIG. 1. Mean excitation patterns of the A (filled lines) and B (dashed lines) stimuli in the separated condition for each experiment. Each curve is the average of the excitation patterns of the ten stimuli used in the experiment. The stimuli were synthesized with impulse responses recorded for sources located at 0.65 m/+25° (A) and 10 m/0° (B) in a lecture hall for Experiment 1, +30° (A) and −30° (B) in an anechoic room for Experiment 2, and 0.65 m/0° (A) and 1.25 m/0° (B) in a meeting room for Experiment 3.

the differences in thresholds measured in Experiments 1–3 resulted from a difference of segregation or from an effect of dissimilarity between stimuli. For example, Grose *et al.* (2001) showed that a spectral difference between leading and trailing stimuli can affect gap detection, and it is possible that this kind of effect influenced the current results. Since the rhythmic discrimination procedure is an indirect measure of stream segregation, the subjective procedure was used to investigate if thresholds were consistent with perceived-segregation judgments (Micheyl and Oxenham, 2010).

III. RESULTS

Table I summarizes the mean thresholds obtained in Experiments 1–3 for each condition, and the corresponding percentage of segregated percepts measured in the subjective experiment. For Experiment 1, the mean thresholds for the colocated and the separated condition were 8 and 13.5 ms, respectively. For Experiment 2, they were 15.4 and 18.8 ms, respectively, and for Experiment 3, they were 11.1 and 14.6 ms, respectively. Table I shows the standard deviations (SD) and the standard errors (SE) of each threshold. The log values of the thresholds for each experiment were assessed using matched-pair t-tests. The main effect of condition was significant in the three experiments ($p < 0.05$ in each case).

The results of the subjective experiment show that in the separated conditions, listeners largely perceived two rhythms (87.9%, 86.4%, and 83.6% of the time for Experiments 1, 2, and 3, respectively). Conversely, listeners rarely perceived two rhythms in the colocated conditions (6.4%, 5.0%, and 6.4% of the time for Experiments 1, 2, and 3, respectively).

IV. DISCUSSION

In Experiments 1–3, the thresholds were significantly lower in the colocated condition compared to the separated condition in which there were spectral differences between A and B. This result suggests that the streams were significantly more segregated when introducing head and/or room coloration. This inference was confirmed by the subjective experiment for which segregation was largely perceived when stimuli were synthesized from impulse responses measured at different positions.

TABLE I. Results of the subjective experiment associated with the mean temporal discrimination thresholds, standard errors (SE) and deviations (SD) obtained in Experiments 1–3. Each of the three objective experiments involved 14 listeners. For the subjective experiment, 14 listeners judged the ten AB pairs of each condition of Experiments 1–3. Thus, the results presented are the averages of 140 responses and correspond to the number of times listeners heard two rhythms (i.e., two streams, so a segregated percept) in percentage.

	Colocated condition		Separated condition	
	Threshold	Segregation	Threshold	Segregation
Exp. 1	8.0 ms (SE = 1.15; SD = 1.7)	6.4%	13.5 ms (SE = 1.13; SD = 1.6)	87.9%
Exp. 2	14.5 ms (SE = 1.15; SD = 1.7)	5.0%	18.3 ms (SE = 1.13; SD = 1.6)	86.4%
Exp. 3	11.1 ms (SE = 1.17; SD = 1.8)	6.4%	14.6 ms (SE = 1.14; SD = 1.6)	83.6%

The spectral differences tested in the present study were of only a few dB (Fig. 1). They were smaller than what had been tested in previous streaming studies, which used different bandwidths with limited overlap to investigate the influence of spectral differences (Hartmann and Johnson, 1991). Nevertheless, the spectral differences tested in the present study were sufficient to produce obligatory streaming. These results are in agreement with those obtained by Middlebrooks and Onsan (2012) who parametrically measured stream segregation in the free field. They showed that voluntary streaming could be obtained in the median plane where head filtering induce subtle spectral changes which depend on position.

The study of Gaudrain *et al.* (2008) on vowel segregation due to fundamental frequency differences suggested that not all spectral cues are effective in inducing segregation of speech signals. They found that under certain conditions, vowels grouped together even if there were spectral differences between them due to differences in formant positions. It would be interesting to consider which spectral cues are relevant and useful for speech segregation. In the present work, stationary noises whose spectral differences were only of a few dB were studied. It is likely that these differences are small compared to the spectral variability of speech. However, these differences are constant across time since they are associated with a given position in space. In a cocktail-party context, if it can be assumed that the spatial configuration of speakers and listeners remains sufficiently constant over a given period of time, the regularities of the spectral differences could be exploited by the auditory system. Thus, the consistency of spatial differences could be relevant for segregation of competing voices.

Gaudrain *et al.* (2012) showed a significant correlation between stream segregation and intelligibility, and Martin *et al.* (2012) showed that modest spectral differences due to different locations in the median plane can improve speech segregation. The stimuli used in the present study did not contain large spectral variations like those found in real speech, and the sequences used consisted of regular repeated units unlike speech. However, the present study suggests that reverberation (through coloration) can favor the obligatory segregation of sounds coming from sources placed at different positions. The results could be extended to speech stimuli in the future, for example examining the effects of differences in coloration on simultaneous sentence perception in cocktail-party situations.

ACKNOWLEDGMENTS

The authors would like to thank the listeners who took part in the experiments, Christian Füllgrabe, Laurent Demany, and Brian Moore for their helpful comments. This work was supported by an institutional grant from the LabEX CeLyA ("Centre Lyonnais d'Acoustique," ANR-10-LABX-60) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

ANSI (**1989**). ANSI S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), Chap. 1, pp. 1–45.
Gardner, W. G., and Martin, K. D. (**1995**). "HRTF measurements of a KEMAR," J. Acoust. Soc. Am. **97**, 3907–3908.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2008**). "Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation," J. Acoust. Soc. Am. **124**, 3076–3087.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2012**). "The relationship between concurrent speech segregation, pitch-based streaming of vowel sequences, and frequency selectivity," Acta Acust. Acust. **98**, 317–327.
Grimault, N., and Gaudrain, E. (**2006**). "The consequences of cochlear damages on auditory scene analysis," Curr. Top. Acoust. Res. **4**, 17–24.
Grose, J. H., Hall, J. W., Buss, E., and Hatch, D. (**2001**). "Gap detection for similar and dissimilar gap markers," J. Acoust. Soc. Am. **109**, 1587–1595.
Hartmann, W. M., and Johnson, D. (**1991**). "Stream segregation and peripheral channeling," Music Percept. **9**, 155–183.
Houtgast, T., Steeneken, H., and Plomp, R. (**1980**). "Predicting speech-intelligibility in rooms from the modulation transfer-function. 1. General room acoustics," Acustica **46**, 60–72.
Lavandier, M., and Culling, J. F. (**2008**). "Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer," J. Acoust. Soc. Am. **123**, 2237–2248.
Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and Makin, S. J. (**2012**). "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," J. Acoust. Soc. Am. **131**, 218–231.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G., and Brungart, D. S. (**2012**). "Spatial release from speech-on-speech masking in the median sagittal plane," J. Acoust. Soc. Am. **131**, 378–385.
Micheyl, C., and Oxenham, A. J. (**2010**). "Objective and subjective psychophysical measures of auditory stream integration and segregation," J. Assoc. Res. Otolaryngol. **11**, 709–724.
Middlebrooks, J. C., and Onsan, Z. A. (**2012**). "Stream segregation with high spatial acuity," J. Acoust. Soc. Am. **132**, 3896–3911.
Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. **112**, 2074–2085.
van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tone sequences," Ph.D. thesis, University of Technology, Eindhoven.

When considering realistic configurations, the experiments described above present a major limitation since they focused on monaural processes. In fact, using a diotic presentation of the stimuli does not illustrate the perception of a normal hearing listener who receives information from his/her two ears. In order to get a step closer to a realistic configuration, the following study introduced progressively the binaural cues to investigate the potential benefice provided by having a second ear.

2 Sequential streaming, binaural cues and lateralization

2.1 Abstract - Résumé

Abstract

The second question raised in this chapter concerns the binaural listening. Spatial differences induce both differences in spectral coloration and in interaural level and time differences (ILD and ITD). The aim of this second study was to investigate to what extent binaural cues could influence stream segregation.

In a first experiment, the influence of the binaural cues was assessed introducing progressively ILD and ITD. Concerning the level differences, two different conditions were tested, in order to investigate the monaural intensity variations across time at each ear on one hand and the interaural level differences on the other hand. The thresholds in a rhythmic discrimination task using broadband noises were significantly higher for stimuli with ILD and ITD meaning that the auditory system could rely on the interaural cues to segregate competing streams. The results also suggested that the monaural intensity variations across time at each ear were more relevant for stream segregation than the interaural level variations.

Besides, interaural differences allow the listener to lateralize the coming sounds. A second experiment was run to investigate whether the influence of ITD was due to the change of interaural differences *per se* and/or to the corresponding differences in lateralization. The underlying aim was to understand if the auditory system exploits the low level information (i.e., based on the acoustical differences across the ears, the ITD *per se*) and/or the higher level information (i.e., based on the interpretation of a difference in perceived positions associated with ITD). To do so, the stimuli were split into high- and low- frequency bands which were presented either with a consistent ITD across frequencies (leading to a clear lateralization) or with ITDs of equal magnitude but opposite signs (leading to a blurred percept of lateralization).

The results of this experiment showed first that the percept of lateralization helped segregation when the lateralization was salient enough. Besides, the results also suggest that ITD *per*

se could favour segregation up to a ceiling magnitude. These results indicate that the auditory system can rely on interpreted information of the acoustical parameters, but also that those acoustical parameters can be directly exploited to some extent.

Résumé

La seconde question abordée dans ce chapitre concerne le caractère binaural d'une écoute spatialisée. Les différences spatiales entraînent non seulement de la coloration spectrale, mais également des différences interaurales de niveau (ILD) et de temps d'arrivée (ITD) aux deux oreilles. L'objectif de cette seconde étude était de déterminer dans quelle mesure l'introduction d'indices binauraux pouvait influencer la ségrégation de flux.

Dans une première expérience, l'influence des indices binauraux a été étudiée en introduisant progressivement l'ILD puis l'ITD. Pour les différences de niveau, deux conditions différentes ont été testées de manière à évaluer l'influence d'une variation monaurale d'intensité au cours du temps à chaque oreille d'une part, et la différence interaurale de niveau d'autre part. Les résultats d'une mesure de discrimination rythmique ont montré que les différences interaurales pouvaient être utilisées par le système auditif pour séparer les sources sonores. En particulier, les résultats ont indiqué une influence significative de l'ITD. Ils suggèrent également que les variations d'intensité au cours du temps à chaque oreille seraient plus pertinentes pour la ségrégation que les différences interaurales de niveau.

D'autre part, les différences interaurales permettent la latéralisation de sources sonores. Une seconde expérience a été menée de manière à comprendre si l'influence de l'ITD était due à la différence interaurale *per se* ou à la différence de position perçue correspondante. L'objet de cette expérience visait à étudier si le traitement de l'information spatiale pour la ségrégation était un processus de bas niveau (i.e., basé sur les différences acoustiques entre les deux oreilles, l'ITD *per se*) ou alors un processus de plus haut niveau (i.e., basé sur les différences de position associées à une différence d'ITD). Les stimuli ont été filtrés et présentés soit avec ITD consistant sur deux bandes de fréquence (hautes et basses fréquences), ce qui conduit à un percept clairement latéralisé, soit avec un ITD d'égale amplitude mais de signe opposé sur ces deux bandes de fréquence, ce qui conduit à un percept flou de latéralisation.

Les résultats de cette expérience ont montré d'une part que la position perçue pouvait induire de la ségrégation lorsque le percept de latéralisation est suffisamment saillant. Les résultats montrent également que l'ITD *per se* peut être exploité par le système auditif pour séparer des flux jusqu'à une valeur seuil. Ces résultats suggèrent que le système auditif peut

se baser sur une information interprétée des paramètres acoustiques mais que ces paramètres peuvent également être directement exploités dans une certaine mesure.

2.2 Article submitted to J. Acoust. Soc. Am.

2.2.1 Abstract

The present study assessed the influence of spatial cues on stream segregation. Experiment 1 investigated whether interaural time and level differences (ITDs and ILDs), introduced separately, could enhance auditory streaming. The thresholds in a rhythmic discrimination task using speech-shaped noises (SSNs) were significantly higher for stimuli with ITDs and ILDs, indicating that interaural differences could favor segregation. The results also suggested that the monaural intensity variations across time at each ear were more relevant for streaming than the interaural level differences. Experiment 2 investigated whether the influence of ITD was due to the difference of interaural difference and / or to the corresponding difference in lateralization. SSNs were split into high- and low- frequency bands which were presented either with a consistent ITD across frequency, leading to a clear sound lateralization, or with ITDs of equal magnitude but opposite signs, leading to blurred perceived positions. A significant lower threshold was observed in the inconsistent condition compared to the consistent condition, indicating that perceived position could help segregation when the percept of lateralization is sufficiently salient. The results also suggested that ITD *per se* might favor segregation up to a ceiling magnitude.

Pacs Numbers: 43.66.Mk, 43.66.Qp, 43.66.Pn

2.2.2 Introduction

In a context of competing sound sources, the main objective of auditory scene analysis (ASA; Bregman, 1990) consists in grouping sound events coming from a same source (i.e., integrated percept) and separating those coming from different sources (i.e., segregated percept). Many sound properties can influence stream formation, such as differences in fundamental frequency, in temporal envelope, in spectrum or in lateralization (see Moore and Gockel (2002, 2012) for reviews). The main purpose of the present study was to focus on the potential influence of spatial differences on stream segregation.

Sounds coming from different locations in space present both monaural and binaural differences. First, the source spectrum produced at each ear depends on the listener and source positions. Indeed, sound is submitted to several frequency-dependent reflections during its

propagation, and the sound at the ears result from the addition of the direct sound an a given combination of filtered reflections (Collin and Lavandier, 2013; Flanagan and Lummis, 1970; Larsen *et al.*, 2008). It has been shown that the slight monaural spectral differences associated with head and room coloration can induce segregation of broadband noises (David *et al.*, 2014). Sound coming from different locations present also binaural differences: interaural time and level differences (ITDs and ILDs). The present study first investigated whether those binaural cues could strengthen broadband sound segregation.

Previous studies have shown that ILD could help segregate sequences of pure tones. For example, Hartmann and Johnson (1991) showed better results in a melody recognition task when target and interleaved masker differed in ILD by 8 dB. Sach and Bailey (2004) showed similar results using a task where listeners had to identify a target rhythm interleaved with arrhythmic masked tones (referred to as a rhythmic masking release task, RMR). Indeed, listeners reached higher performance when target and interferer had different ILDs (4 and 0 dB, respectively). Besides, when two alternate sounds differ in ILDs, they also differ monaurally, in intensity across time at each ear. These monaural intensity differences could help segregation (Plomp, 1964). For example, Stainsby *et al.* (2004); van Noorden (1975) showed that monaural level differences larger than 5 dB could induce segregation. The present study assessed whether level differences at each ear and / or across ears were useful cues in the streaming process (Experiment 1).

The influence of ITD on stream segregation is more contrasted in the literature. Some studies showed that segregation could be induced by providing an ITD difference between streams. Hartmann and Johnson (1991) showed that listeners reached higher performance in a pure tone melody recognition task when target and interferer differed in ITD ($\pm$ 500 μs). Sach and Bailey (2004) investigated the influence of ITD in a rhythmic masking release (RMR) task. Listeners reported a better identification accuracy when target and masker differed in ITD by 100 to 600 μs, for target and masker without ILD. Gockel *et al.* (1999) measured to what extent ITD could influence the threshold for detecting a change in F_0 of a complex tone. In some conditions, the target was preceded and followed by harmonic complexes temporally adjacent to the target sound (i.e., temporal "fringes"). The results showed that the impairment induced by the fringes was reduced when they were shifted in perceived position due to ITD away from the target. Darwin and Hukin (1999) also demonstrated that ITD could influence sequential grouping of speech sounds: listeners tended to group a target word with a sentence more often if they shared the same ITD. Finally, Kidd *et al.* (2008) presented two sequences of speech words to the listeners. The target and masker words were interleaved (odd number words and even number words, respectively), and the listeners had to track the target words.

The results showed that the percentage of correct target-words identification increased when a difference in apparent location, induced by a difference in ITD (from ±150 and ±700 μs), was applied to the target words.

Other studies showed only a weak or no effect of ITD on stream segregation. Boehnke and Phillips (2005) found no significant improvement in segregation for broadband noises differing in ITDs in a gap discrimination task. Stainsby *et al.* (2011) showed that ITD could help segregating harmonic complex tones in a rhythmic discrimination procedure (Roberts *et al.*, 2002), but only for ITD values outside the physiological range. Finally, Füllgrabe and Moore (2012) replicated the experiment of Stainsby and colleagues with pure tones and ITDs below 500 μs. They found only a weak effect of ITD on stream segregation.

There is no clear agreement between those studies investigating the influence of ITD on stream segregation. It seemed that experiments where listeners had to segregate streams (i.e., experiments measuring voluntary streaming for which task performance is favored by segregation) were markedly influenced by ITD (Darwin and Hukin, 1999; Gockel *et al.*, 1999; Hartmann and Johnson, 1991; Kidd *et al.*, 2008; Sach and Bailey, 2004). This effect was obtained with different types of stimuli such as pure tones, harmonic complex tones and speech sounds. Conversely, ITDs within the physiological range seemed to have only a weak effect on stream segregation in tasks measuring obligatory streaming (i.e., situations where segregation occurs even when the listeners tries to fuse the streams, task performance being impaired by segregation). For example, the results of Füllgrabe and Moore (2012) using pure tones did not show a strong effect of ITD. This result might be explained first by the fact that pure tones provide less binaural information than broadband stimuli as they only contain one frequency component. Second, if the period of the pure tone is inferior to twice the considered ITD, the perceived position will be ambiguous since the phase corresponds to more than one entire period. For instance, if the period is 250 μs (i.e., $f = 4000$ Hz), an ITD of 500 μs induces a delay of two whole circles, so leads to two possible phases: 0 or 360° (Moore, 2007). Thus, for low frequency pure tones, ITD provides unambiguous information about sound lateralization, but for higher frequencies, ITD can lead to an ambiguous perceived position for pure tones. In the study of Boehnke and Phillips (2005), sound sequences of only 330 ms were used. This might have been too short for the build-up of segregation to occur (Anstis and Saida, 1985; Roberts *et al.*, 2008). The present study evaluated whether realistic ITDs could influence obligatory stream segregation using 4.8-s sequences of broadband noise bursts (Experiment 1).

Binaural cues allow for sound localization. In the horizontal plane, the sound localization accuracy is largely based on the spatial dependence of interaural difference cues. Thus, if binaural cues could facilitate segregation, it might be explained by a difference in binaural

cues *per se* (ILD and / or ITD) but also by the corresponding difference in perceived position. Another aim of the present study was to assess independently the influence of ITD and the corresponding perceived position on the segregation of broadband noise bursts (Experiment 2).

2.2.3 General Methods

Rhythmic discrimination paradigm

The rhythmic discrimination procedure introduced by Roberts *et al.* (2002) was used in the present study to evaluate how binaural cues and the associated perceived positions could enhance obligatory streaming. This procedure has been widely used to objectively measure obligatory streaming (Füllgrabe and Moore, 2012; Roberts *et al.*, 2008; Stainsby *et al.*, 2011, 2004; Thompson *et al.*, 2011). It involved the presentation of two intervals of alternate noise bursts [A-B-A-B...]. In the target interval, the first six AB pairs were regularly spaced by 40 ms (i.e., the Bs were placed at the temporal midpoint between two consecutive As). The Bs were then progressively delayed by equal steps for the next four pairs. Thus, the seventh B was delayed by δT, and the three next Bs were delayed by $2\delta T$, $3\delta T$ and $4\delta T$, respectively. Finally, the cumulative delay ΔT ($\Delta T = 4\delta T$) was kept constant for the last two pairs. In the reference interval, the silence duration between consecutive stimuli was always 40 ms, leading to a regular rhythm. The silence duration between two intervals was set to 1 s. Figure I.1 illustrates the rhythmic discrimination paradigm. The listener's task was to identify the target sequence with the delayed Bs among the two intervals.

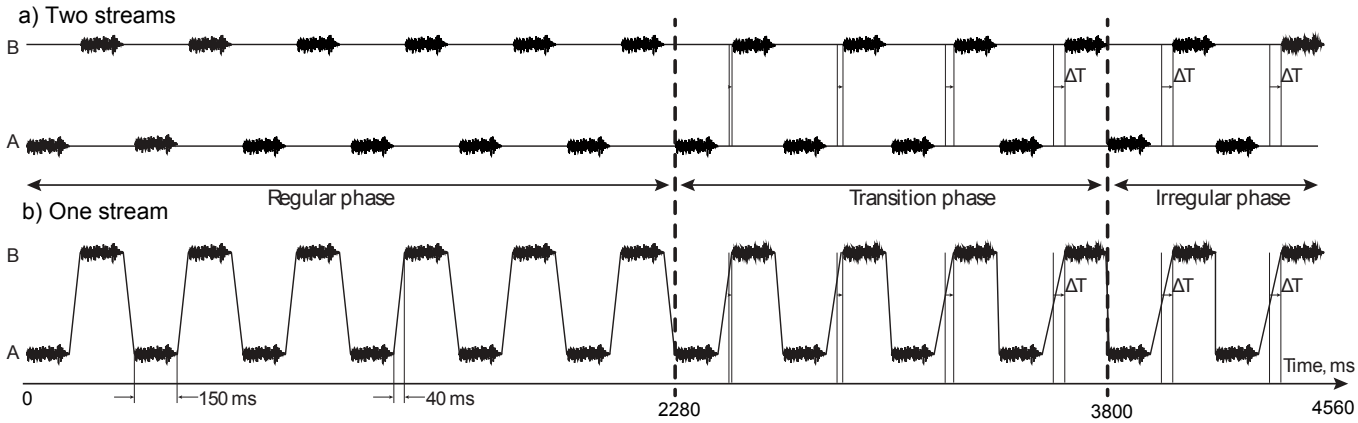

Figure I.1 – Schematic representation of irregular sequences presented for the rhythmic discrimination task. The sequences consisted of twelve pairs of alternate noise bursts. In the irregular sequence, the B bursts were initially positioned at the exact temporal midpoint between two successive A bursts (regular phase), then they were progressively delayed in the transition phase. In the irregular phase the cumulative delay applied to the B bursts was kept constant. In the regular sequence (not plotted), the B bursts remained at the temporal midpoint between the A bursts for the entire sequence. The irregular sequence could lead to two different percepts depending on the segregation state of the streams A and B [segregated in the top panel (a) and integrated in the bottom panel (b)].

According to this paradigm, the perceived rhythm of the target interval depends on whether

the listener hears a single stream or two segregated streams (see Figure I.1). The delay applied to the Bs is more easily detectable when a single stream is heard because successive time intervals [A-B] and [B-A] are compared and δT is not negligible compared to these interval durations. Conversely, the delay applied to the Bs is more difficult to detect when the streams are segregated because successive [B-B] intervals are compared and these intervals are longer compared to δT. Thus, the rhythmic irregularities are better detected when the percept is integrated and segregation impairs task performance.

Global characteristics of the stimuli

Bursts of speech-shaped noise (SSN) were used to generate the stimuli A and B. SSNs were stationary noises with a spectrum similar to the long-term spectrum of speech, so approximately flat from 0 to 1000 Hz and then decreasing by 20 dB per octave (ANSI S3.6, 1989). The rhythmic discrimination procedure requires rapid sequences and short stimuli duration (van Noorden, 1975). For example, Roberts *et al.* (2002) used 60-ms stimuli, separated by 40 ms of silence. Thus, short samples of SSN were extracted from a long SSN. However, the spectrum of a short excerpt is highly dependent on which segment of the long noise is chosen. For example, Figure I.2 shows the excitation patterns of a 60-sec SSN and of two different SSN of 150 ms taken from this long-duration SSN. The patterns of the short duration SSNs differed from each other and from the pattern of the long-duration SSN. A compromise was reached by generating stimuli lasting 150 ms, which led to less spectral variability than stimuli lasting 60 ms. In order to further limit the influence of the choice of a particular SSN, ten different SSNs were used instead of a single one. Thus, as in David *et al.* (2014), ten AB pairs were synthesized.

Adaptive procedure

Thresholds for detecting anisochrony were estimated with a two-interval, two alternative forced-choice method (Levitt, 1971). The delay applied to the B bursts was adapted according to a three-down one-up rule and varied on a logarithmic scale. This method determines the 79.4% of correct responses on the psychometric curve. It is worth noting that higher thresholds suggest a greater tendency for the streams to be heard as segregated. The maximum delay was 40 ms, which corresponded to the maximum delay without overlap between two consecutive stimuli. The initial value of the delay was 28.28 ms. A measurement reached saturation when ten consecutive incorrect answers were provided with a ΔT of 40 ms. Saturated measurements were assigned a threshold of 40 ms (Roberts *et al.*, 2002). If more than 50 percent of the

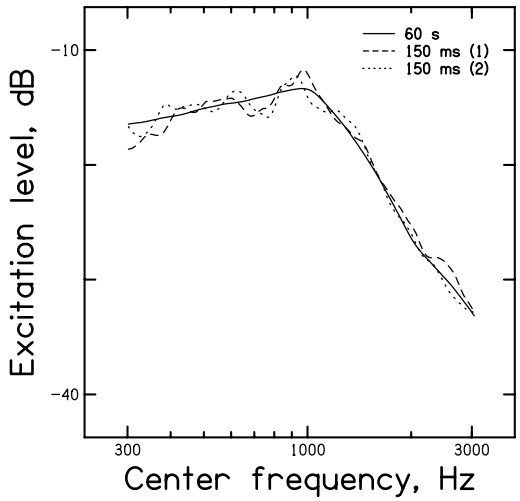

Figure I.2 – Excitation patterns of a long speech-shaped noise (SSN) of 60 s and of two different SSN excerpts of 150 ms extracted from this long SSN. The patterns of the short-duration SSNs depend on the particular time epoch where the excerpt was extracted.

measurements saturated for a given listener, his / her data were discarded from the analysis (Devergie *et al.*, 2011). This saturation condition occurred for one participant in Experiment 2, thus the corresponding results were not considered in the analysis.

Since a set of ten pairs of stimuli was used, each pair had to have the same probability of being presented to the listener, thus the estimation of thresholds was slightly different from the one used by Roberts *et al.* (2002). Each run (i.e., each adaptive staircase) was divided into successive blocks of ten trials. For each trial, one SSN segment was randomly chosen without replacement among the set of available SSNs (ten to start with) to synthesize A and B. For the next trial, a different SSN was randomly chosen without replacement from the remaining samples and so on for the next trials until the ten samples had been used.

When the listeners gave three correct answers, the delay decreased by a factor of 1.414, and it increased by the same factor when listeners gave one wrong answer. If at least two reversals were obtained at the end of a block of ten trials, the step factor was reduced to 1.189; otherwise, it was kept constant for another block until at least two reversals were obtained. Once the step factor was reduced, the number of reversals was reset, and the procedure continued until an even number of reversals greater than or equal to four was obtained at the end of a block. Finally, thresholds were estimated using the geometric mean of the reversals from the entire blocks which used the smaller step factor. As the number of reversals which could be obtained in one block varied from zero to four, thresholds were calculated at the end of each run with four, six, eight or ten reversals. With this respect, the present procedure was at least as accurate

as the procedure of Roberts *et al.* (2002) where thresholds were estimated using four reversals.

The adaptive procedure was similar in Experiments 1 and 2, only the tested stimuli differed. Experiments were run in a sound attenuated booth. Listeners used a computer keyboard and mouse to enter their answers on a graphical interface visible on a screen placed outside the booth. One run lasted approximately ten minutes and five runs were completed for each condition in each experiment. The experiments were divided in five one-hour sessions and all the conditions were tested once in each session. In order to get familiar with the task, participants did twenty trials before each session, where ΔT took pseudo-random values between 0 and 40 ms. For this familiarization session, diotic SSN samples of 150 ms were presented, and visual feedback was given by displaying a green or a red square after a correct or a wrong answer. No feedback was given during the test session.

2.2.4 Experiment 1: Realistic ITDs and ILDs

Rationale

The aim of Experiment 1 was to investigate whether obligatory streaming could be enhanced by binaural cues. ITDs and ILDs corresponding to real positions in an anechoic room were introduced to assess to what extent each cue is useful in the streaming process. Four conditions were tested. Condition 1 was the reference condition where A and B were identical and diotic. In Condition 2, stimuli were diotic but monaural spectral differences (induced by head coloration) were introduced. ILDs were introduced in Condition 3, and stimuli containing both ILDs and ITDs in Condition 4.

Stimuli

Spatial information was conveyed to the stimuli A and B by convolving the bursts of SSN with head-related impulse responses (HRIRs). The HRIRs were measured by Gardner and Martin (1995) using loudspeakers mounted at a distance of 1.4 m from a KEMAR mannequin (Knowles Electronic model DB-4004) in an anechoic chamber. SSNs of 700 ms were convolved with the HRIRs, then a window with 12.5-ms raised-cosine on- and off-sets was applied in the middle of the convolved SSNs, leading to stationary bursts of 150 ms duration.

Figure I.3 illustrates the configuration of Experiment 1. The HRIRs used to generate the stimuli were recorded in the left ear of a mannequin with sources at three positions (L, R and F). L, R and F corresponded to a source located at 30° to the left-hand side of the head (+30°), at 30° to the right-hand side (-30°) and at 0°, respectively. For the chosen positions L and R,

the broadband ILD and ITD were 6 dB and 272 μsec. X, Y and Z on Figure I.3 represent the signals (i.e., the HRIR convolved with the burst of speech-shaped noise) received at the left ear for a source located in L, R and F, respectively. By symmetry, X, Y and Z also represent the signals received at the right ear for a source located in R, L and F, respectively. Thus, when X was sent to the left ear and Y was sent to the right ear, the source was perceived on the left (i.e., position L). Conversely, when Y was sent to the left ear and X was sent to the right ear, the source was perceived on the right (i.e., position R). When the two ears received the same signal (diotic presentation), whichever it was, the source was perceived in front (i.e., position F).

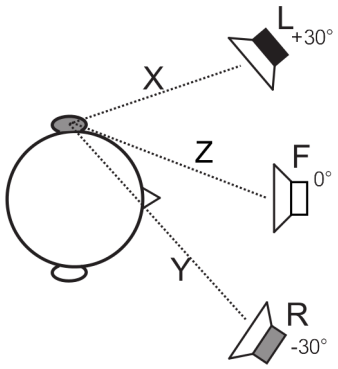

Figure I.3 – Configuration tested in Experiment 1. X, Y and Z correspond to the signals recorded in the left ear for a source placed at +30° (position L), -30° (position R) and 0° (position F), respectively.

Table I.1 presents the four conditions tested in Experiment 1 and describes which signal (X, Y or Z) was sent to each ear in each condition. In Condition 1, A and B were synthesized as coming from a source located in front of the listener. So, for both stimuli, the left and right ears received Z. In Condition 2, stimuli were also presented diotically so both ears received X (stimuli A) or Y (stimuli B). In Condition 3, stimuli were presented dichotically. The left ear received X and the right ear received Y (stimuli A) or the left ear received Y and the right ear received X (stimuli B). In this condition, all ITDs were eliminated in each frequency band by rotating all the FFT (Fast Fourier Transform) components of both X and Y into a null phase before computing an inverse FFT. In Condition 4, stimuli were presented dichotically as in Condition 3 except that ITDs were preserved.

Binaural signals were prepared beforehand for Conditions 1, 3 and 4. The left-ear and right-ear channels of these stimuli were divided by the mean RMS power level of the two channels. This equalization led to a mean level across the ears of 70 dB SPL, as measured with an artificial ear (Larson Davis AEC101 and 824; ANSI S3.7, 1995). Thus, in Conditions 3 and 4,

Tableau I.1 – Conditions tested in Experiment 1 with the corresponding signals presented at each ear. In condition 1, both ears received the same F-signal, so the source was perceived in front of the listener. In condition 2, both ears received the same alternation of X- and Y-signals, leading to a monaural intensity difference across time but no difference in perceived position (source perceived in front). For conditions 3 and 4, according to the symmetric configuration, when the X- and Y-signals were sent to the left and right ears, respectively, the source was perceived on the left-hand side of the listener. Conversely, when the X- and Y-signals were sent to the right and left ears, respectively, the source was perceived on the right-hand side of the listener. The asterisk in condition 3 indicates that ITDs were removed from the stimuli, all their spectral components were set into a null phase.

		Sequences				
C1: Reference	Left ear	F	F	F	F	...
	Right ear	F	F	F	F	...
C2: Diotic differences	Left ear	X	Y	X	Y	...
	Right ear	X	Y	X	Y	...
C3: ILD only	Left ear	X*	Y*	X*	Y*	...
	Right ear	Y*	X*	Y*	X*	...
C4: Binaural	Left ear	X	Y	X	Y	...
	Right ear	Y	X	Y	X	...

the stimuli original ILDs were preserved. In order to preserve the broadband level difference between the stimuli in Condition 2, diotic signals were prepared using the equalized binaural signals of Condition 4. So, stimuli A consisted of the left channel of the equalized signal of Condition 4 sent in both ears, and stimuli B consisted of the right channel of the equalized signal of Condition 4 sent in both ears.

A single couple of positions was used (+30° and -30°) so that the spectral differences at each ear between A and B were kept constant in Conditions 2, 3 and 4. Since David *et al.* (2014) showed that the spectral differences induced by head coloration could induce obligatory streaming, Conditions 2, 3 and 4 were supposed to show more segregation (i.e., higher thresholds) than Condition 1. Besides, as shown in Table I.1, the alternation of noise bursts [A-B-A-B...] in the sequences induced a difference between X and Y at each ear, as well as a difference between X and Y across the ears in Conditions 3 and 4. If Experiment 1 shows more segregation in Condition 3 and 4 compared to Condition 2, it would be due to the binaural cues.

Listeners

Experiment 1 involved fourteen listeners. They were students, aged between 20 and 27 yrs (ten females, mean age = 22 yrs, standard deviation SD = 2 yrs), they signed a general consent form before the experiment, and self-reported normal hearing. They were paid an hourly wage for their participation, and came for five one-hour sessions.

Results

Figure I.4 presents the geometric mean across listeners of the temporal discrimination thresholds measured in Experiment 1. From left to right, the bars correspond to the Conditions 1 to 4, and the error bars represent the geometric standard errors across listeners.

The log values of these thresholds were assessed using a one-way repeated-measures analysis of variance (ANOVA), which showed that the effect of tested condition was significant ($F(3,52)=17.11$, $p<0.001$). A post-hoc analysis (Bonferroni test) indicated that the mean threshold obtained in Condition 1 was significantly lower than those obtained in Conditions 2 to 4 ($p<0.0001$ in each case) and that the threshold was higher in Condition 4 compared to Conditions 2 and 3 ($p=0.0015$ and $p=0.0106$, for the comparison between Condition 4 and Conditions 2 and 3, respectively). There was no significant difference in thresholds between Conditions 2 and 3.

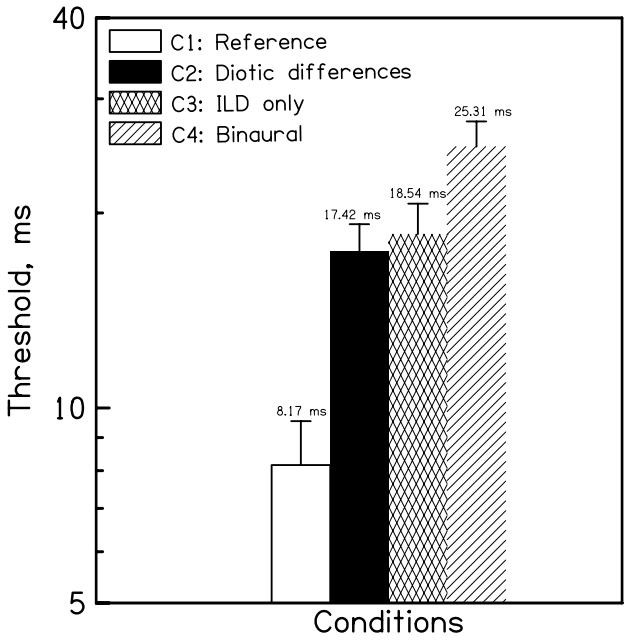

Figure I.4 – Mean thresholds in ms on a log scale with geometric standard errors across participants for detection of the delay of the B bursts in Experiment 1. From left to right the bars correspond to Conditions 1 to 4 (see Table I.1).

Discussion

The significantly lower threshold obtained in Condition 1 compared to Conditions 2 to 4 suggests that the task was easier when the sources were co-located (i.e., stimuli without either spectral differences nor binaural cues) than when they were spatially separated (i.e., stimuli with spectral differences, ILDs or ITDs). The rhythmic discrimination task is an indirect measure of obligatory streaming for which performance are impaired by segregation. Micheyl and Oxenham (2010) and David *et al.* (2014) showed consistency between this measure and perceived-segregation judgements. Thus, Experiment 1 showed a greater tendency to hear two streams when the sources were spatially separated than when sources were co-located.

The mean threshold obtained in Condition 1 was significantly lower compared to Condition 2. In this condition, the stimuli were presented diotically, and presented monaural spectral differences (including a broadband level difference and spectral differences induced by coloration) associated with the difference in spatial locations. These results are consistent with those obtained by David *et al.* (2014), since they observed obligatory streaming introducing only monaural spectral differences associated with coloration. The effect on segregation of the within-band spectral differences and of the broadband level difference seem to be additive, be-

cause the difference in thresholds induced by coloration was larger in the present study (the difference in thresholds between Condition 1 and 2 was 9.25 ms) compared to the one obtained by David *et al.* (2014) (3.40 ms) with the same stimuli without the broadband level difference. The present results are also in agreement with those obtained by Middlebrooks and Onsan (2012) who showed that voluntary streaming could be obtained in the median plane where head filtering induce subtle spectral differences which depend on position.

The mean threshold was significantly lower in Condition 1 compared to Condition 3. This result confirms that SSN bursts can be segregated when their ILDs are different, in agreement with Stainsby *et al.* (2004). In addition, there was no significant difference in mean threshold between Conditions 2 and 3. This suggests that adding the "binaural component" of ILD did not influence streaming. The monaural spectral differences induced by coloration at each ear are sufficient to explain the increase of discrimination thresholds and the corresponding improvement in segregation. Thus, listeners seemed to organize the coming sounds based on the spectral variations across time at each ear rather than on the spectral difference across ears.

The mean threshold was significantly higher in Condition 4 compared to Condition 3. This result suggests that ITDs significantly favored obligatory streaming between the sounds coming from competing sources which were spatially separated. The stimuli were generated using recorded HRIRs, so the ITDs were in the physiological range ($\pm$ 272 μs). According to this result, realistic ITDs could favor segregation whereas previous studies showed that ITD had no or only a limited influence on obligatory streaming (Füllgrabe and Moore, 2012; Stainsby *et al.*, 2011). This difference with the literature might be explained by the nature of the stimuli used: the broadband noises used in the present study might have lead to stronger binaural cues than the pure or complex tones used in the previous studies.

2.2.5 Experiment 2: ITDs versus lateralization

Rationale

Experiment 1 showed a significant effect of ITD on stream segregation which could be due to the interaural difference *per se* and / or to the associated perceived position. The aim of Experiment 2 was to investigate separately the potential influences of the binaural information and of the perceived position associated with this information. The distinction between interaural differences and the corresponding perceived positions might be useful to determine whether the signals were segregated before the localization of the corresponding sources (so based only on interaural differences) or after that sources were localized. Experiment 2 intended to determine if the two mechanisms were independent or if they can be both relevant

for stream segregation.

Stimuli

In Experiment 2, artificial ITDs were used instead of realistic ones in order to facilitate their manipulation across frequencies. The stimuli were spectrally divided into high- and low-frequency bands below and over 550 Hz. In Conditions 1 to 3, the two frequency bands of the stimuli had the same consistent ITDs, while in Condition 4 the ITDs were of opposite sign in each frequency band. When ITDs were consistent across the whole spectrum, perceived positions should be clearly identified. When ITDs were inconsistent between high and low frequencies, perceived position should be blurred (Edmonds and Culling, 2005).

Figure I.5 represents the conditions tested in Experiment 2. In Condition 1, the A and B bursts had the same ITD of 0 μs across the whole spectrum with the hypothesis that they would both be perceived as coming from the same direction: in front of the head. In Condition 2, A and B had consistent ITDs across the whole spectrum of 272 and -272 μs, respectively, which correspond to the broadband ITDs of Experiment 1. Condition 3 was identical to Condition 2 except that the ITD magnitude was increased to 500 μs. Finally, in Condition 4, the high-frequency band of stimuli A and the low-frequency band of stimuli B were presented with an ITD of 500 μs while the low-frequency band of stimuli A and the high-frequency band of stimuli B were presented with an ITD of -500 μs. In this condition, the amount of interaural differences was kept identical to Condition 3. So, if interaural differences are the main factor influencing stream segregation, one would expect the same thresholds in Conditions 3 and 4. However, the fact that the ITD was not consistent across frequency bands was hypothesized to result in reduced lateralization. So, if perceived position is the main factor for stream segregation, one would expect a significant higher threshold in Condition 3 compared to Condition 4.

The synthesis of the stimuli involved different steps. The SSNs of 10 s were first high-pass and low-pass filtered using fourth-order Butterworth high-pass and low-pass filters and a cutoff frequency of 550 Hz. The high and low stop frequencies were 592 and 507 Hz, respectively. This gap avoided to have an overlap between the two ITDs (Edmonds and Culling, 2005). The cutoff frequency was chosen first to have approximately the same energy at high and low frequencies, and second to be sure that ITD was usable in the two frequency bands (Feddersen *et al.*, 1957). A delay corresponding to the tested ITD was introduced in each frequency band. Then, the two parts of the spectrum were concatenated and an inverse FFT was performed. Finally, the stimuli were time-windowed in the middle of the 10-s signal using 12.5 ms on- and off-cosine ramps, and each left- and right-ear channels were equalized independently in RMS

to obtain stationary bursts of 150 ms at 70 dB SPL. This windowing led to stimuli presenting only ongoing ITDs. The fact that they did not present onset nor offset ITDs should not impair the results since ongoing delays are usually much more relevant in determining location than onset/offset delays (Buell and Hafter, 1991; McFadden and Pasanen, 1976).

Temporal discrimination thresholds were measured using the rhythmic discrimination procedure described in the General methods. The listeners also did a short subjective lateralization task (about 10 min long). They were asked to evaluate the perceived position of the different stimuli used in the rhythmic discrimination task, ten different As and Bs presented in a random order. The sequences played during the lateralization task corresponded to a single stream of the rhythmic discrimination task (the As or the Bs), and were generated by repeating a single stimulus twelve times, spaced by 190 ms. Twenty one-stream sequences were generated, ten using the As stimuli and ten using the Bs stimuli. Listeners could play the sequences as many times as they wanted. They had to draw on a protractor the perceived position of the corresponding sound source. The protractor was graduated from +90° (left hand side) to -90° (right-hand side) with 5° steps. Listeners had to judge the position of all the sequences twice. No specific indication was given, thus they could indicate either a precise position, a single or several area(s), more or less wide, from which direction(s) they perceived the sounds.

Listeners

Experiment 2 involved fourteen listeners, but the results of one participant had to be discarded because saturation was reached in more than 50% of the trials, as previously indicated. The thirteen remaining listeners were students, aged between 20 and 30 (six females, mean age = 26, standard deviation SD = 3), and signed a general consent form before the experiment. All listeners had audiometric thresholds of 20 dB hearing level or less in each ear at octave frequencies between 250 and 4000 Hz. They were paid an hourly wage for their participation, and they came for five sessions lasting roughly one hour.

Results

Figure I.6 shows the geometric mean across listeners of the temporal discrimination thresholds measured in Experiment 2. From left to right, the bars correspond to the Conditions 1 to 4, and the error bars represent the geometric standard errors across listeners.

The log values of these thresholds were assessed using a one-way repeated-measures analysis of variance (ANOVA), which showed that the effect of tested conditions was significant

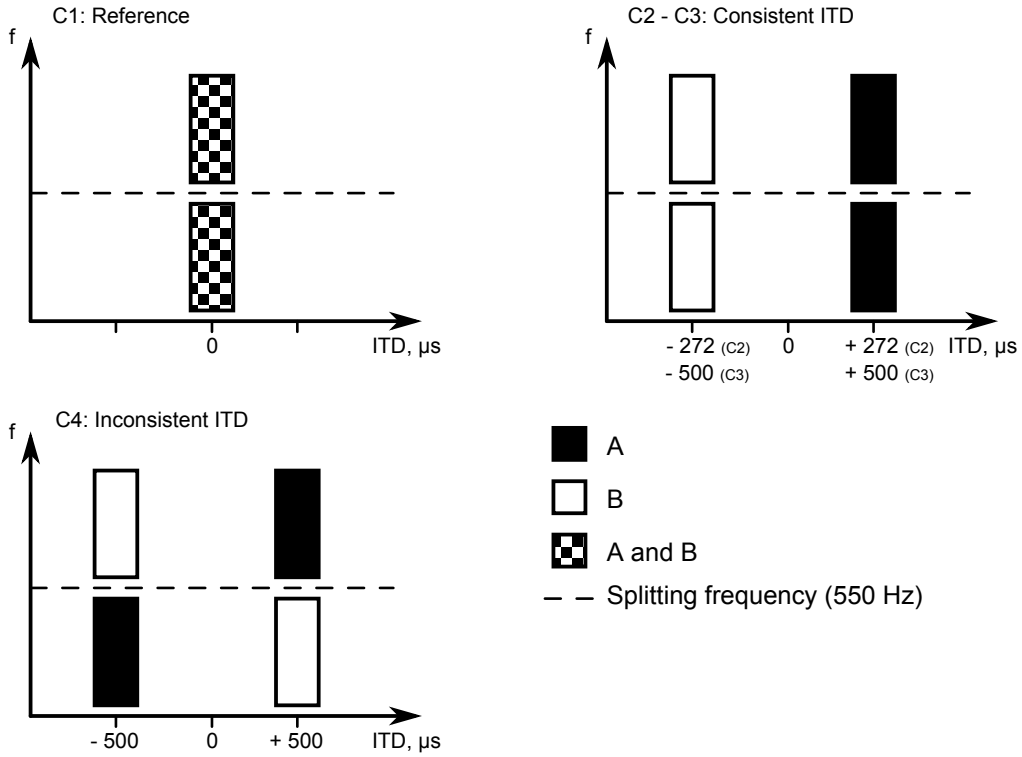

Figure I.5 – Conditions tested in Experiment 2. Stimuli were spectrally divided into high-and low-frequency bands, with a splitting frequency of 550 Hz. In Conditions 1 to 3, the ITD was consistent across frequency, so the two bands had the same ITDs: 0, 272 and 500 μs, respectively. In Condition 4, perceived position was blurred by manipulating the ITD independently in each frequency band (i.e., the high-frequency band of stimuli A and the low-frequency band of stimuli B were presented with a $+500\mu s$ ITD while the low-frequency band of stimuli A and the high-frequency band of stimuli B were presented with a $-500\mu s$ ITD).

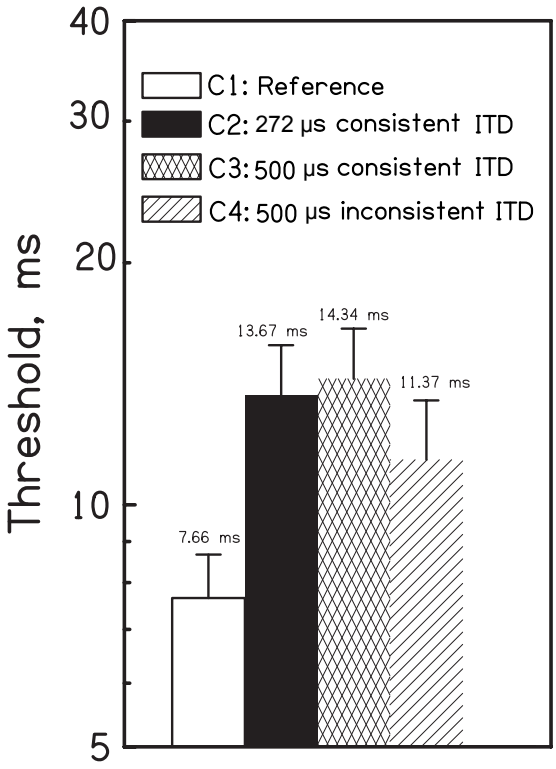

Figure I.6 – Mean thresholds in ms on a log scale with geometric standard errors across participants for detection of the delay of the B bursts in Experiment 2. From left to right the bars corresponded to Conditions 1 to 4 (see Figure I.5).

($F(3,48)=3.528$, $p<0.05$). A post-hoc analysis (Bonferroni test) indicated that the mean threshold obtained in Condition 1 was significantly lower than those obtained in Conditions 2 to 4 ($p<0.001$ in each case), and that the mean threshold obtained in Condition 3 was significantly higher than the mean threshold obtained in Condition 4 ($p=0.041$). There was no significant difference in threshold between Conditions 2 and 3, nor between Conditions 2 and 4.

In the subjective lateralization task, the hypothesis was made that listeners' responses followed one (or more) Gaussian distribution(s). If a listener perceived a single stimulus as coming from two different positions at a same time - which happened sometimes with the stimuli of Condition 4 - his/her response was fitted by two Gaussian distributions. The mean(s) and the spread(s) of the raw individual responses corresponded to the mean(s) and standard deviation(s) of the distribution(s). The individual distributions were then averaged across listeners. Figure I.7 presents the mean distributions of the responses as a function of the direction. The mean perceived positions of stimuli A and B correspond to the dotted grey and filled black lines, respectively. Some listeners (6 over 13) often lateralized the stimuli at very precise positions and even pointed at only one position. Thus, the standard deviation of their response was null, resulting in peaks in the mean distributions.

A Kolmogorov-Smirnov test showed that the mean distributions were normal in each condition ($p<0.001$ in each case). The parameters of these mean distributions were estimated using a maximum likelihood estimation (MLE) procedure. For the consistent 0 μsec-ITD (Condition 1, top panel), the mean distribution of the responses was centred at 0° (i.e., in front of the head), and the standard deviation (std) was equal to 4.6 and 4.8° for A and B, respectively. For the consistent $\pm$272 and $\pm$500 μsec-ITD (Conditions 2 and 3, second and third panels), mean distributions were centred at 52.7 and 59.2° (stimuli A) and at -62.4 and -62.1° (stimuli B). Std was equal to 9.3 and 8.2° (stimuli A) and 2.6 and 2.7° (stimuli B) for Conditions 2 and 3, respectively. Note that the MLE procedure seemed to under-estimate the standard deviation of stimuli B in these conditions. An explanation might be that the data were truncated at -90° and larger angle might be needed to obtain a better estimation. For the inconsistent 500 μsec-ITD (Condition 4, bottom panel), mean distributions were centred around -40 and 0° for A and B, respectively, and were flattened compared to the other distributions (std = 20 and 8.7° for A and B, respectively).

The values of these mean distributions were assessed using a one-way repeated-measures ANOVA, which showed that the effect of the tested conditions was significant ($F(7,14400)=193.3$, $p<0.001$). A post-hoc analysis (LSD) indicated that there was no significant difference between the mean distributions of A and B in Condition 1. The difference between the distributions of A and B was significant in Conditions 2, 3 and 4 ($p<0.0001$ for Conditions 2 and 3, and $p=0.015$

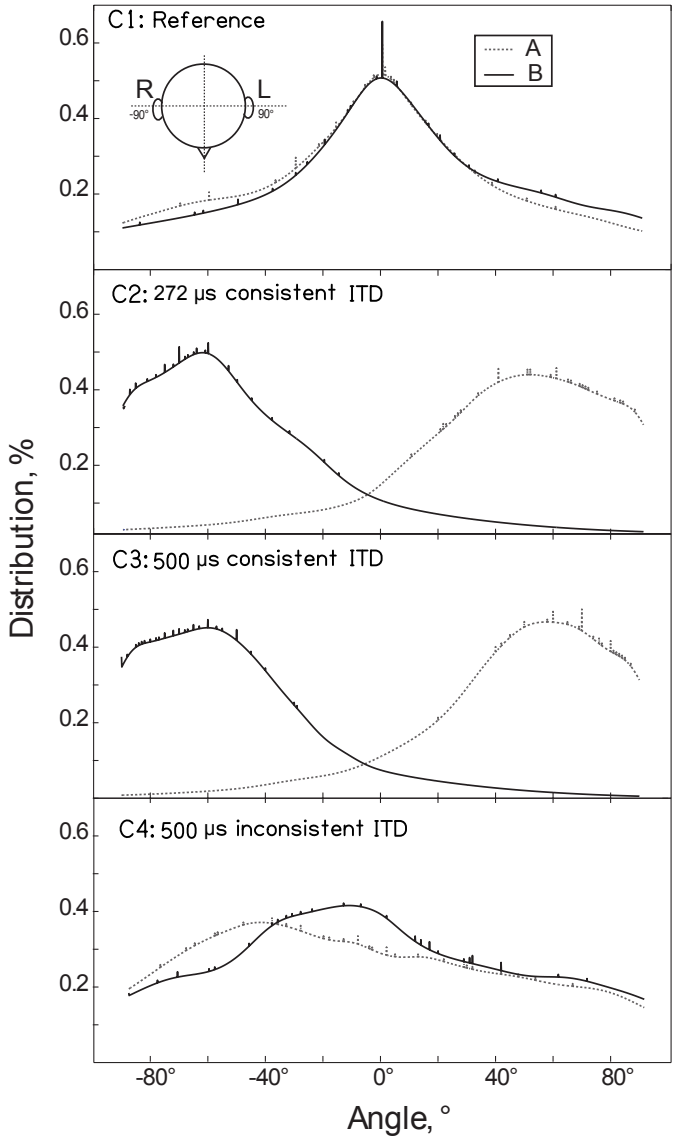

Figure I.7 – Mean results for the subjective lateralization task where listeners were asked to draw on a protractor (from $-90°$ to $+90°$) the direction(s) from which they perceived the sounds. Gaussian distributions were then fitted to the raw individuals data, the mean point(s) and the spread(s) of their responses corresponded to the mean(s) and the standard deviation(s) of the distributions. The curves represent the mean of these distributions for the thirteen listeners of Experiment 2, for the stimuli A (dotted lines) and B (solid lines) in each condition. From top to bottom, the panels correspond to Conditions 1 to 4.

for Condition 4). Moreover, there was no significant difference between the distributions of stimuli A nor between the distributions of stimuli B across Conditions 2 and 3.

Discussion

The mean threshold was significantly lower in Condition 1 compared to Conditions 2, 3 and 4. This results confirms the results of Experiment 1, indicating that ITD can favor obligatory stream segregation.

The mean thresholds were significantly higher in Condition 3 compared to Condition 4. In this last condition, the ITDs were swapped between high and low frequencies and that led to blurred perceived positions. Even though the difference in perceived position was not eliminated, the percept of lateralization was strongly reduced (see Figure I.7). Since the amount of interaural differences was constant in Conditions 3 and 4 in each frequency band, the results suggested that streams tended to be more segregated when the difference in perceived position was more salient.

The rhythmic discrimination task showed no significant difference in threshold in Condition 2 compared to Condition 3. Thus, the increase of ITD from 272 to 500 μs did not significantly improve the segregation. This result indicated that if the interaural difference *per se* has an influence on stream segregation, then this effect reached a ceiling. This hypothesis is not in accordance with previous results showing that an increase of ITD above the physiological range increased the effect on segregation (Füllgrabe and Moore, 2012; Stainsby *et al.*, 2011). Once again, this difference might be due to the nature of the stimuli used (see III.E.). Besides, the lateralization task did not show any difference in perceived position between the As and between the Bs in Conditions 2 and 3. This result supports the idea of a ceiling effect of ITD *per se*.

The lateralization task also showed that the estimated positions in Condition 3 were consistent with the ITD values (an ITD of $\pm$500 μs corresponds to an azimuth of $\pm$60°, Feddersen *et al.* (1957)) but that the perceived positions in Condition 2 were overestimated (an ITD of $\pm$272 μs should correspond to an azimuth of $\pm$30°). An explanation for this poor accuracy in the lateralization task could be that a constant ITD was applied across frequency, while real ITDs depend on frequency because of the diffraction effects around the head (Kuhn, 1977). Thus, the lateralization task used in the present study was presumably less accurate than a pointing task for example, where listeners have to move a narrow band of noise to match the perceived position of the target stimulus (Bernstein and Trahiotis, 1985). However, the present lateralization task allowed to check that the perceived position was actually reduced from a

condition with a consistent ITD across frequency to a condition with an inconsistent ITD across frequency.

2.2.6 General discussion

The percept of lateralization was strongly reduced between Conditions 3 and 4 in Experiment 2. This was associated with a significant decrease in threshold between these two conditions meaning a reduced tendency to segregate the streams. These results supported the hypothesis that the notion of perceived position is a cue used by the auditory system to organize the streams as long as the percept is salient enough. This interpretation is coherent with the results of Darwin and Hukin (1999). Indeed, they founded that when the listeners attend for a particular source, they track the particular location of the auditory object instead of tracking the frequency components that share the same ITD.

The weak effect of ITD found by Füllgrabe and Moore (2012) and Stainsby *et al.* (2011) might be explained by the fact that their stimuli did not induce enough perceived positions. Sandel *et al.* (1955) presented two localization experiments where the listeners had to move a pointer (a loudspeaker producing a broadband noise) to match the location of a target (a loudspeaker producing a pure tone). As a control condition, the pure tone target was replaced by a broadband noise filtered from 100 to 5000 Hz in Experiment 1. They found that noises can be localized with greater accuracy than pure tones.

Experiment 2 of the present study suggested that the auditory system can rely on a difference in perceived position associated with an ITD to segregate auditory streams (Condition 3 versus Condition 4). However, when ILDs were introduced in the stimuli of Experiment 1 (Condition 3 versus Condition 2), no segregation enhancement was observed compared to the diotic condition, even though ILD produces a percept of lateralization. This result suggests that the perceived position induced by the ILDs in Condition 3 of Experiment 1 was not salient enough to produce obligatory streaming. This result was in agreement with the findings of Wightman and Kistler (1992) who presented two experiments of lateralization using stimuli manipulated such that ITD gave a localization cue toward one direction and ILD gave cue toward an other direction. The listeners had to judge the perceived position of these conflicting sounds. In a first experiment, the lateralization task was made with broadband gaussian noises and in a second experiment with high-pass-filtered noises. The results showed that the perceived position was determined based principally on ITD as long as the stimuli presented low frequencies. Thus, for the present study, using broadband speech-shaped noises, ITD was the dominant cue for lateralization compared to ILD.

Even if the percept of lateralizatzion was reduced - but not eliminated - from Condition 3 to Condition 4 in Experiment 2, the difference in thresholds between Condition 1 and Condition 4 remained significant, meaning that the streams were significantly more segregated in Condition 4 compared to Condition 1. This segregation could probably not be explained only by a difference in perceived position, the ITDs available in this condition might have been used to segregate the streams. An interpretation of the results of Experiments 1 and 2 would be that ITDs are useful to obligatory segregate bursts of speech-shaped noises until a critical value above which the influence reaches an asymptote, and that differences in perceived positions can provide additional segregation when the percept of lateralization is salient enough.

Edmonds and Culling (2005) investigated the influence of ITD on spatial unmasking. They measured speech reception thresholds (SRTs) for a target speech presented with concurrent speech masker, and assessed whether the mechanisms underlying spatial unmasking rely on a difference in ITD within each frequency channel or a difference in ITD consistent across frequencies. ITD is a useful cue for sound localization as long as it is consistent within frequencies. The results showed that listeners can rely on ITD in each frequency band to reach high performance, spatial unmasking was not impaired by inconsistent ITDs across frequency as long as target and interferer differ in ITD. Thus, ITD differences can be exploited within each frequency band, even if this leads to unclear perceived position, to segregate target and masker. Their results seemed on opposition with the results of the present study suggesting that consistent ITD across frequency can favor segregation of auditory streams. However, the stimuli used by Edmonds and Culling (2005) contained strong streaming cues other than perceived position, such as differences in pitch, timbre and level. The differences in perceived positions might then be weaker in this situation than other cues, so that it was not significant in the study of Edmonds and Culling.

In a multi-talkers environment, if it can be assumed that the spatial configurations of speakers and listeners remain sufficiently constant over a given period of time, the consistency of the location cues associated with a difference in source position could be used by the auditory system. Thus, the consistency of the spatial differences could be relevant for segregation of competing voices. The present study does not allow to conclude on this particular point as frozen stationary noises were used. In fact, speech sounds are generally unfrozen. In this respect, an outlook of this study would be to assess the influence of binaural cues using unfrozen stimuli to get a step closer to real-sound situations.

2.2.7 Summary and Conclusions

The main purpose of this study was to investigate whether binaural cues (ILDs and ITDs) and the associated perceived position could enhance stream segregation of speech-shaped noises. The results of Experiments 1 and 2 showed a greater tendency to segregate the streams when the sources were spatially separated compared to when they were co-located. The results of Experiment 1 suggested that the perceptual organization of sounds was rather based on the monaural spectral level variations across time than on the interaural level differences associated with ILDs. Experiments 1 also showed a significant influence of ITD on obligatory stream segregation of broadband speech-shaped noises.

Experiment 2 investigated if the influence of ITD was due to the interaural difference *per se* and / or to the corresponding differences in perceived position. The results confirmed that ITDs can favored stream segregation but also suggested that the difference in perceived position associated with an ITD has a stronger influence on stream segregation. An interpretation of both experiments would be that the differences in perceived position can help to segregate sounds when the percept of lateralization is salient enough, and that ITDs are also useful *per se* until a critical value above which their influence reaches an asymptote.

Chapter II

Towards studying the segregation of speech sounds

So far in this PhD work, the results showed that spatial differences can be used by the auditory system to separate streams of broadband noises. The main idea of this second Chapter was to replicate these experiments with speech items instead of frozen noises. To reach this goal, two preliminary studies were conducted. First, because running speech presents a high degree of acoustical variability, the robustness of stream segregation based on a frequency difference to variability on the same perceptual scale (i.e., pitch) was assessed (using pure tones as a first step, section 1). Second, as the results of the first studies in Chapter I suggest that the effect of spatial cues are of a small magnitude, it is assumed that those cues could favour segregation of speech items but only in interaction with another factor, such as a difference in fundamental frequency. Thus, the F0 difference needed to separate speech items was evaluated (section 2).

These two studies were conducted in collaboration with Andrew Oxenham at the Auditory Perception and Cognition Lab at the University of Minnesota.

1 Influence on segregation of a random frequency variability within pure tone streams

1.1 Abstract - Résumé

Abstract

Most studies investigating auditory stream segregation have been focused on varying a single parameter between two constant stimuli (i.e., frozen) and the influence of this parameter on perceptual segregation was observed. However, considering more realistic stimuli such as speech sounds, multiple parameters vary at the same time (spectral envelope, F0...). A motivation

for this study was to get closer to everyday acoustic environments by introducing frequency variability within pure tone stimuli. This way, the two considered parameters (frequency variability within the streams and frequency difference between the streams) varied on a single perceptual scale: pitch. The listeners were asked to report their perception (one stream or two streams) after hearing the sequences of A-B-A pure tones. The frequency of each stimulus A_i and B_i in the sequences randomly varied, and the mean frequency difference between the A- and B-streams (ΔF) was set to different values.

The results showed that a large frequency variability within the streams induced a complicated percept, most presumably consisting of more than two streams, with pop-out sounds. The response to the task (one stream or two streams) seemed to depend on the listeners' strategy to fuse the pop-out sounds or to segregate them into isolated streams. It appeared that two groups of listeners could be identified. In the first group, the listeners seemed to respond "one stream" independently of the pop-out sounds, and in the second group, listeners seemed to respond "two streams" because of the pop-out sounds even if no ΔF was introduced. For the listeners of Group 1, the proportion of two streams increased when ΔF increased, from 2 semitones up to 5. The listeners of Group 2 tended to be less sensitive to a difference in ΔF to segregate the streams as the frequency variability within the streams increased.

Résumé

Jusqu'à présent, la plupart des études portant sur la ségrégation de flux auditifs se concentraient sur un unique paramètre qui variait entre deux stimuli constants (dans le sens gelés). Les expérimentateurs observaient alors dans quelle mesure ce paramètre influençait la ségrégation. Cependant, si l'on considère des stimuli plus réalistes, comme des signaux de parole, plusieurs paramètres varient en même temps (par exemple l'enveloppe spectrale, la F0...). Cette étude a donc été motivée par la volonté d'étudier des situations plus proches de la réalité, en introduisant dans un premier temps de la variabilité spectrale dans des séquences de sons purs. Cette variabilité a été introduite à la fois au sein de chaque flux, mais aussi entre les flux. De cette manière, les deux paramètres considérés (c'est à dire la variabilité spectrale intra-flux et la différence spectrale entre les flux) variaient sur une même dimension perceptive: la hauteur. Les auditeurs écoutaient des séquences de sons purs A-B-A, et devaient indiquer quel percept avait été prédimonant (un ou deux flux). La fréquence de chaque stimulus A_i et B_i dans les séquences variaient de manière aléatoire, et la différence moyenne de fréquence entre les flux (ΔF) était fixée à différentes valeurs.

Les résultats ont montré qu'une large variabilité spectrale intra-flux conduisaient à un percept complexe, vraisemblablement constitué de plus que deux flux, avec des sons qui semblent ressortir. La réponse à la tâche (choisir entre un ou deux flux) dépendait de la stratégie développée par les auditeurs d'intégrer les sons qui ressortent ou bien de les séparer en flux distincts. Deux groupes d'auditeurs ont été identifiés. Dans le premier groupe, les auditeurs ont eu tendance à répondre "un flux" indépendamment des sons qui ressortaient, et dans le second groupe, les auditeurs ont eu tendance à répondre "deux flux" à cause des sons qui ressortaient même si aucune différence de fréquence n'avait été introduite entre les flux (i.e., $\Delta F = 0$). Pour les auditeurs du groupe 1, la proportion de réponses "deux flux" augmentait lorsque ΔF augmentait à partir de 2 demi-tons jusqu'à 5. Les auditeurs du groupe 2 semblaient être moins sensibles à une différence de ΔF pour la ségrégation à mesure que la variabilité spectrale intra-flux augmentait.

1.2 Rationale

So far, there is limited work on stream segregation with sequences of varying stimuli except those using melody recognition tasks (see the Introduction, section 3.2.3). Dowling (1973) presented two melodies to the listeners, separated by a silence gap of 2 s. The first melody was "standard" while the second was interleaved with a masker. The two melodies were either identical or different and the task consisted of detecting if the second melody was identical to the first one or not. The mean frequency difference (ΔF) between the masker and the second melody was set to 0, 6 or 12 semitones. The results showed that the listeners obtained better results when ΔF increased. Hartmann and Johnson (1991) found that listeners reached 90% of correct answers when ΔF was of 12 semitones for the same task. In the same way, Vliegen and Oxenham (1999) showed that musicians obtained 85% of correct answers when ΔF was of 11 semitones. It is worth noting that the ΔF required to separate two melodies is larger than the ΔF required to separate two streams of constant pure tones: around 4 semitones (van Noorden, 1975). However, the studies using melody recognition tasks did not investigate how the spectral variability within the melodies influence the ΔF required to separate the melodies. Indeed, the variability was inherent to the melodies.

van Noorden (1975) evaluated the temporal coherence boundary (i.e., the critical value of the considered acoustical parameter above which listeners are no longer able to hear coherence, TCB, see Introduction) and varied the interval (I) between each stimuli. In some conditions, the interval was either constant (i.e., I was of 1 or 2 semitones), alternate (i.e., I was of 1 or 2 semitones but its sign was alternately positive and negative) or random (i.e., the next tone in

the sequence was determined by randomly choosing a positive or negative I and adding it to the frequency of the previous tone). The TCB was evaluated by asking the listeners to modify the tempo of the sequences (technically, they had to move a potentiometer that controlled the tone repetition time) until they could just hear the tones as an integrated sequence. The tone duration was kept constant at 40 ms. The results showed that the three types of sequence have the same TCB (van Noorden, 1975, Fig. 5.5 pg. 46), suggesting that the TCB would not be influenced by irregularities.

Some neurophysiological studies have shown that regular temporal patterns can stabilize the perceptual organization of auditory streams. Measuring event-related brain potentials (ERPs), it has been shown that the auditory system can detect any violation of a regularity (for a review, see Näätänen *et al.* (2001)). Bendixen *et al.* (2010) assessed the influence of temporal regularities on streaming. The listeners were presented with sequences of [ABA-ABA] triplets and were asked to continuously indicate how they perceived the sets of tones (integrated or segregated). The frequency and intensity of A and B were manipulated independently, leading to different regular or irregular patterns. The patterns of regularities were the same for all the listeners. The regular intensity pattern consisted of introducing an emphasis once every four As, or once every three Bs, and the regular frequency pattern consisted of a set of two repeating pair of A tones (i.e., $A_1A_1A_2A_2$) or a set of three ascending B tones (i.e., $B_1B_2B_3$). The proportion of time the listeners pressed the "integrated" (or "segregated") key was evaluated in ten conditions: a control condition without regularity and nine test conditions introducing the regularities in one or the two features (intensity or frequency or both), in one or the two stimuli sets (A, B or both). The results indicated that the percept was significantly influenced by the "amount of regularity" (number of regular features, number of streams containing a regular pattern), even though the type of feature (intensity or frequency) had no significant influence. These results indicated that the auditory system could use the temporal regularities to stabilize an established percept of segregation, but would not rely on these regularities to organize the streams.

In real acoustic environments like cocktail-party situations (Cherry, 1953), even though speech sounds can present acoustic regularities (harmonic vowels, spatial positions...), any sentence consists of a concatenation of different speech sounds. Moreover, running speech presents large acoustical variability that enables to convey information. In order to get closer to real stimuli, the aim of the present study was to evaluated the robustness of the pure tones segregation based on a frequency difference by introducing variability in the same perceptual scale (i.e., pitch).

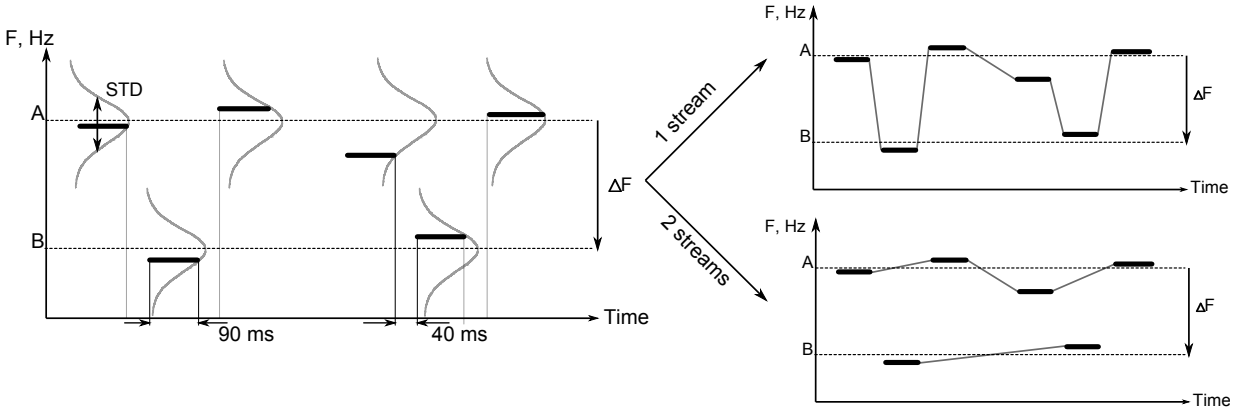

Figure II.1 – Rationale of the experiment: introduction of a random frequency variability within a sequence of pure tone triplets. The horizontal dotted lines correspond to the mean frequencies of the A and B tones, and ΔF is the mean frequency difference between A and B. F_A and F_B are defined by a Gaussian distribution centred on the mean frequencies. The standard deviations (STDs) of the distributions and the mean ΔF are set to 0, 1, 2 or 5 semitones and 0, 1, 3, 6, 9, or 15 semitones, respectively, leading to 24 different stimulus conditions. The right panel indicates the different rhythms perceived depending on the percept. When one stream is heard, the listeners hear a galloping rhythm whereas when two streams are heard, they hear two regular rhythms.

1.3 Method and stimuli

In the present experiment, subjective judgments of perceived segregation were collected. In each trial, a sequence of 14 pure tone triplets [ABA-ABA] was presented to the listeners. As shown in Figure II.1, the rhythm of the sequence depended on the listeners' perception. If the sequence was perceived as an integrated stream, the listeners would hear a single galloping rhythm (top right panel of Figure II.1), otherwise, if the sequence was perceived as two segregated streams, the listeners would hear two regular streams that differ in tempo (van Noorden, 1975). The task for the listeners was to indicate whether they heard one or two streams (i.e., a galloping or a regular rhythm) and then the proportion of two-stream responses was calculated.

Each tone lasted 90 ms and was windowed on and off with 18-ms squared ramps. The silence duration between two stimuli within a triplet was 40 ms so that each triplet lasted 350 ms. The triplets were regularly spaced by 170 ms in the sequence, so the A-tones were spaced by 170 ms and the B-tones by 430 ms. The left panel of Figure II.1 synthesizes the structure of the sequences.

To introduce a random spectral variation within the streams, the frequency of each stimulus in the A- and B-stream was a sample of normal distributions centered around fixed values, F_A and F_B. ΔF was the mean difference between the streams, so $\Delta F = F_A - F_B$. Two stimulus parameters varied, ΔF and the standard deviation (STD) of the normal distributions. ΔF was

equal to 0, 1, 3, 6, 9 or 15 semitones and STD was equal to 0, 1, 2 or 5 semitones. F_A was kept constant equal to 1000 Hz and F_B was set ΔF semitones below F_A (i.e., approximately 1000, 944, 841, 707, 595 and 420 Hz). The same STD was applied to the A- and B-stream in each trial. Once selected, ΔF and STD were kept constant within a sequence.

The total combination of all ΔF and STD resulted in 24 conditions (6 ΔF * 4 STD). A run consisted of 2 repetitions of each ΔF, in a random order, for a given STD. 30 runs were completed for each STD, also in random order. Thus, the 24 conditions were repeated 60 times (2 * 30). Each sequence lasted around 5 s, so the whole experiment lasted about 4 hours split into two 2-hour sessions. The stimuli were presented diotically through Sennheiser HD650 headphones at 65 dB SPL in a double-walled sound-attenuated booth. Listeners could use either the computer mouse or the keyboard to enter their answers on a graphical interface visible inside the booth.

1.4 Listeners

Fifteen listeners participated in the present experiment (8 females, mean age = 22 yrs., ranged between 18 and 43 yrs., standard deviation = 7 yrs.). They all had pure tone thresholds of less than 20 dB HL at octave frequencies between 200 and 8000 Hz. They were paid an hourly wage for their participation.

1.5 Results

The proportion of two-stream responses was measured based on the listeners' subjective responses. When no average frequency difference was introduced between the streams, one might expect the percept to be integrated, and thus the proportion of two-stream responses to be close to 0. However, eight listeners showed discrepant results when ΔF was null depending on the value of STD. For each listener, the number of one-stream and two-streams responses were calculated in the condition $\Delta F = 0$ for each STD value. Two groups of listeners could be identified. The listeners for who the number of two-streams responses remained inferior to 12 (i.e. 20% of the total number of responses, arbitrarily chosen) were grouped in Group 1, the others were grouped in Group 2. Figure II.2 shows the mean repartition of the responses across the listeners when $\Delta F = 0$ for each STD condition and for each group. Groups 1 and 2 counted 7 and 8 listeners, respectively. The two groups were significantly different as a result of a two-way measure ANOVA where the two parameters were the group (2 levels) and the standard deviation STD (4 levels).

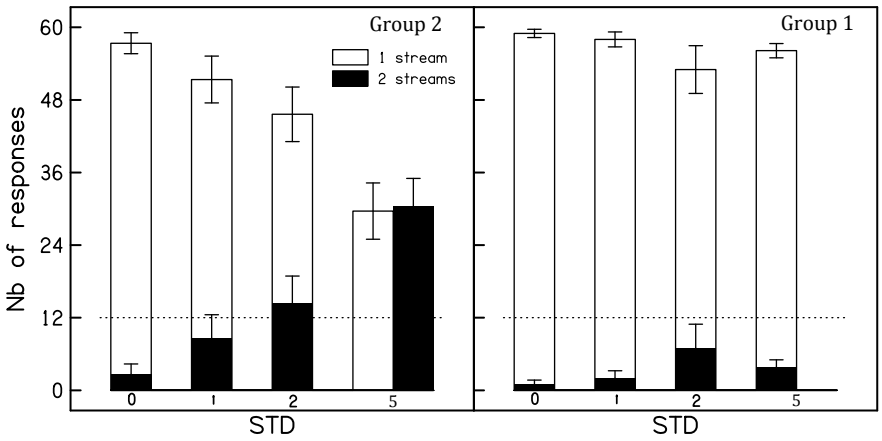

Figure II.2 – Mean repartition of the responses across listeners when $\Delta F = 0$ for each STD condition and for each group. The white bars correspond to the one-stream responses while the black bars correspond to the two-streams responses. The error bars represent the standard errors.

The mean results of each listener across repetitions were fitted using a maximum likelihood estimation (MLE) method. This method enabled us to estimate two parameters of the psychometric curves: F_{50} and the slope. F_{50} corresponds to the frequency difference leading to the 50% correct point on the psychometric curve. F_{50} gives an estimation of the bi-stable point where the percept "flips" between one stream and two streams. The slope is associated with the sensitivity to the variable parameter, here ΔF (see section I.3.2.1 for more details concerning the psychometric functions and the corresponding parameters).

The mean results across the listeners of the subjective streaming task are shown in Figure II.3 (symbols). The proportion of two-stream responses is plotted as a function of ΔF for each STD condition. The lines correspond to the psychometric curves calculated with the parameters estimated with the MLE method[1]. The results were assessed with a two-way ANOVA where ΔF and STD were the two parameters with 6 and 4 levels, respectively. The main effects of ΔF, STD and their interaction were significant ($p < 0.01$ in each case) in each groups of listeners.

Influence of STD on the 50% correct point

The right panel of Figure II.4 represents the mean values (across listeners for each group) of F_{50} and the corresponding slope as a function of STD. A one-way repeated measures ANOVA

[1] An estimation of the parameters (i.e., the slope and the F_{50}) was done for each listeners and then these estimations were averaged across all the listeners. The psychometric curves were calculated with the mean values of the estimated parameters.

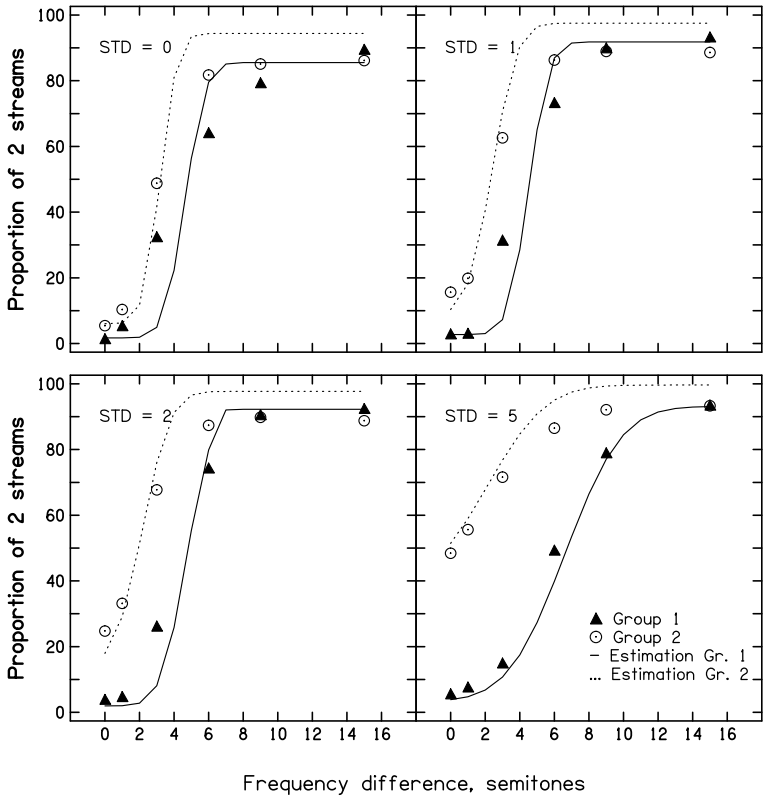

Figure II.3 – Mean results across the listeners of the subjective streaming task for the two groups. The proportion of two-streams responses is plotted as a function of the frequency difference between the streams for each STD condition. The lines correspond to the psychometric curves calculated with the parameters estimated with the MLE method.

was processed on the data estimated with the MLE method[2] for each group of listeners. For the listeners of Group 1, the frequency difference required to reach bi-stability increased significantly as STD increased ($F(1,6) = 24.1; p < 0.001$). The linear contrast[3] was also significant ($F(1,6) = 24.3; p < 0.003$). A post-hoc analysis (LSD) indicated that the F_{50} (for the listeners of group 1) obtained with $STD = 5$ semitones was significantly higher than those obtained with $STD = 0, 1$ and 2 semitones (p=0.004, 0.0001 and 0.0001 when the STDs were 0 vs 5, 1 vs 5 and 2 vs 5 semitones, respectively). The F_{50} obtained with $STD = 0, 1$ and 2 semitones were not significantly different. For the listeners of Group 2, STD had no significant influence on the frequency difference required to reach the 50% correct point.

Influence of STD on the slopes of the psychometric curves

The left panel of Figure II.4 shows the mean slopes (across the listeners for each group) of the psychometric curves as a function of STD. A one-way repeated measures ANOVA was processed on the data estimated with the MLE method[4] for each group of listeners. The slopes significantly decreased when STD increased for the listeners of Group 2 ($F(1,7) = 5.4; p = 0.007$). A post-hoc analysis (LSD) indicated that the mean slope (for the listeners of group 2) obtained with $STD = 5$ semitones was significantly more flat than those obtained with $STD = 0, 1$ and 2 semitones (p=0.028, 0.020 and 0.011 when the STDs were 0 vs 5, 1 vs 5 and 2 vs 5 semitones, respectively). The mean slopes obtained with $STD = 0, 1$ and 2 semitones were not significantly different. For the listeners of Group 1, STD had no significant influence on the slope of the psychometric curves ($p = 0.1$).

1.6 Discussion and prospectives

The results suggested that the segregated percept increased for $\Delta F = 0$ when STD increased for eight listeners among fifteen (Group 2). A possible hypothesis to explain these results would be that when the STD increased, the perceptual organization became more complicated, some sounds might be grouped together and some others might pop out. The percept should be broken into many isolated and incoherent sound events rather than two separate coherent streams. It seemed that for the other listeners (i.e., seven over fifteen, Group 1), an increase of STD had no influence on stream segregation as long as $\Delta F = 0$. Thus, two types of segregation might be distinguished, the segregation for which each of the two streams was fully coherent

[2]The analysis of the raw data showed the same results.
[3]A linear contrast analysis is similar as a linear regression as far as the distance between each level is equal.
[4]The analysis of the raw data showed the same results.

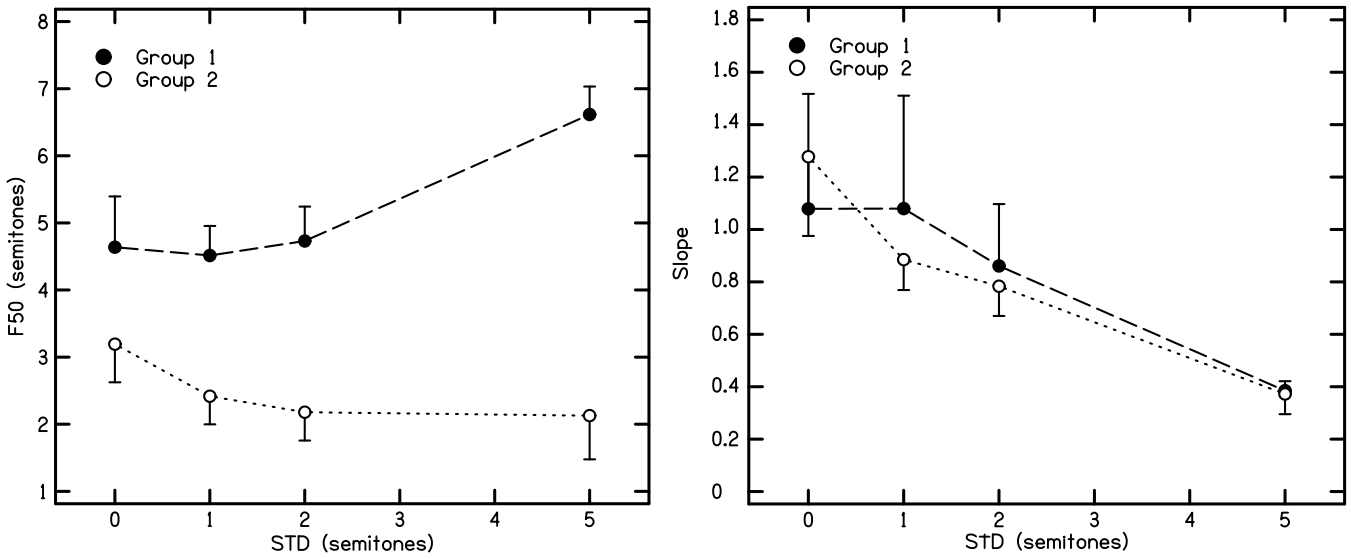

Figure II.4 – Evolution of the mean F_{50} (left panel) and corresponding slope (right panel) of the psychometric curves as a function of the STD for each group of listeners. F_{50} corresponds to the frequency difference which leads to the 50% correct point on the psychometric curve (i.e., the maximum of bi-stable percept). The slope is associated to the sensitivity to the variable parameter, here ΔF. In both graphs, the error bars correspond to the standard errors.

(without any sound outside these two streams), and the segregation for which each single sound could generate an independent stream. These results raised the question of the bias introduced by a subjective procedure. In fact, even if this type of methodology enables one to access directly the listener's percept, the experimenter has only a weak control of how the task is understood and has little access to the response strategy. For instance, if the listeners of Group 1 heard pop-out sounds, they perhaps managed to ignore them judging the sequence as a whole. Conversely, as the procedure was a two-alternative forced-choice measure (one stream or two streams), the listeners of Group 2 might have responded "two streams" when they wanted to respond "several streams".

Besides, the results showed that the slope of the psychometric curves significantly decreased as STD increased for the listeners of Group 2. The slope of a psychometric curve is related to the sensitivity to the variable parameter, here the mean frequency difference between the streams (i.e., ΔF). A steeper slope corresponds to a greater sensitivity since a small difference of the parameter leads to a large difference in percept. Conversely, a flatter slope corresponds to a weaker sensitivity. Thus, the listeners of Group 2 were less sensitive to a difference in ΔF to segregate the streams as STD increased. These results suggested that when a large frequency variability was introduced within the streams, the perceptual organization of the sounds in one or two streams based on frequency did not appear clearly and segregation seemed to be dismantled.

For the listeners in Group 1, when the frequency variability introduced within the streams was less than or equal to 2 semitones, the influence of STD was not significant, while the influence of ΔF was. Indeed, the percept tended to be more segregated by increasing ΔF independently of STD. In these three cases (i.e., STD = 0, 1 or 2 semitones), a frequency difference of 4 semitones or more between the streams led to a segregated percept in more than 50% of the trials. This result is consistent with the previous studies that showed the influence of a frequency difference on stream segregation of pure tones (for example, see Hartmann and Johnson (1991); van Noorden (1975)). However, the results also showed that when the STD was equal to 5 semitones, a greater frequency difference was required for the streams to be segregated (i.e., more than 6 semitones). Thus, introducing a large frequency variability within the streams tend to make the segregation weaker, or tend to stabilize the integrated percept. This result complements the finding of Bendixen *et al.* (2010) who showed a stabilization of the segregated percept by increasing the pattern regularities. Moreover, this result might explain why the ΔF needed to separate melodies in melody recognition task (Dowling, 1973) was higher than the one needed to separate streams of constant pure tones van Noorden (1975). However, this effect is not significant for the listeners of Group 2.

It would be worth doing this experiment again using an objective task such as a rhythmic discrimination task (Roberts *et al.*, 2002). In this case, if some sounds pop-out because of the large spectral variability introduced within the streams, the ability to detect the irregularity would be strongly reduced. Thus, one might expect that the irregular rhythm, which requires integration to be detected, would be more and more difficult to find as STD would increase, even with a small frequency separation between the streams. The results should tend towards those of Group 2 because the potential pop-out sounds could break the rhythm.

1.7 Conclusions

- For both groups, the proportion of two-streams responses increased when the frequency difference (ΔF) between the streams increased. When no spectral variability was introduced, a ΔF comprised between 3 and 5 semitones led to 50% of two-streams responses.

- Two groups of listeners were distinguished. For some listeners (7 out of 15, Group 1), a greater frequency separation was required between the streams to hear segregation when a large spectral variability was introduced within the streams. The other listeners (8 out of 15, Group 2) seemed to become less sensitive to a frequency difference between the streams as the spectral variability increased within the streams.

- The results of the present study showed that when a large spectral variability was introduced within streams of pure tones, the percept could be broken into two separate and coherent streams (Group 1). For some listeners (Group 2), it is possible that the percept broke into many isolated and incoherent sound events (more than two streams).

- These results were obtained with a subjective streaming task. This type of procedure is highly dependent on how the listeners understood the task and on the strategy developed to do it. Thus, the results could be complemented with an additional objective experiment such as a rhythmic discrimination task. An objective task would prevent the bias introduced by the pop out sounds in the listeners' response to a subjective task.

2 Stream segregation with speech sounds

2.1 Abstract - Résumé

Abstract

The natural ability to perceptually organize the coming sounds into coherent streams is relevant for speech intelligibility. Previous work on auditory streaming has focused on pure or complex tone stimuli. Little work has been done using broadband noises and even less using speech material. This study was the second step towards a wider aim consisting of investigating to what extent spatial differences could influence the segregation of speech sounds. Thus, the main purpose of the present study was to evaluate the required F0 difference between the streams to reach an ambiguous percept (i.e., 50% of segregation). Further experiments will investigate if spatial cues could bias this maximum of bi-stable percept towards segregation.

The speech sounds used consisted of a fricative consonant plus a vowel recorded from two native American English speakers (a female and a male). A set of 45 sounds were recorded and their pitch contour was flattened, thus the voices sounded monotone. The listeners were presented with sequences of alternate speech sounds [A_i-B_j]. The F0 of the As was kept constant in all sequences equal to 220 Hz (female voice) or 110 Hz (male voice); while the F0 of the Bs was set 0, 1, 3, 5, 7 or 9 semitones above F_A. In half of the presentations, the sequences contained only different stimuli selected in random order. In the second half, a sound was repeated twice in a row. The repeat was either introduced across streams (RAS) or within streams (RWS). After each sequence presentation, the task was to detect whether or not a repeat had been introduced. The RAS measure is linked to a measure of obligatory streaming as the listeners reached higher performances when they managed to integrate the streams. Conversely, the RWS measure is linked to a measure of voluntary streaming as the performances were higher when the listeners could segregate the streams.

The results were expressed in terms of sensitivity (d-prime) as a function of the $\Delta F0$ between the streams. The sensitivity increased as $\Delta F0$ increased for the RWS measure, suggesting that voluntary streaming becomes easier when the pitch difference between the streams increased. Besides, the sensitivity tended to decrease with an increase of $\Delta F0$ in the RAS measure. This result indicates that the integration becomes harder as $\Delta F0$ increases. For both measures, a $\Delta F0$ comprised between 3 and 5 semitones between the streams of random speech sounds led to an ambiguous percept.

Résumé

La capacité naturelle à pouvoir organiser les sons perçus en différents flux auditifs est pertinente pour l'intelligibilité de la parole. Les précédentes études traitant de la ségrégation auditive ont principalement utilisé des sons purs ou des sons complexes. Seulement très peu d'études

ont utilisé des bruits large-bande et encore moins des signaux de parole. Cette étude représente la seconde étape d'un objectif plus large consistant à évaluer l'influence des différences spatiales sur la ségrégation de parole. Ainsi, l'objet principal de cette étude était d'évaluer la différence de F0 entre les flux pour atteindre un percept ambigu (c'est à dire 50% de ségrégation). D'autres expériences seront menées pour étudier si les indices spatiaux peuvent faire tendre ce maximum de bistabilité vers un percept ségrégé.

Les signaux de parole utilisés étaient constitués d'une consonne et d'une voyelle enregistrées par deux locuteurs Américains (une femme et un homme). Un ensemble de 45 stimuli ont été enregistrés et le contour de fréquence fondamentale (F0) a été applati, ainsi les voix ont été monotonisées. Les auditeurs écoutaient des séquences de sons alternés [A_i-B_j]. La F0 des sons A a été gardée constante à 220 Hz (pour la voix de femme) ou à 110 Hz (pour la voix d'homme) et la F0 des sons B a été fixée à 0, 1, 3, 5, 7 ou 9 demi-tons au-dessus de F_A. Dans la moitié des présentations, les séquences étaient composées uniquement de stimuli différents choisis de manière aléatoire. Dans la seconde moitié des présentations, un même stimuli était répété deux fois de suite. La répétition pouvait soit être introduite inter-flux (RAS) ou intra-flux (RWS). À la fin de chaque séquence présentée, la tâche consistait à détecter si une répétition avait été introduite ou non. La mesure RAS est reliée à une mesure de ségrégation obligatoire puisque les auditeurs obtenaient de meilleures performances lorsqu'ils parvenaient à intégrer les flux. Par ailleurs, la mesure RWS est reliée à une mesure de ségrégation volontaire puisque les performances étaient meilleures lorsque les auditeurs pouvaient séparer les flux.

Les résultats ont été exprimés en terme de sensibilité (d-prime) en fonction de la différence de F0 entre les flux ($\Delta F0$). La sensibilité augmentait avec $\Delta F0$ pour la mesure RWS, indiquant que la ségrégation volontaire devient plus facile quand la différence de hauteur tonale augmente entre les flux. D'autre part, la sensibilité avait tendance à diminuer lorsque $\Delta F0$ augmentait dans la mesure RAS. Ce résultat indique que l'intégration devient plus difficile quand $\Delta F0$ augmente. Pour les deux mesures, une différence de F0 comprise entre 3 et 5 demi-tons entre les flux de parole aléatoires a conduit à un percept bi-stable.

2.2 Rationale

Solving an auditory scene can be useful for speech intelligibility. In a context of cocktail-party (Cherry, 1953), where a listener tries to understand a message delivered by a target speaker in a noisy background, the sounds coming from the target and the competing sources have to be separated for the message to be intelligible. These situations are very common in everyday sound environments, however, little work has been done in auditory stream segregation using

speech material.

Gaudrain *et al.* (2007) investigated the influence of a F0 difference on vowels separation. Blocks of 6 random vowels were repeated during a sequence presented to the listeners. The F0 of the successive items alternated leading to a $\Delta F0$. The task consisted of reporting the order of appearance of the vowels. The results showed that obligatory stream segregation of vowels was induced by a fundamental frequency separation between the streams. This effect increased as $\Delta F0$ increased from 0 to 9 semitones.

In natural speech, some words - or parts of words - consist of vowels (relatively low frequency bands) and fricative consonants (relatively high frequency bands). David *et al.* (2014) found that small spectral differences associated with a difference in location could introduce obligatory stream segregation of broadband noises which can be assimilate to fricative consonants. This study consisted of a preliminary work to further investigate how spatial differences could influence the segregation of speech sounds. So, the main purpose was to investigate the influence of a F0 difference on the segregation of speech sounds consisting of concatenation of a fricative consonant and a vowel.

2.3 General methods

The detection of repetition method was used in this experiment (see the State of art section, paragraph 3.2.2). The listeners were presented with multiple sequences of 28 alternate speech sounds [A-B-A-B...]. Each speech sound consisted of a unique combination of a consonant and a vowel (see section Stimuli) and lasted 160 ms. The silence duration between each stimuli was set to 40 ms. In half of the trials, the sequences contained stimuli which were all different, and in the second half, a repeat was introduced (i.e. one repetition of a single sound just after its first presentation) at a random position in the sequence after the 12^{th} sound. The repeat was never introduced before the 12^{th} sound in order to let time for segregation, if any, to build up (Anstis and Saida, 1985; Roberts *et al.*, 2008). This method enabled us to measure both voluntary and obligatory stream segregation. In fact, in the sequences with repeat, when it was introduced across the A and B streams, the task was favoured by integration and thus the measure evaluated obligatory stream segregation. When the repeat was introduced within the B stream, the task was favoured by segregation and thus the measure evaluated voluntary stream segregation. The two measures are named in the following manuscript RAS (Repeat Across Streams) and RWS (Repeat Within Streams). The left panel of Figure II.5 shows the structure of the sequences for each measure (top: RAS, bottom: RWS). The circled grey symbols correspond to a repeated speech sound, the other symbols correspond to different sounds.

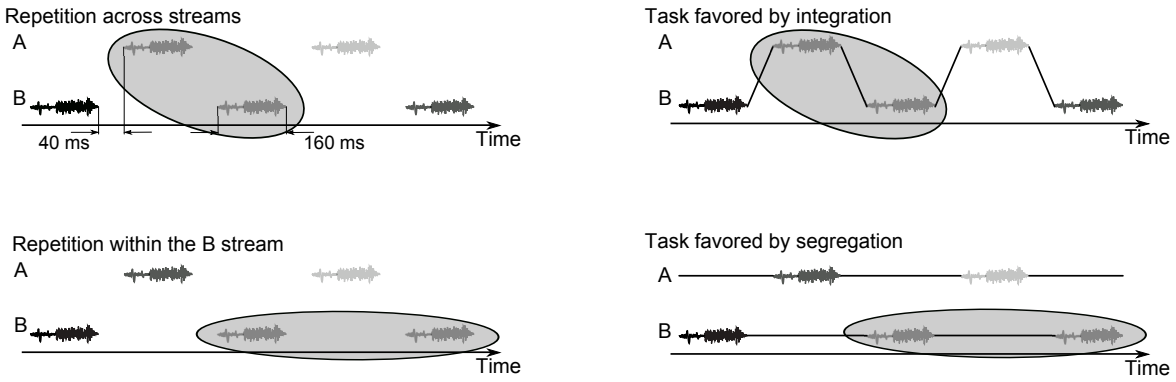

Figure II.5 – Left panel: structure of the sequences in the RAS (top-left) and in the RWS measures (bottom-left). In half of the presentations, the sequences consisted of only different stimuli and in the other half, a repeat was introduced. The circled grey symbols correspond to a repeated speech sound (i.e., the same consonant and the same vowel), all the other sounds are different. The right panel shows how the repetition could be detected. In the RAS measure (top-right), the performances were higher when the streams were integrated whereas in the RWS measure (bottom-right), the performances were higher when the streams were segregated.

The task in both measures was to indicate if the sequence contained a repeat or not. The sensitivity d-prime was estimated comparing the hit rate (i.e., proportion of a repeat detection when the repeat was actually introduced) and the false alarm (proportion of repeat detection when no repeat was introduced). For the RAS measure, the repeat was detectable (if any) only if the streams were integrated as integration enabled to follow the alternation of stimuli across the streams (top right panel of Figure II.5). For the RWS measure, the repeat was detectable (if any) only if the streams were segregated as segregation enabled to follow the stimuli within a single stream (bottom panel of Figure II.5). It is worth noting that for the RWS measure, the listeners were instructed to attend to the B stream (i.e., they had to focus on the low pitch stream) since if segregation occurs, he/she could also attend to the A stream where no repeat was introduced.

2.4 Stimuli and conditions

The stimuli used in this experiment were speech sounds consisting of a unique combination of a fricative consonant and a vowel pronounced by two native American English talkers (1 male and 1 female) and recorded with a portable recorder (Model PMD670, Marantz). The speech material consisted of forty-five combinations of five fricative consonants and nine vowels. Table II.1 shows the speech material and the corresponding sounds. The recordings were made in a sound attenuated booth using a sample rate of 48 kHz. The talkers were asked to pronounce

Tableau II.1 – Speech stimuli consisting of 45 combinations of five fricative consonants and 9 vowels of American English. The phonetic notations are shown between forward slashes. All of the stimuli are sounds used in real words of American English which are shown in bolt font in the table.

/fi/ → **fi**sh	/fI/ → **fi**t	/feI/ → **fai**l	/fae/ → **fa**d	/fνω/ → **foo**l	/fν/ → **foo**t	/fa/ → **fo**dder	/fΛ/ → **fu**n	/fϵ/ → **fe**d
/θi/ → **thi**s	/θI/ → **thi**ck	/θeI/ → **the**y	/θae/ → **tha**n	/θ νω/ → **thoo**l	/θ ν/ → **thou**gh	/θa/ → **tho**ng	/θ Λ/ → **thu**s	/θ ϵ/ → **the**n
/si/ → **si**ck	/sI/ → **si**t	/seI/ → **sai**l	/sae/ → **sa**d	/sνω/ → **soo**n	/sν/ → **soo**t	/sa/ → **so**ft	/sΛ/ → **su**dden	/sϵ/ → **se**nd
/∑i/ → **shi**p	/∑I/ → **shi**ft	/∑eI/ → **sha**pe	/∑ae/ → **sha**dow	/∑ νω/ → **shoo**t	/∑ ν/ → **shoo**k	/∑a/ → **sho**ck	/∑ Λ/ → **shu**t	/∑ ϵ/ → **she**ll
/hi/ → **hee**d	/hI/ → **hi**t	/heI/ → **hai**l	/hae/ → **ha**d	/hνω/ → **hoo**ligan	/hν/ → **hoo**k	/ha/ → **ho**d	/hΛ/ → **hu**t	/hϵ/ → **hea**d

nine speech sounds (i.e., a single consonant and all the vowels) spaced by a second or two in a single track. Five tracks corresponding to the five lines of Table II.1 were recorded. Each track was repeated twice and the recording with the best quality for each speech sound was kept as a stimulus for the experiment.

The forty-five stimuli were isolated, resampled at 44.1 kHz and time-windowed on and off with sine-squared ramps to last 160 ms. The mean pitches for the male and female voice were 110 and 220 Hz, respectively. The pitch variations from one speech sound to another were cancelled by flattening the pitch contours. Then, a pitch value was set for each stimulus according to the tested conditions. Although this signal processing introduce slight distortions in the formant's frequency, this approach enable us to control the pitch differences between the streams while getting closer to natural speech sounds.

In the two measures (RAS and RWS), the varied parameters were the gender of the speaker (male or female) and the F_0 difference between the As and Bs. ΔF_0 was equal to 0, 1, 3, 5, 7 or 9 semitones. For the male and female voices, the F_0 of the A-sounds was kept constant equal to 110 and 220 Hz, respectively and the F_0 of the B-sounds was set ΔF_0 semitones above (i.e., approximately 110, 117, 131, 147, 165 and 185 Hz for the male voice and 220, 233, 262, 264, 330 and 370 Hz for the female voice).

A run consisted of 2 repetitions of each ΔF_0 (6 items) and each sequence condition (2 items, with or without repeat). 26 runs had to be completed for each gender condition, so a total of 52 runs, and 624 trials. As each sequence lasted 5.6 s, the whole measure lasted around 1.5 hours.

2.5 Listeners

Fourteen listeners participated in both measures (6 females, mean age = 20 yrs. ranged between 18 and 36 yrs., standard deviation = 4 yrs.). American English was the first language of all the listeners, and they all had normal hearing (i.e., pure tone threshold of less than 20 dB HL at octave frequencies between 200 and 8000 Hz). They were paid an hourly wage for their participation.

Half of the listeners started with the RAS measure and the other half started with the RWS measure. They all came twice to the lab to complete the two measures for approximately 2 hours. The stimuli were presented diotically through Sennheiser HD 650 headphones at 65 dB SPL in a double-walled sound-attenuated booth. Listeners could use either the computer mouse or the keyboard to enter their answers on a graphical interface visible inside the booth.

2.6 Results

The detection rates are expressed in terms of d-prime scores. Using the signal detection theory (Macmillan and Creelman, 2004), a hit corresponded to a detection of a repeat when a repeat was actually introduced in the sequence, and a false alarm (FA) corresponded to a detection of a repeat when no repeat was introduced in the sequence. The rate of hits was evaluated as the number of hits divided by the number of sequences containing a repeat. Respectively, the rate of FA was evaluated as the number of FA divided by the number of sequences without repeat. Then, the d-prime score was the difference between the z-transforms of these two rates: $d' = z(rate(hit)) - z(rate(FA))$. Figure II.6 shows the mean results of d-prime across the fourteen listeners for the RAS (left panel) and RWS (right panel) measures. For the RAS measure, the gender of the speaker had no significant difference, so the results obtained with the female and male voices were averaged. However, the interaction gender-ΔF was significant for the RWS measure (see next paragraph), so the results were not averaged. The error bars represent the inter-listeners standard error of the mean results.

The mean d-prime scores across listener were assessed using a two-way repeated measure ANOVA. The two considered parameters were $\Delta F0$ (6 levels) and the gender of the talker (2 levels). For the RAS measure, there was no significant effect of these parameters. For the RWS measure, $\Delta F0$ significantly influence the ability to detect the repeat ($F(5, 65) = 5.557$ and $p < 0.001$). Indeed, the repeat was more accurately detected when ΔF increased. The effect of the speaker was not significant but the interaction of the effects of ΔF and speaker was significant ($F(5, 65) = 2.494$ and $p = 0.04$).

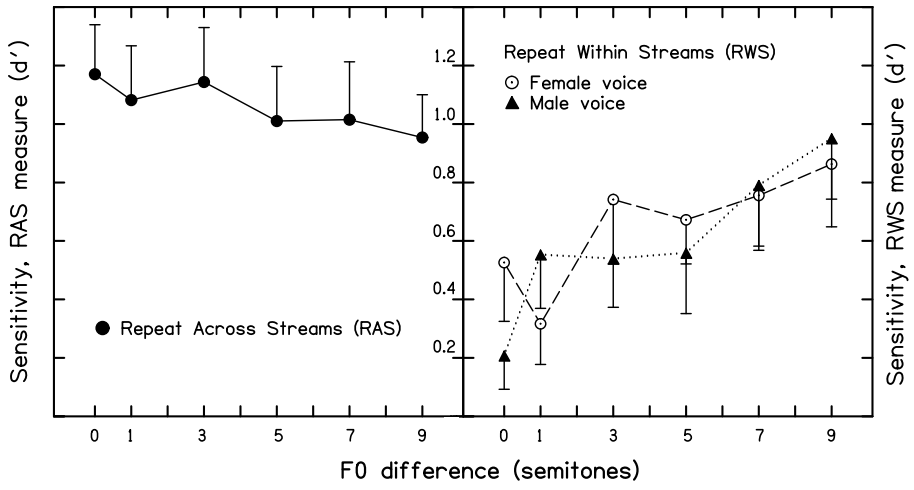

Figure II.6 – Mean detection rates across the 14 listeners expressed in terms of d-prime scores for the RAS (right panel) and RWS (left panel) measures. In the RAS measure, the results obtained with the female and male voices were averaged for each listener because the influence of the gender was not significant on the ability to detect the repeat. This was not the case for the RWS measure. The error bars represent the inter-listeners standard errors of the mean results.

2.7 Discussion and prospectives

The RWS measure is a measure of voluntary stream segregation since the listeners tried to segregate the streams in order to detect the potential repeat. High d-prime scores in the RWS measure is interpreted as a greater ability to segregate the streams. The results indicated that the task became easier (i.e., higher d-prime scores) as the fundamental frequency difference between the streams increased. This result is in agreement with the previous studies which found that a difference in F0 is relevant for the auditory system to segregate the streams.

High d-prime scores in the RAS measure are interpreted as a greater tendency to integrate the streams. Indeed, since the repeat was introduced across the streams, the repeat was detected if the listeners managed to fuse the streams and to follow the alternation of stimuli across time. Thus, this measure evaluate the $\Delta F0$ above which the listeners were no longer able to integrate the streams (obligatory stream segregation). The results indicated that $\Delta F0$ had no significant influence on the ability to detect the repeat. However, it is worth noting that even if the effect was not significant, there was a tendency for the sensitivity (d-prime) to decrease as $\Delta F0$ increased ($p = 0.10$). This tendency suggested that as $\Delta F0$ increased, the integration of the streams became more difficult. The fact that the effect was only marginally significant

might be explained either by the large variability among listeners or by a too small fundamental frequency difference between the streams, or by a tempo too slow.

Gaudrain *et al.* (2007) investigated whether auditory stream segregation of vowels could be influenced by a fundamental frequency difference. The pitch of each individual vowel was held constant and the F0 of the successive items alternated. $\Delta F0$ was constant for a given sequence and varied from trial to trial. The values of $\Delta F0$ were 0, 10, 21, 34, 47, 62, 78, 96, 116 and 138 Hz (approximately 0, 1.5, 3.5, 5.0, 6.5, 8.5, 10, 11.5, 13.5 and 15 semitones above 100 Hz). The listeners were asked to report the order of appearance of the vowels. An across score, which corresponded to the proportion of correct identifications across streams, and a within score, which corresponded to the proportion of correct identifications within streams, were measured. The results of the across measure showed that the performances significantly decrease with an increase of $\Delta F0$ (see Figure 7 in section I.3.1.1). The performances decreased strongly from 100 to 162 Hz (i.e., from 0 to 8.5 semitones above 100 Hz) and the slopes decreased for larger $\Delta F0$. This result suggests that the main effect of a fundamental frequency difference should be observed between 0 to 9 semitones. The $\Delta F0$ tested in the present study were precisely between 0 and 9 semitones. Thus, according to the results obtained by Gaudrain *et al.* (2007), an increase of $\Delta F0$ above 9 semitones should not strongly enhance the effect on obligatory stream segregation[5].

The large variability among listeners might explain the non-significant effect in the RAS measure. The task was presumably too difficult to obtain consistent results. Some solutions might be considered to make the task easier, and thus to reduce the standard errors in the results. First, a visual cue could be displayed on the screen when the repeat might occur (for example between the 12^{th} and the 24^{th} stimulus) to catch the attention of the listener. Another option would be to introduce a second repeat during the sequence to increase the chances of detection.

2.8 Conclusions

- The results of this study indicated that a difference of fundamental frequency ($\Delta F0$) can induce stream segregation of speech stimuli consisting of a fricative consonant (mostly high frequency) and a vowel (mostly low frequency).

[5]The rate of the sequences were comparable between the present study and the study of Gaudrain *et al.* (2007). Indeed, the TRT was either 175 or 135 ms in the previous study and it was 200 ms (160 ms sound duration plus 40 ms of silence between the stimuli) in the present study. Thus, the results of the two studies are comparable

- The procedure of repeat detection allowed to measure obligatory and voluntary stream segregation, depending on where the repeat was introduced. The results showed that the segregation of the streams became easier, and that the integration tended to become harder, as $\Delta F0$ increased from 0 to 9 semitones.
- For both measures (voluntary and obligatory streaming), the results suggested that a fundamental frequency difference comprised between 3 and 5 semitones led to a maximum of bi-stable percept. This F0 difference could be used in a future experiment to investigate the extent to which spatial cues could bias this ambiguous percept towards segregation.

General conclusions

The aim of this PhD thesis was to investigate to what extent differences in spatial locations could influence the perceptual organization of sequential speech sounds. It is worth noting that such a difference implies several phenomena. First, colouration corresponds to the fact that the spectrum of the sounds produced at the listener's ears depend on the listener and speaker positions. In fact, the sounds are submitted to multiple frequency-dependent reflections during their propagation, and the sounds at each ear result from the addition of these filtered reflections and the direct sound. Second, sounds coming from different location in space present binaural cues: interaural time and level differences (ITD and ILD). And third, these binaural cues allow for sound localization. Thus, a difference in spatial position induces differences in colouration, binaural cues and perceived position.

Our aim was to assess speech signals, thus, several steps were considered to approach the different characteristics of speech. On average, the long-term speech spectrum is approximately flat from 100 to 1000 Hz, and then decreases by around 20 dB per octave. The high frequency bands in the 2-4 kHz region correspond to the fricative consonant sounds while the low frequency bands correspond to the vowel sounds. As a first approximation of speech, the stimuli used were frozen speech-shaped noises which had a long-term spectrum similar to speech. However, running speech present large acoustical variability which enables to convey information. Thus, the variable nature of speech was assessed using pure tones containing spectral variability. The following step consisted of using recorded speech which pitch contour had been flattened. All the sounds used were a concatenation of a fricative consonant and a vowel, and were parts of real words. That allowed us to study unfrozen stimuli with speech spectrum and a fundamental frequency. A further step would be to conserve the pitch contour of the recorded sounds, thus

the voices could sound more “natural”.

The main results indicated that stream segregation of speech-shaped noises can occur based on spatial differences. Especially, we found that the slight monaural spectral differences induced by head and room colouration can induce segregation, and that this segregation can be enhanced by adding binaural cues. We also showed that the monaural intensity variations across time at each ear were predominant for stream segregation compared to the interaural level differences, and that the interaural time differences can also favour segregation. Moreover, the percept of lateralization - associated with a given ITD - was shown to be relevant for stream segregation.

When we introduced spectral variability within sequences of pure tone stimuli, the perceptual organization was disturbed. Two types of segregation were observed, the segregation for which each of the two streams was fully coherent (without any sound outside these two streams), and the segregation for which each single sound could generate an independent stream. These results raised the question of the bias introduced by a subjective procedure where the experimenter is not able to control the strategy used by the listeners to perform the task. According to some results, it seemed that the listeners became less sensitive to a frequency difference between the stimuli to segregate the streams as the spectral variability within the streams increased. However, these results have to be confirmed using an objective task.

Finally, the experiment using speech sounds indicated that voluntary segregation tended to become easier, and integration tended to become harder, when the frequency separation between the streams increased. The results suggested that a frequency difference of about 4 semitones between the streams would lead to a maximum of bi-stable percept. This last study raised several questions, and especially the investigation of the bi-stable percept. How could this perceptual point be biased towards segregation or integration? Can the consistency of spatial differences be used by the auditory system as a regular cue to organize the streams?

L'objectif de cette thèse était d'étudier dans quelle mesure des différences de positions pouvaient influencer l'organisation perceptive de séquences de signaux de parole. Il est important de noter qu'une telle différence implique plusieurs phénomènes différents. Tout d'abord, la coloration correspond au fait que le spectre des sons produits au niveau de l'oreille de l'auditeur dépend des positions respectives de l'auditeur et du locuteur. En effet, les sons sont soumis à de multiples réflexions qui dépendent de la fréquence pendant leur propagation. Les sons à chaque oreille résultent de l'addition de toutes ces réflexions filtrées et du son direct. De plus, les sons provenant de différentes positions dans l'espace présentent des indices binauraux: des différences interaurales de temps et de niveau (ITD et ILD). Finalement, ces indices binauraux permettent la latéralisation des sons. Ansi, une différence de positions induit des différences de coloration, d'indices binauraux et de position perçue.

Notre objectif était d'utiliser des signaux de parole, et pour cela, plusieurs étapes préliminaires ont été considérées pour approcher les différentes caractéristiques de la parole. En moyenne, le spectre de la parole est approximativement plat de 100 à 1000 Hz, puis décroit d'environ 20 dB par octave. Les hautes fréquences (entre 2 et 4 kHz) correspondent aux sons de consonnes fricatives alors que les basses fréquences correspondent aux sons de voyelles. En première approximation de la parole, nous avons utilisé des "speech-shaped noises" gelés dont le spectre est similaire à celui de la parole. Cependant, la parole présente une grande variabilité de ses paramètres acoustiques, ce qui permet de transmettre de l'information. Ainsi, le caractère variable de la parole a été étudié en utilisant des sons purs avec de la variabilité spectrale. Pour l'étape suivante, de la parole enregistrée a été utilisée. Pour ces stimuli, l'enveloppe de hauteur a été aplati. Tous les sons utilisés était constitués d'une consonne fricative associée à une voyelle. Ils étaient tous extraits de mots réels. Cette étape a permis d'étudier des stimuli non gelés avec un spectre de parole et une fréquence fondamentale. La prochaine étape serait de conserver l'enveloppe de hauteur des sons enregistrés, pour que les voix sonnent de manière plus naturelles.

Les principaux résultats ont montré que la ségrégation séquentielle de "speech-shaped noises" peut être induite par des différences spatiales. En particulier, nous avons montré que les fines différences spectrales monaurales dues à la coloration de la tête et de la salle peuvent induire de la ségrégation. Cette ségrégation peut être renforcée par l'ajout d'indices binauraux. Nous avons également mis en évidence que les variations monaurales d'intensité au cours du temps à chaque oreille avaient un rôle prédominant pour la ségrégation en comparaison avec les différences interaurales de niveau, et que les différences interaurales de temps peuvent également favoriser la ségrégation. Par ailleurs, il a été montré que le percept de latéralisation - associé à

un ITD donné - était pertinant pour la ségrégation.

Lorsque de la variabilité spectrale a été introduite au sein de séquences de sons purs,l'organisation perceptive a été dérangée. Deux types de ségrégation ont été observées, la ségrégation pour laquelle chacun des deux flux était pleinement cohérent (sans aucun sons en dehors de ces flux), et la ségrégation pour laquelle chaque son pouvait générer un flux indépendant. Ces résultats soulèvent la question du biais introduit par une procédure de mesure subjective, où l'expérimentateur ne peut pas contrôler la stratégie utilisée par l'auditeur pour faire la tâche. Les résutlats indiquent que certains auditeurs étaient moins sensibles à une différence de fréquence entre les stimuli pour les séparer quand la variabilité spectrale intra-flux augmentait. Cependant ces résultats devraient être confirmés par une tâche objective.

L'expérience utilisant des signaux de parole a montré que la ségrégation volontaire des flux avait tendance à devenir plus facile, et que l'intégration avait tendance à devenir plus difficile, lorsque la différence de fréquence augmentait entre les flux. Les résultats suggèrent qu'une différence de fréquence de 4 demi-tons environ entre les flux devrait conduire à un maximum de bistabilité. Comment ce percept ambigu peut-il être orienté vers de la ségrégation ou de l'intégration ? Est-ce que la constance des différences spatiales peut être utilisée par le système auditif comme un indice régulier pour organiser les flux ?

Abstract - Résumé

Abstract

In a context of competing sound sources, the auditory scene analysis aims to draw an accurate and useful representation of the perceived sounds. Solving such a scene consists of grouping sound events which come from the same source and segregating them from the other sounds. This PhD work intended to further our understanding of how the human auditory system processes these complex acoustic environments, with a particular emphasis on the potential influence of spatial cues on perceptual stream segregation. All the studies conducted during this PhD endeavoured to rely on realistic configurations.

In a real environment, the diffraction and reflection properties of the room and the head lead to distortions of the sounds depending on the source and receiver positions. This phenomenon is named colouration. Speech-shaped noises, as a first approximation of speech sounds, were used to evaluate the effect of this colouration on stream segregation. The results showed that the slight monaural spectral differences induced by head and room colouration can induce segregation. Moreover, this segregation was enhanced by adding the binaural cues associated with a given position (ITD, ILD). Especially, a second study suggested that the monaural intensity variations across time at each ear were more relevant for stream segregation than the interaural level differences. The results also indicated that the percept of lateralization associated with a given ITD helped the segregation when the lateralization was salient enough. Besides, the ITD *per se* could also favour segregation.

The natural ability to perceptually solve an auditory scene is relevant for speech intelligibility. The main idea was to replicate the first experiments with speech items instead of frozen noises. A characteristic of running speech is a high degree of acoustical variability used to convey information. Thus, as a first step, we investigated the robustness of stream segregation based on a frequency difference to variability on the same acoustical cue (i.e., frequency). The second step was to evaluate the fundamental frequency difference that enables to separate speech items. Indeed, according to the limited effects measured in the two first experiments, it was assumed that spatial cues might be relevant for stream segregation only in interaction with another "stronger" cue such as a F0 difference.

The results of these preliminary experiments showed first that the introduction of a large spectral variability introduced within pure tone streams can lead to a complicated percept, presumably consisting of multiple streams. Second, the results suggested that a fundamental frequency difference comprised between 3 and 5 semitones enables to separate speech item. These experiments provided results that will be used to design the next experiment investigating how an ambiguous percept could be biased toward segregation by introducing spatial cues.

Keywords: Auditory scene analysis, sequential segregation, spectral differences, spatial cues, speech sounds

Résumé

Dans un contexte sonore constitué de plusieurs sources sonores, l'analyse de scène auditive a pour objectif de dresser une représentation précise et utile des sons perçus. Résoudre ce type de scènes consiste à regrouper les sons provenant d'une même source et de les séparer des autres sons. Ce travail de thèse a eu pour but d'approfondir nos connaissances du traitement de ces scènes auditives complexes par le système auditif. En particulier, il s'agissait d'étudier l'influence potentielle des indices spatiaux sur la ségrégation. Une attention particulière a été portée tout au long de cette thèse pour intégrer des éléments réalistes dans toutes les études menées.

Dans un environnement réel, la salle et la tête entraînent des distorsions des signaux de parole en fonction des positions de la source et du récepteur. Ce phénomène est appelé coloration. Comme première approximation de la parole, des bruits avec un spectre de parole ont été utilisés pour évaluer l'effet de la coloration. Les résultats ont montré que les fines différences spectrales monaurales induites par la coloration due à la tête et à la salle peuvent engendrer de la ségrégation. De plus, cette ségrégation peut être renforcée en ajoutant les indices binauraux associés à une position donnée (ILD, ITD). En particulier, une deuxième étude a suggéré que les variations monaurales d'intensité au cours du temps à chaque oreille étaient plus utiles pour la ségrégation que les différences interaurales de niveau. Les résultats ont également montré que le percept de latéralisation, associé à un ITD donné, favorise la ségrégation lorsque ce percept est suffisamment saillant. Par ailleurs, l'ITD *per se* peut induire de la ségrégation.

La capacité naturelle à résoudre perceptivement une scène auditive est pertinente pour l'intelligibilité de la parole. L'objectif était de répliquer ces premières expériences, donc évaluer l'influence des indices spatiaux sur la ségrégation de signaux de parole à la place de bruits gelés. Une caractéristique de la parole est la grande variabilité de ses paramètres acoustiques qui permettent de transmettre de l'information. Ainsi, la première étape a été d'étudier dans quelle mesure la ségrégation basée sur une différence de fréquence peut être influencée par l'introduction de variabilité spectrale au sein des stimuli. L'étape suivante a été d'évaluer la différence de fréquence fondamentale requise pour séparer des flux de parole. En effet, il a été supposé que des indices de position pourraient être utiles pour renforcer la ségrégation basée sur un indice plus robuste comme une différence de F0 du fait de leur stabilité au cours du temps dans des situations réelles.

Les résultats de ces expériences préliminaires ont montré que l'introduction d'une large variabilité spectrale au sein de flux de sons purs pouvait entraîner un percept compliqué, probablement constitué des multiples flux sonores. De plus, les résultats ont indiqué qu'une différence de F0 comprise entre 3 et 5 demi-tons permettait de séparer des signaux de parole. Les résultats de ces expériences pourront être utilisés pour concevoir la prochaine expérience visant à étudier dans quelle mesure un percept ambigu peut évoluer vers de la ségrégation par l'introduction d'indices de position.

Mots-clé: Analyse de scènes auditives, ségrégation séquentielle, différences spectrales, indices de position, signaux de parole

Bibliography

ANSI S3.6 (**1989**). "American national standard specification for audiometers", American National Standards Institute, New York . 50

ANSI S3.7 (**1995**). "Methods for coupler calibration of earphones", American National Standards Institute, New York . 53

Anstis, S. M. and Saida, S. (**1985**). "Adaptation to auditory streaming of frequency-modulated tones", J. Exp. Psychol.: Human Percept. Perf. **11**, 257–271. 21, 47, 83

Beauvois, M. W. and Meddis, R. (**1996**). "Computer simulation of auditory stream segregation in alternating-tone sequences", J. Acoust. Soc. Am. **99**, 2270–2280. 15

Bendixen, A., Denham, S. L., Gyimesi, K., and Winkler, I. (**2010**). "Regular patterns stabilize auditory streams", J. Acoust. Soc. Am. **128**, 3658–3666. ix, 22, 23, 24, 28, 72, 79

Bernstein, L. R. and Trahiotis, C. (**1985**). "Lateralization of low-frequency, complex waveforms: The use of envelope based temporal disparities", J. Acoust. Soc. Am. **77**, 1868–1880. 64

Boehnke, S. E. and Phillips, D. P. (**2005**). "The relation between auditory temporal interval processing and sequential stream segregation examined with stimulus laterality differences", Percept. Psychophys. **67**, 1088–1101. 19, 47

Bregman, A. S. (**1978**a). "Auditory streaming: Competition among alternative organizations", Perception and Psychophysics **23**, 391–398. 30

Bregman, A. S. (**1978**b). "Auditory streaming is cumulative", J. Exp. Psychol.: Human Percept. Perf. **4**, 380–387. ix, 20, 21, 27

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA). 8, 10, 11, 45

Bregman, A. S., Ahad, P. A., Crum, P. A., and O'Reilly, J. (**2000**). "Effects of time intervals and tone durations on auditory stream segregation", Perception and Psychophysics **62**, 626–636. viii, 16

Bregman, A. S. and Campbell, J. (**1971**). "Primary auditory stream segregation and perception of order in rapid sequences of tones", J. Exp. Psychol. **89**, 244–249. 30

Bregman, A. S. and Rudnicky, A. (**1975**). "Auditory segregation: stream or streams?", J. Exp. Psychol.: Human Percept. Perf. **1**, 263–267. 30

Buell, T. N. and Hafter, E. R. (**1991**). "Combination of binaural information across frequency bands", J. Acoust. Soc. Am. **90**, 1894–1900. 59

Carlyon, Robert, P., Cusack, R., Foxton, J. M., and Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation", J. Exp. Psychol.: Human Percept. Perf. **27**, 1393–1402. 21, 22

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears", J. Acoust. Soc. Am. **25**, 975–979. 72, 82

Collin, B. and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location and modulation depth of noise interferers", J. Acoust. Soc. Am. **134**, 1146–1159. 46

Dannenbring, G. L. and Bregman, A. S. (**1976**). "Stream segregation and the illusion of overlap", J. Exp. Psychol.: Human Percept. and Perf. **2**, 544–555. 25, 27

Darwin, C. J. and Hukin, R. W. (**1999**). "Auditory objects of attention: the role of interaural time differences", J. Exp. Psychol.: Human Percept. Perf. **20**, 617–629. 46, 47, 65

David, M., Lavandier, M., and Grimault, N. (**2014**). "Room and head coloration can induce obligatory stream segregation (L)", J. Acoust. Soc. Am. **136**, 5–8. 46, 50, 54, 56, 57, 83

Deutsch, D. (**1975**). "Two-channel listenning to musical scales", J. Acoust. Soc. Am. **57**, 1156–1160. 29

Devergie, A., Grimault, N., Gaudrain, E., Healy, E. W., and Berthommier, F. (**2011**). "The effect of lip-reading on primary stream segregation", J. Acoust. Soc. Am. **130**, 283–291. 51

Devergie, A., Grimault, N., Tillmann, B., and Berthomier, F. (**2010**). "Effect of rhythmic attention on the segregation of interleaved melodied", J. Acoust. Soc. Am. **128**, EL1–EL7. 23

Dowling, W. J. (**1973**). "The perception of interleaved melodies", Cognitive Psychology **5**, 322–337. 29, 71, 79

Edmonds, B. A. and Culling, J. F. (**2005**). "The spatial unmasking of speech: evidence for within-channel processing of interaural time delay", J. Acoust. Soc. Am. **117**, 3069–3078. 58, 66

Feddersen, W. E., Sandel, T. T., Teas, D. C., and Jeffress, L. A. (**1957**). "Localization of high-frequency tones", J. Acoust. Soc. Am. **29**, 988–991. 20, 58, 64

Flanagan, J. L. and Lummis, R. C. (**1970**). "Signal processing to reduce multipath distortion in small rooms", J. Acoust. Soc. Am. **47**, 1475–1481. 46

Füllgrabe, C. and Moore, B. C. J. (**2012**). "Objective and subjective measures of pure-tone stream segregation based on interaural time differences", Hear. Res. **291**, 24–33. 20, 47, 49, 57, 64, 65

Gardner, W. G. and Martin, K. D. (**1995**). "HRTF measurements of a KEMAR", J. Acoust. Soc. Am. **97**, 3907–3908. 52

Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2007**). "Effect of spectral smearing on the perceptual segregation of vowel sequences", Hearing Research **231**, 32–41. viii, ix, 13, 14, 30, 32, 83, 88

Gockel, H., Carlyon, R. P., and Micheyl, C. (**1999**). "Context dependence of fundamental-frequency discrimination: Lateralized temporal fringes", J. Acoust. Soc. Am. **106**, 3553–3563. 19, 46, 47

Gregory, A. H. (**1994**). "Timbre and auditory streaming", Music Perception **12**, 161–174. 29

Grimault, N., Bacon, S. P., and Micheyl, C. (**2002**). "Auditory stream segregation on the basis of amplitude modulation rate", J. Acoust. Soc. Am. **111**, 1340–1348. 25

Hartmann, W. M. and Johnson, D. (**1991**). "Stream segregation and peripheral channeling", Music Percept. **9**, 155–183. 13, 15, 18, 19, 29, 46, 47, 71, 79

Heise, G. A. and Miller, G. A. (**1951**). "An experimental study of auditory patterns", J. Acoust. Soc. Am. **64**, 68–77. 27

Iverson, P. (**1995**). "Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes", J. Exp. Psychol.: Human Percept. Psychophys. **21**, 751–763. 25, 29

Kidd, G. J., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech", J. Acoust. Soc. Am. **124**, 3793–3802. 19, 46, 47

Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane", J. Acoust. Soc. Am. **62**, 157–167. 64

Larsen, E., Iyer, N., Lansing, C. R., and Feng, A. S. (**2008**). "On the minimum audible difference in direct-to-reverberant energy ratio", J. Acoust. Soc. Am. **124**, 450–461. 46

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics", J. Acoust. Soc. Am. **49**, 467–477. 26, 50

Macmillan, N. A. and Creelman, D. C. (**2004**). *Detection theory: A user's guide* (Psychology Press). 86

McCabe, S. L. and Denham, M. J. (**1997**). "A model of auditory streaming", J. Acoust. Soc. Am. **101**, 1611–1621. 15

McFadden, D. and Pasanen, E. G. (**1976**). "Lateralization at high frequencies based on interaural time differences", J. Acoust. Soc. Am. **59**, 634–639. 59

Micheyl, C. and Oxenham, A. J. (**2010**). "Objective and subjective psychophysical measures of auditory stream integration and segregation", J. Assoc. Res. Otolaryngol. **11**, 709–724. 34, 56

Middlebrooks, J. C. and Green, D. M. (**1991**). "Sound localization by human listeners", Ann. Rev. Psychol. **42**, 135–159. 17

Middlebrooks, J. C. and Onsan, Z. A. (**2012**). "Stream segregation with high spatial acuity", J. Acoust. Soc. Am. **132**, 3896–3911. 29, 57

Miller, G. A. and Heise, G. A. (**1950**). "The trill threshold", J. Acoust. Soc. Am. **22**, 637–638. 27

Moore, B. C. J. (**2007**). *An Introduction to the Psychology of Hearing - Fifth edition* (Elsevier Academic Press, London, UK). 17, 47

Moore, B. C. J. and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", J. Acoust. Soc. Am. **74**, 750–753. 13

Moore, B. C. J. and Gockel, H. (**2002**). "Factors influencing sequential stream segregation", Acta Acustica United with Acustica **(88)**, 320–332. 45

Moore, B. C. J. and Gockel, H. E. (**2012**). "Properties of auditory stream formation", Phil. Trans. R. Soc. B **(367)**, 919–931. 45

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., and Winkler, I. (**2001**). "'primitive inteligence' in the auditory cortex", Trends Neurosci. **24**, 283–288. 72

Plack, C. J. (**2005**). *The sense of hearing* (Psychology Press). viii, ix, 11, 12, 29

Plomp, R. (**1964**). "Rate of decay of auditory sensation", J. Acoust. Soc. Am. **36**, 277–282. 46

Pressnitzer, D. and Hupé, J.-M. (**2006**). "Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization", Current Biology **16**, 1351–1357. 9

Rayleigh, L. (**1907**). "On our perception of sound direction", Philos. Mag. **13**, 214–232. 17

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband", J. Acoust. Soc. Am. **112**, 2074–2085. 25, 29, 47, 49, 50, 51, 52, 80

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2008**). "Effects of the build-up and resetting of auditory stream segregation on temporal discrimination", J. Exp Psychol.: Human Percept. Perf. **34**, 992–1006. 21, 47, 49, 83

Rogers, W. L. and Bregman, A. S. (**1993**). "An experimental evaluation of three theories of auditory stream segregation", Perception and Psychophysics **53**, 179–189. 29

Rogers, W. L. and Bregman, A. S. (**1998**). "Cumulation of the tendency to segregate auditory streams: Resetting by changes in location and loudness", Perception and Psychophysics **60**, 1216–1227. 21

Rose, M. M. and Moore, B. C. J. (**1997**). "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners", J. Acoust. Soc. Am. **102**, 1768–1778. 27

Sach, A. J. and Bailey, P. J. (**2004**). "Some characteristics of auditory spatial attention revealed using rythmic masking release", Percept. Psychophys. **66**, 1379–1387. 18, 19, 46, 47

Sandel, T. T., Teas, D. C., Feddersen, W. E., and Jeffress, L. A. (**1955**). "Localization of sound from single paired sources", J. Acoust. Soc. Am. **27**, 842–852. 65

Schwartz, A., Mc Dermott, J. H., and Shinn-Cunningham, B. (**2012**). "Spatial cues alone produce inacurate sound sesgregation: The effect of interaural time differences", J. Acoust. Soc. Am. **132**, 357–368. 9

Singh, P. G. (**1987**). "Perceptual organization of complex-tone sequences: a tradeoff between pitch and timbre?", J. Acoust. Soc. Am. **82**, 886–899. 29

Stainsby, T. H., Füllgrabe, C., Flanagan, H. J., Waldman, S. K., and Moore, B. C. J. (**2011**). "Sequential streaming due to manipulation of interaural time differences", J. Acoust. Soc. Am. **130**, 904–917. 19, 47, 49, 57, 64, 65

Stainsby, T. H., Moore, B. C. J., Medland, P. J., and Glasberg, B. R. (**2004**). "Sequential streaming and effective level differences due to phase-spectrum manipulations", J. Acoust. Soc. Am. **115**, 1665–1673. 18, 46, 49, 57

Summers, V. and Leek, M. R. (**1998**). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss", J. Speech, Lang. and Hear. Res. **41**, 1294–1306. 10

Thompson, S. K., Carlyon, R. P., and Cusack, R. (**2011**). "An objective measurement of the build-up of auditory streaming and of its modulation by attention", J. Exp. Psychol.: Human Percept. Perform. **37**, 1253–1262. 22, 49

van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tone sequences", PhD thesis, University of Technology, Eindhoven. viii, 9, 13, 15, 16, 27, 46, 50, 71, 72, 73, 79

van Noorden, L. P. A. S. (**1977**). "Minimum differences of level and frequency for perceptual fission of tone sequences abab", J. Acoust. Soc. Am. **61**, 1041–1045. 27

Vliegen, J., Moore, B. C. J., and Oxenham, A. J. (**1999**). "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task", J. Acoust. Soc. Am. **106**, 938–945. 29

Vliegen, J. and Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues", J. Acoust. Soc. Am. **105**, 339–346. 23, 27, 71

Wightman, F. L. and Kistler, D. J. (**1992**). "The dominant role of low-frequency interaurl time differences in sound localization", J. Acoust. Soc. Am. **91**, 1648–1661. 65