



Towards a binaural model for predicting speech intelligibility among competing voices in rooms

Thibaud Leclère

► To cite this version:

Thibaud Leclère. Towards a binaural model for predicting speech intelligibility among competing voices in rooms. Acoustics [physics.class-ph]. École Nationale des Travaux Publics de l'État [ENTPE], 2015. English. NNT : 2015ENTP0008 . tel-01277871v2

HAL Id: tel-01277871

<https://hal.science/tel-01277871v2>

Submitted on 27 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

DOMAINE DE RECHERCHE : Psychoacoustique, Acoustique des Salles

Préparée au Laboratoire Génie Civil et Bâtiment
Présentée par

Thibaud Leclère

Towards a binaural model for predicting speech intelligibility among competing voices in rooms

Directeur de thèse : **Mathieu Lavandier**

Directeur de thèse (HDR) : **Dominique Dumortier**

Soutenue publiquement le 9 décembre 2015

JURY

Pr. Torsten Dau	Danmarks Tekniske Universitet, Copenhague	Rapporteur
Pr. Christian Lorenzi	École Normale Supérieure, Paris	Rapporteur
Pr. Etienne Parizet	Institut National des Sciences Appliquées, Lyon	Examineur
Dr. Olivier Macherey	Laboratoire de Mécanique et Acoustique, Marseille	Examineur
Dr. Dominique Dumortier	École Nationale des Travaux Publics de l'État, Lyon	Directeur de thèse
Dr. Mathieu Lavandier	École Nationale des Travaux Publics de l'État, Lyon	Directeur de thèse

Remerciements

Je tiens en premier lieu à adresser mes plus sincères remerciements à Torsten Dau et Christian Lorenzi pour me faire l'honneur d'avoir accepté et pris le temps d'évaluer mes travaux, ainsi qu'à Olivier Macherey et Etienne Parizet pour faire partie de mon jury de thèse.

Je suis reconnaissant à Jean-Baptiste Lesort, Luc Delattre, Claude Henri Lamarque et Dominique Dumortier pour m'avoir donné l'opportunité de réaliser cette thèse au sein du Laboratoire Génie Civil et Bâtiment à l'École Nationale des Travaux Publics de l'État. Merci également à tout le personnel administratif de l'école qui accompagne les doctorants et veille au bon déroulement de leurs thèses. Je pense en particulier à Sonia Cenille, Emilie Enrico, Chantal Durand, Monique Darnand, Marie-Victoire Baussant, Francette Pignard. Merci pour vos nombreux services et conseils.

Je remercie la Société Française d'Acoustique (SFA), le Centre Lyonnais d'Acoustique (CeLyA) ainsi que l'école doctorale MEGA pour avoir financé des conférences nationales et internationales pendant mon doctorat, me permettant de présenter mes travaux à la communauté scientifique.

Je remercie John Culling et Mickael Deroche pour avoir collaboré avec moi sur mes travaux, vous avez été une aide précieuse sur le plan scientifique et très constructifs. Je remercie également Nicolas Grimault pour ses nombreux conseils, pour les riches discussions que nous avons pu avoir, pour sa bonne humeur permanente et pour son aisance naturelle.

Je remercie également les doctorants de l'école (et de Gerland !) avec qui j'ai pu partager de nombreux moments sportifs et festifs : Pierre, Marine, Kévin, Lucas, Salvo, Étienne, Guillaume, Clem, Fred, Nicole, Mathieu, Diego, Abhilash, Clélia, Adriana, Erij, Maïté et j'en oublie... Cette cohésion a été importante pour moi donc j'imagine qu'elle l'était pour vous.

Un esprit de labo est important et en particulier durant trois ans de travail de recherche. De près ou de loin, vous avez contribué à ce que ces années soient une réussite sur le plan humain, ce dont je me souviendrai longtemps. Je vous dis donc merci à tous : Fulbert, Achim, Guillaume, Riccardo, Mathias, Niko, Letizia, Sophie, Manu, Pascale, Joachim, Cathy, Arnaud, Thierry, Joris, Dominique, Raphaël... et tous ceux qui étaient de passage (et que j'ai pu oublier). Une pensée particulière à ma frangine de thèse : Marion, tu m'as beaucoup apporté sur le plan professionnel, culturel et personnel. Je te suis très reconnaissant des moments passés et tu sais à quel point j'apprécie la chercheuse et l'amie que tu es.

Ces trois ans ont également été l'occasion de tisser des liens forts et de belles complicités avec les étudiants de l'école via le sport, les TPs, mes tests d'écoute ou autre... Merci donc à Benjamin, Laurent, Stan, Greg, Vano, Soi, Eric, Lucas, Mél, Marine et Emilie d'avoir rendu ces années si mémorables. Je remercie de manière générale tous les volleyeurs et volleyeuses que j'ai pu croiser et/ou avec qui j'ai pu jouer pour le plaisir ou en compétition, c'était un plaisir chaque semaine. Je remercie au passage Gérard Taboulet pour son investissement dans la pratique du sport à l'école aussi bien pour les étudiants que les permanents. Un immense merci également à tous les étudiant(e)s qui ont participé à mes tests d'écoute, vos oreilles sont à la source de mon travail et m'ont donc été très précieuses.

Je tiens à remercier David pour m'avoir accompagné dans mes travaux de thèse le temps d'un stage rempli d'expériences (dans tous les sens du terme) et lui souhaite le meilleur pour la suite. J'en viens enfin à remercier du fond du coeur mon directeur de thèse, Mathieu Lavandier pour la confiance qu'il m'a accordée durant ces années. Travailler avec toi a été un réel plaisir pendant trois ans. Tu as été disponible, réceptif, attentif, constructif, motivant et optimiste. J'ai énormément appris à tes côtés et, au-delà de tes qualités scientifiques, ton humanité a fait que tu es bien plus qu'un directeur de thèse à mes yeux aujourd'hui. Outre les sciences, nous avons échangé sur bien des plans et partagé plus que je n'aurais imaginé. Je repars de Lyon avec bien plus qu'un diplôme en ayant le sentiment d'une expérience profondément enrichissante et une énorme motivation pour continuer de travailler avec toi. Merci pour toutes les digressions et discussions qu'on a pu avoir (et qu'on aura)... merci pour tout.

Je remercie ceux qui ont fait de mon parcours ce qu'il est aujourd'hui. Merci Bruno pour ces années de guitare et pour m'avoir transmis cette passion. Merci Gianni pour ton soutien dans mon parcours, pour ta bonne humeur et pour ton intérêt envers recherches. Je remercie Mathieu Paquier et Vincent Khoel pour m'avoir encouragé vers cette poursuite en doctorat. Enfin merci Rozenn pour m'avoir fait confiance en stage et pour avoir déclenché en moi cette envie de poursuivre en thèse. Tu m'as dit avec une grande honnêteté avoir « un bon feeling » à propos de cette offre à Lyon en juillet 2012... je ne saurai te dire à quel point tu as eu raison.

Ces trois années n'auraient pas été les mêmes sans vous. J'ai une chance énorme de pouvoir compter sur des potes comme vous : savoir dès le réveil qu'on va se marrer dans la journée est un luxe que j'ai eu la chance d'avoir auprès de vous depuis quelques années maintenant. Du fond du coeur, merci Orso, Charles, Yann, Romain, Momo, Rémi, Ben pour toutes ces années à vos côtés. Que vous ayez été à quelques arrêts de métro ou à des centaines de kilomètres, votre amitié était précieuse et indispensables dans les moments difficiles. Merci également à tous mes Gigaoctets de musique et mon casque qui m'ont accompagné toutes ces années et en particulier ces trois derniers mois.

Je remercie toute ma famille, mes cousins/cousines, mes oncles et tantes mais surtout mon frère Pierre, ma soeur Ségolène et mes parents. Vous m'avez toujours soutenu et encouragé quelles que soient les épreuves, vous m'avez énormément apporté malgré les distances et c'est en grande partie grâce à vous que j'en suis là aujourd'hui, à m'épanouir au quotidien. Enfin, je remercie ceux qui me regardent sûrement de là-haut, qui m'ont beaucoup transmis et que je n'oublie pas...

Je remercie enfin celle qui a su me supporter, me soutenir et m'encourager pendant ce travail de recherche et en particulier pendant la rédaction de ce manuscrit. Hélène, du fond du coeur, je te remercie pour tout ce que tu m'apportes chaque jour, pour l'affection que tu me témoignes ainsi que pour ton soutien et tes encouragements inconditionnels pour mes travaux. Merci pour tous les moments passés et à venir. Ta joie de vivre, ton sourire, ta compréhension, tes rires, ton humour, ton regard, ta tendresse et ton naturel occupent une place privilégiée dans mon coeur et font de moi un homme heureux qui, selon les Beatles, « a tout ce dont il a besoin »...

« La vie est un mystère qu'il faut vivre et non une énigme à résoudre. »
Gandhi

Acronymns, symbols and notations

AI	Articulation Index
BMLD	Binaural Masking Level Difference
EC	Equalization-Cancellation
F0	Fundamental Frequency
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
MR	Masking Release
MTF	Modulation Transfer Function
RMSE	Root Mean Square Error
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SRT	Speech Reception Threshold
STI	Speech Transmission Index
TMR	Target-to-Masker Ratio
USM	Upward Spread of Masking
$\bar{\epsilon}$	Absolute mean error
$\overline{F0}$	Mean F0
$\Delta F0$	F0 difference
$\overline{\Delta F0}$	Mean F0 difference
r	correlation coefficient
ϕ	interaural phase difference
ρ	interaural coherence

Table of Contents

Introduction	1
I State of the art	3
1 The cocktail party problem	3
2 Speech intelligibility in noise	4
2.1 Definitions	4
2.2 Unmasking mechanisms	6
2.2.1 Spatial unmasking	7
2.2.2 Temporal dip listening	10
2.2.3 F0-segregation	11
2.3 Effects of reverberation	13
2.3.1 Intrinsic influence of reverberation	14
2.3.2 Reverberation and unmasking mechanisms	17
3 Speech intelligibility models	19
3.1 Monaural models	19
3.1.1 Articulation Index (AI)	19
3.1.2 Speech Intelligibility Index (SII)	21
3.1.3 Rhebergen and Versfeld	22
3.1.4 Speech Transmission Index (STI)	22
3.1.5 Copenhagen model	24
3.2 Binaural models	27
3.2.1 Van Wijngaarden and Drullman (2008)	27

3.2.2	Madison model	27
3.2.3	Oldenburg model	29
3.2.4	Boston model	32
3.2.5	Cardiff/Lyon model	34
3.3	Summary	37
4	Aims of the PhD	38
 II Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation		41
1	Introduction	41
2	The integrated model	44
2.1	Model structure	44
2.2	Early/Late separation parameters	45
3	Validation of the room-dependent model and definition of the room-independent model	47
3.1	Data from the literature	47
3.1.1	Temporal smearing and spatial unmasking	47
3.1.2	Binaural de-reverberation	47
3.2	Model parameters and performance criteria	48
3.3	Results	48
4	Discussion	52
5	Room-Independent model validity	56
5.1	Experimental data	56
5.2	Scores transformation	56
5.3	Results	57
5.4	Discussion	57
6	General Discussion	59
6.1	Limitations of the U/D approach	59
6.2	Unified interpretation of spatial unmasking, temporal smearing and binaural de-reverberation	60
7	Conclusion	62
 III Speech intelligibility for a target and masker with different spectra		63
1	Introduction	63
2	General methods	65

2.1	Design of the stimuli	65
2.1.1	Target sentences	65
2.1.2	Maskers	65
2.1.3	Filters	66
2.2	Tested conditions	66
2.3	Measurement of SRTs	67
2.4	Determination of floor and ceiling values	68
2.5	Equipment	68
2.6	Listeners	68
3	Results	69
3.1	HP target	69
3.2	LP target	70
3.3	HP masker	71
3.4	LP masker	72
4	First discussion	73
5	Modelling	75
5.1	New implementations	75
5.2	Model performance	78
5.3	Preliminary test of backward compatibility	82
6	Second discussion	84
7	Conclusions and perspectives	87

IV Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location 89

1	Introduction	89
2	General Methods	91
2.1	Stimuli	91
2.1.1	Target sentences	91
2.1.2	Maskers	92
2.2	Procedure	93
2.3	Equipment	95
2.4	Listeners	95
3	Experiment 1	95
3.1	Aim and design	95
3.2	Results	95

Table of contents

3.3	Discussion	96
3.3.1	Influence of the target F0 contour	96
3.3.2	Benefit of spatial separation	97
3.3.3	Benefit of F0 differences	99
4	Experiment 2	101
4.1	Aim and design	101
4.2	Results	102
4.3	Discussion	104
4.3.1	Comparison with experiment 1	104
4.3.2	Benefit of envelope modulations of the buzz	105
5	Conclusions	108
General conclusions and perspectives		111
Résumé en français		115
1	Introduction	115
2	Intelligibilité de la parole dans le bruit	116
2.1	Définitions	116
2.2	Mécanismes et modèles	118
2.2.1	Démasquage spatial	118
2.2.2	Écoute dans les creux de modulation	119
2.2.3	Ségrégation par F0	119
2.3	Effet de la réverbération	120
2.4	Modèles d'intelligibilité	120
3	Modèle de prédiction pour les sources distantes	121
4	Intelligibilité pour des sources de spectres différents	123
5	Intelligibilité de la parole en présence de masqueurs harmoniques	125
6	Conclusions et perspectives	126
Bibliography		129

Introduction

More than ever, communication occupies a key place in the nowadays society. Understanding speech is particularly crucial for social interactions, security (e.g., alert messages through public address systems) or accessibility to buildings or transportations. Speech intelligibility may be strongly disrupted by the presence of noise sources or other competing conversations in enclosed spaces, which might lead to an increase of listening effort, annoyance or tiredness.

To improve speech intelligibility in noisy situations, it is then necessary to understand the different auditory and cognitive mechanisms operating at the different stages of the auditory pathway while listening to speech disturbed by either surrounding noises or by competing voices. If modelling these mechanisms can lead to accurate predictions of speech intelligibility in many situations with a limited number of parameters, buildings could be designed in order to provide good listening conditions to the users and algorithms for hearing aids could be developed to improve deficient mechanisms for hearing-impaired people.

The scope of this PhD is limited to some auditory mechanisms and to normal-hearing people. Even at the peripheral level, many years of research in hearing, room acoustics, and psychoacoustics have allowed to identify acoustic properties of the target speech source, the masking source and the room which can influence speech intelligibility. Several models aiming to predict speech intelligibility emerged from these different studies. At first, only basic situations of speech disturbed by a single noise source were considered by monaural models. Many scientific studies, including this PhD work, are interested in extending these models to more complex and realistic speech perception situations encountered in everyday life, i.e. to consider multiple competing voices located in space instead of only ambient noise. To converge towards such a model, it is first necessary to investigate if the mechanisms considered in the

case of noises are still relevant in the presence of masking voices which present different acoustic properties (fundamental frequency, intonation, envelope modulations). In addition, it has to be determined if these new acoustic properties could activate other auditory mechanisms than those operating while listening to speech in noise. The acoustic properties of the room also need to be investigated in order to determine to which extent reverberation can influence speech intelligibility.

This manuscript is composed of four chapters. A state of the art regarding the scientific knowledge concerning speech intelligibility in noise is first presented in chapter I by describing the auditory mechanisms involved and the different existing models. Chapters II, III and IV present three studies where speech intelligibility was investigated by considering the influence of the room and by progressively transforming the noise maskers into “speech-like” maskers, converging towards cocktail-party situations. The influence of the room on speech intelligibility is first examined in chapter II by extending the model of Lavandier and Culling (2010) to the case of a reverberant target. Then, differences between the target and masker spectra are introduced in chapter III since they rarely match in real life. Some additional acoustic cues available with speech maskers and the associated auditory mechanisms (F0 segregation, temporal dip listening and spatial unmasking) were investigated in chapter IV by examining to which extent listeners can benefit from each mechanism: do they interact? Are they independent? Finally, general conclusions about the different studies are summarized by highlighting the main scientific findings and potential perspectives for future scientific work are suggested.

Chapter I

State of the art

1 The cocktail party problem

In a crowded room, a listener can encounter difficulties in extracting and understanding a target speaker surrounded by competing voices. Such a situation has been previously referred to as the “cocktail party problem” ([Cherry, 1953](#)). The complexity of this problem results from the different forms of masking at a peripheral level, and from disturbances in higher cognitive processes, e.g. attentional effects, voice recognition, linguistic confusion, etc...

Energetic masking results from the physical overlap between target and masker acoustic signals at the periphery of the auditory system ([Durlach *et al.*, 2003](#)). The more energetic the masking signal in the cochlea or auditory nerve, the more difficult to extract the target signal from the acoustic mixture. The overlap can occur in the time domain, frequency domain or in the modulation domain.

Modulation masking occurs when an amplitude-modulated source prevents the detection of the temporal fluctuations in a target signal, which could be relevant for speech perception ([Houtgast and Steeneken, 1973](#)). In the presence of envelope-modulated target and masker, the ability to detect the target modulations is impaired by a masker which presents similar modulations properties. [Bacon and Grantham \(1989\)](#) investigated the influence of the frequency and depth of modulations of the masking source as well as the masker/target phase relationship on the target modulation detection. They measured thresholds for detecting the sinusoidal modulation of a target broadband white noise in the presence of another sinusoidally modulated masking white noise. They observed that modulation masking was the most effective when target and masker modulation frequencies were close to each other. Similarly to tone-on-tone masking, these results would indicate a selectivity in the modulation-frequency domain.

Even if the masking signal does not overlap across frequency, time or modulation frequency with the target signal, some informational masking can occur. Higher cognitive mechanisms can be disrupted by the presence of competing voices which convey intelligible discourses. The attention of the listener may switch from one source to another, leading to some masking effect even though the sources are both audible. Disturbances also occur on the ability to gather the speech information delivered by the target voice and perceive them as a single stream along

time. Informational masking is mediated by the similarity of the characteristics of the target and competing voices (Brungart *et al.*, 2001), the language used by the masking speaker (Rhebergen *et al.*, 2005) and other features related to high-level mechanisms, e.g. the semantic content of the discourse.

This PhD thesis will focus on energetic masking occurring at a peripheral level and will use noise masker (except if mentioned differently) to avoid informational masking.

2 Speech intelligibility in noise

2.1 Definitions

To study the cocktail party problem in terms of masking effects, many researches, including this PhD work, consider the situation of a perfect locutor (the target) talking to a perfect listener (except for studies dealing with hearing impairment) through a transmission channel (which is the air and/or the room in cocktail party situations). The target speech is not perfectly transmitted to the listener because of the presence of masking sources sharing the same transmission channel as the target (Fig. I.1), i.e. only a fraction of the spoken words emitted by the locutor will be correctly understood by the listener. This fraction, often expressed as a percentage, quantifies the intelligibility of the speech target. By considering a perfect locutor and listener, speech intelligibility then depends on both the masking sources and the transmission channel.

In the case of a cocktail party situation, the speech target is disturbed by other competing voices (Fig. I.2). Informational masking aside, competing speech presents different acoustic properties as stationary noise such as a harmonic structure with a fundamental frequency (F0) and formants (F1, F2,...), variations of F0 (intonation) and a fluctuating envelope. These properties then need to be investigated in order to determine if they consist in relevant cues for the listener to unmask a speech target among one or several masking speeches.

A common approach to study speech intelligibility is then to conduct experimental measurements by making a subject listen to sentences, words or vowels (speech items) in different masking conditions by having manipulated the masker signal (sometimes, the target signal or

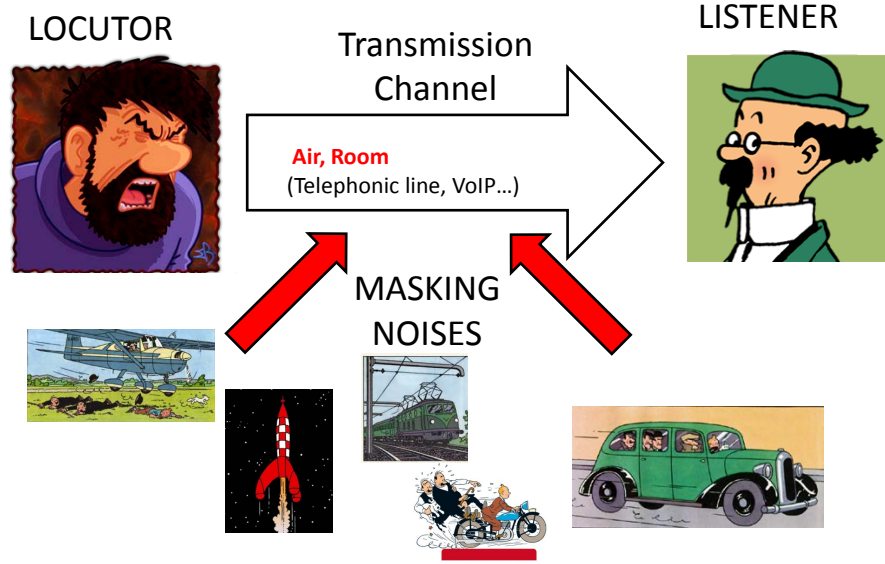


Figure I.1 – Schematic representation of situations investigated by speech in noise studies. The locutor voice (target) reaches the listener’s ears together with masking noises.

the room could be manipulated too). The subject is generally asked to report the sentence, word or vowel he/she heard, leading to a performance score which can be then compared across masking conditions. An improvement of speech intelligibility between two conditions is referred to as “masking release” (MR), “benefit” or “unmasking”.

One of the most obvious factor influencing speech intelligibility is the signal-to-noise ratio (SNR) which is the difference between the power levels of the speech target and the noise (Eq. I.1). The more energetic the target compared to the noise, the higher the speech intelligibility (Fig. I.3).

$$SNR = 10 \log \left(\frac{P_{target}}{P_{noise}} \right) = L_{P/target} - L_{P/noise} \quad (I.1)$$

with P and L_P being the power and the power level, respectively.

Speech intelligibility measurements are obtained by either comparing the correct number of items the subject understood to the total number of presented items (performance score), or

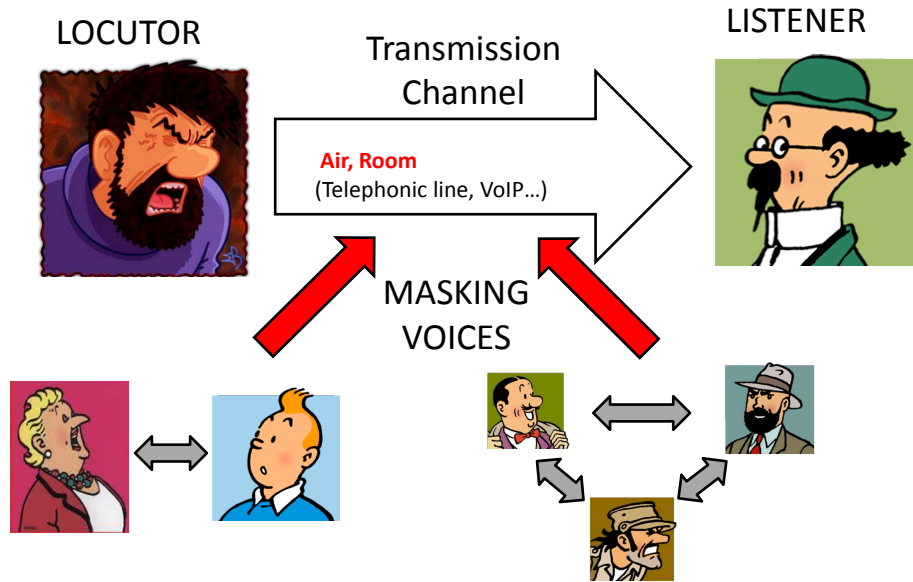


Figure I.2 – Schematic representation of a cocktail party situation, involving competing voices as the masking sources.

by measuring the SNR for which the subject performs a given intelligibility score, e.g. 50%. This last measurement is known as the speech reception threshold (SRT, see Fig. I.3) and is generally obtained by using an adaptive procedure (Levitt, 1971; Brand and Kollmeier, 2002), i.e. the SNR is varied from one item presentation to another, depending on the answer of the subject at the previous presentation. If the subject correctly answered, the task is made more difficult by decreasing the SNR at the following presentation, and conversely if the subject had difficulty to perform the task. The SRT is then obtained by averaging the SNR on a given number of the last trials. A decrease of SRT corresponds to an increase of intelligibility since a lower SNR is needed to reach the same score performance (50%).

2.2 Unmasking mechanisms

Even in the presence of noise, the auditory system can rely on certain mechanisms triggered by acoustic cues in order to unmask the speech target and then improve speech intelligibility. Three of these mechanisms are presented hereafter: spatial unmasking, temporal dip listening and F0-segregation.

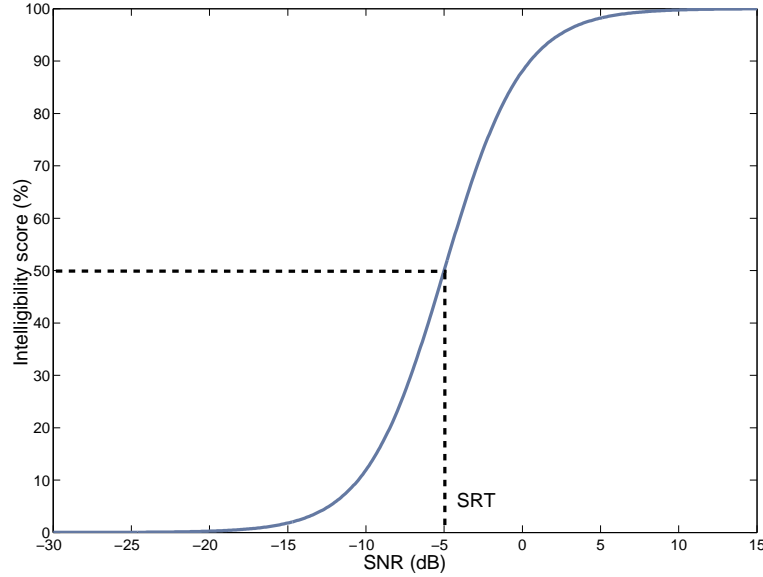


Figure I.3 – Intelligibility score as a function of SNR. The particular SNR yielding 50% intelligibility is called the speech reception threshold (SRT).

2.2.1 Spatial unmasking

Listeners better understand the speech target when it is spatially separated from the masking source. This spatial release from masking has been observed in many studies from the literature (Plomp, 1976; Shinn-Cunningham *et al.*, 2001; Culling *et al.*, 2003; Hawley *et al.*, 2004; Beutelmann and Brand, 2006; Jones and Litovsky, 2011; Rennie *et al.*, 2011; Lavandier *et al.*, 2012). For instance, Plomp (1976) conducted speech intelligibility tests with a target masked by a speech-shaped noise (SSN, noise with the same spectrum as long-term speech). While the target was reproduced through a loudspeaker located at 0° in front of the listener's head, the noise was reproduced through one of five loudspeakers distributed in the horizontal plane over the range $[0^\circ - 180^\circ]$. Figure I.4 (left panel) presents the measured masked thresholds (SNR required for just intelligible speech) as a function of noise azimuth in anechoic conditions. The highest threshold, i.e. the less intelligible condition, was obtained when target and masker sources were colocated. Beutelmann and Brand (2006) reproduced a similar experiment by using head-related impulse responses (HRTFs) to spatially separate target and noise over headphones. The measured SRTs (SNR yielding 50% intelligibility) are replotted in the right panel of Fig. I.4. Beutelmann and Brand (2006) observed the same masking release as Plomp (1976): as soon as spatial separation was introduced between the two sources, speech

intelligibility increased.

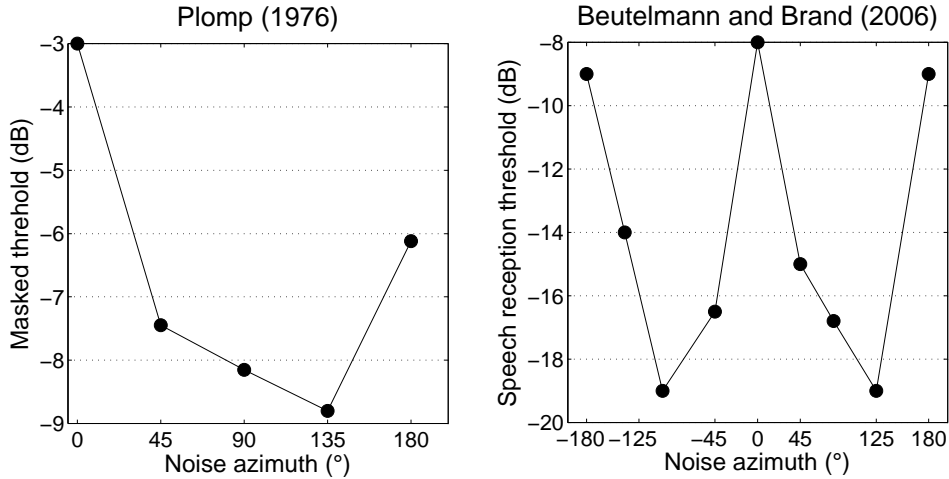


Figure I.4 – Replots from [Plomp \(1976, left panel\)](#) [*Acustica*, **34**, 200-211] and ([Beutelmann and Brand, 2006, right panel](#)), [*J. Acoust. Soc. Am.*, **127**, 2479-2497]. Masked thresholds (left panel, just intelligible speech) and speech reception thresholds (right panel, 50% intelligibility) as a function of the masker azimuth. In both studies, the speech target was located in front of the listener (0°). Spatial unmasking is illustrated by the decrease of thresholds when a spatial separation is introduced between target and masker.

A source located in the azimuthal plane (elsewhere than in front of the listener’s head) creates interaural level differences (ILDs) and interaural time differences (ITDs). These two acoustic factors are referred to as “binaural cues”. Figure [I.5](#) illustrates these differences between the acoustic signals received at each ear. ILDs arise from the acoustic head shadow which attenuates the sound level received on the contralateral ear compared to the ipsilateral one. The acoustic wave reaches the contralateral ear in a longer time than the ipsilateral ear because of the difference in distance from the source to each ear and the diffraction by the head (Fig. [I.5](#)), resulting in ITDs. Since the phase corresponds to a time shift for a given frequency, ITDs can also be expressed as interaural phase differences (IPDs). For large wavelengths (i.e. at low frequencies), the head does not constitute a major obstacle to the wave propagation and does not absorb much energy compared to high frequencies, leading to negligible ILDs at low frequencies (below about 1000 Hz). By considering a high-frequency tone, the ITD is generally larger than a period, which result in an ambiguity for the auditory system in determining which cycle in the left ear corresponds to a given cycle in the right ear. On the contrary, for

low frequencies (below about 1500 Hz), this ambiguity does not occur since the period is longer than the ITD. These two binaural cues allow the listener to localize an acoustic source in the azimuthal plane (Wallach, 1939; Moore, 2003) but they are also strong cues to unmask speech from a spatially-separated noise (Bronkhorst and Plomp, 1988) due to two binaural mechanisms: better-ear listening and binaural unmasking, which rely on ILDs and ITDs, respectively.

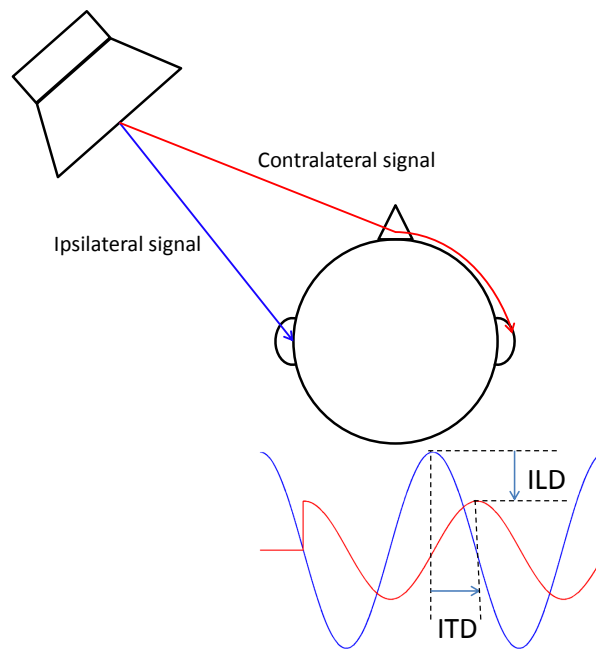


Figure I.5 – Illustration of the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) for a given position of a sound source. The acoustic signal received at the contralateral ear (in red) is delayed in time due to the longer distance for the wave to travel and to the diffraction by the head. It is also attenuated in level compared to the ipsilateral signal (in blue) because of the absorption properties of the head. Both ITD and ILD are frequency-dependent, generally increase with the azimuth angle and are roughly symmetrical around the interaural axis.

By having the target and masker at different azimuths, each source yields a different ILD in a given frequency band, leading to a difference in target-to-masker ratio (TMR) between the two ears. The auditory system is then able to attend to the ear offering the best TMR leading to an improvement of speech intelligibility compared to when target and masker are located at the same position, resulting in the same TMR at both ears. Since ILDs are reduced in low-frequency regions, better-ear listening is the most effective for high-frequency regions.

In addition to better-ear listening which relies on ILDs, masking release also occurs due to ITDs. This advantage is called “binaural unmasking” or “binaural interaction”. Licklider (1948) conducted intelligibility tests with speech against a white noise. The interaural phase of both

the speech target and the masking noise were manipulated. Between the ears, target signals were either in phase ($S0$) or out of phase ($S\pi$) and the noise was either in phase ($N0$) or out of phase ($N\pi$). An improvement of speech recognition scores (up to 30%) was observed in antiphasic ($S0N\pi$ or $S\pi N0$) conditions compared to homophasic ($S0N0$ or $S\pi N\pi$) conditions. Other studies also observed a decrease of the target detection threshold under binaural listening conditions compared to monaural conditions by introducing interaural phase differences on either the target or the masking source (e.g., [Mosko and House, 1971](#)). Interaural phase corresponds to a constant ITD for a given frequency, indicating that spatially-separated sources also lead to binaural unmasking. The amount of masking release associated to binaural unmasking is referred to as Binaural Masking Level Difference (BMLD, in dB). It corresponds to the gain which should be applied on the masking source in binaural listening to reach an equivalent masking effect as in monaural condition. Binaural unmasking is the most effective in low-frequency regions ([Culling *et al.*, 2004](#), below about 1500 Hz).

[Durlach \(1963\)](#) proposed a model of binaural unmasking called “Equalization-Cancellation” (EC) theory in order to quantitatively predict the BMLD. The basic principle of this theory is that the auditory system is able to enhance the TMR in two steps: the acoustic signal received at one ear (containing both target and masker signals) is first amplified/attenuated and translated in time until the masking signal matches at best in both ears; this is the equalization process. From there, the signal from one ear is subtracted to the signal from the other ear, leading to a cancellation of the masking signal; this is the cancellation process. Since both equalization and cancellation processes are applied on both target and masker, the resulting target signal depends on the interaural relationship of the target signal compared to that of the masking signal. For instance, the ideal case would be that the target is in phase between the ears with a noise being out of phase, the resulting target would be perfectly restituted without alteration and the noise would also be perfectly cancelled. But instead, if target and masker have similar interaural phases between the ears, the resulting target would be as cancelled as the noise, indicating that spatial separation leads to a better cancellation of the noise only, and thus to an improvement of speech intelligibility.

2.2.2 Temporal dip listening

In every day life, stationary masking noise are rarely encountered compared to noises with a fluctuating envelope. Many studies investigated the influence of these temporal fluctuations by using either deterministic envelopes ([Gustafsson and Arlinger, 1994](#); [Dubno *et al.*, 2002](#)) or speech-like envelopes ([Festen and Plomp, 1990](#); [Peters *et al.*, 1998](#); [Hawley *et al.*, 2004](#);

Beutelmann *et al.*, 2010; Collin and Lavandier, 2013). All of them observed an improvement of speech intelligibility when amplitude modulations were introduced in the envelope of the masking noise.

Gustafsson and Arlinger (1994) observed that this masking release depended on both modulation depth and rate. Maximum benefit was reached for modulation rates between 10 and 20 Hz and for modulation depth of 100%. In agreement with these findings, Bronkhorst and Plomp (1992) and also Collin and Lavandier (2013) observed that the masking release was the most effective when the noise was modulated by a 1-voice envelope. Envelopes resulting from several simultaneous voices present reduced modulation depths which lead to more masking. All these results indicate that speech intelligibility seems to be strongly related to the width and the magnitude of the temporal dips in the masker signal.

According to Festen and Plomp (1990), two mechanisms may contribute to this masking release: temporal resolution and comodulation masking release. Temporal resolution (or acuity) refers to the ability to detect changes in a sound signal over time (Moore, 2003), for example, to detect temporal gaps within the signal. Fluctuations in the masker envelope generate varying SNRs along time which can be beneficial to the listener, the low sound level of the masker allows the listener to glimpse the target signal, which is reported in the literature as the “dip-listening” or “listening-in-the-dips” effect. This ability relies on how fast the SNR varies compared to the temporal resolution of the listener. For fluctuation rates higher than this temporal resolution, listeners would “miss” the opportunities to glimpse some target signal when the masker level is low (Howard-Jones and Rosen, 1993).

Comodulation masking release occurs when the masker signal presents correlated modulation profiles across frequency channels. Festen (1993) highlighted this effect by measuring SRTs for a speech target masked by a comodulated noise (same temporal envelope in each frequency band) or by a noise presenting different envelopes in each band (which breaks down the comodulation). He observed that more masking release was obtained when the fluctuating masker was presented with correlated envelopes across frequency bands, suggesting that the auditory system could rely on an across-channel process to unmask speech in fluctuating noise.

2.2.3 F0-segregation

Speech presents a harmonic structure with a fundamental frequency (F0) and formants (F1, F2,...) which fluctuate over time. Some studies then used harmonic maskers instead of noise in order to investigate the influence of F0 on speech intelligibility and at the same time limit informational masking.

Speech intelligibility is improved when the target speech and harmonic masker present different F0s. [Brokx and Nootboom \(1982\)](#) examined the influence of this F0 difference ($\Delta F0$) using monotonized (fixed F0) and intonated (fluctuating F0) voices for both target and masker. A better word recognition was observed when F0s (or mean F0 in the intonated case) of the target and masker were separated by more than one semitone. [Bird and Darwin \(1998\)](#) confirmed these findings by using entirely voiced speech rather than natural speech. The voiced parts of speech are the speech sounds which require the vibration of the vocal cords to be produced, i.e. vowels and the voiced consonants such as, for instance, /m/, /b/, /d/ or /z/ (in English). They represent the harmonic parts of speech which justify why competing vowels were often used in F0-segregation experiments. For instance, [Summerfield and Assmann \(1991\)](#), [Culling and Darwin \(1993\)](#) and also [de Cheveigné *et al.* \(1997\)](#) presented concurrent vowels in pairs to the listener and confirmed that the pairs were identified more accurately when the F0s differed.

Different theories have been proposed to interpret this benefit based on the F0 difference between target and masker ([de Cheveigné, 1993](#)). Glimpsing is the ability to gather the spectro-temporal parts of the target within dips in the masker ([Cooke, 2003](#)). A harmonic masker with a fixed F0 presents spectral dips which could be helpful for glimpsing spectrally and listen to the target signal through these dips, providing some substantial masking release ([Deroche *et al.*, 2014](#)). This masking release is increased when target and masker F0s differ because 1) the F0 and the first formants (F1, F2) of the target do not overlap with the resolved or partially resolved partials of the masker, leading to more target signal available in between the masker partials, so then, a better target-to-masker ratio (TMR) is brought to the listener at the output of many auditory filters. 2) Increasing the masker F0 induces larger spectral dips between the partials and a greater amount of target signal could then be available.

Furthermore, some studies proposed that the auditory system would be able to exploit either the harmonicity of the masker in order to cancel it (harmonic cancellation) or that of the target in order to enhance it (harmonic enhancement, [de Cheveigné, 1993](#)). In order to determine to which extent each mechanism is used by the auditory system, [de Cheveigné *et al.* \(1995\)](#) conducted a double-vowel identification experiment in the presence of harmonic and inharmonic vowels. Their results do not support the harmonic enhancement and suggest that listeners likely rely on harmonic cancellation to extract a target source from a harmonic masker.

In the presence of noise or harmonic maskers, the auditory system is then able to rely on spectral, temporal and binaural cues at a peripheral level in order to trigger unmasking mechanisms and then reduce the influence of the masking source over the speech target in

anechoic conditions. The next section will discuss the potential influence of a room on these mechanisms.

Before implementing these mechanisms into speech intelligibility models, it is necessary to determine whether or not they interact when they operate simultaneously. This question is investigated by chapter IV through an experimental work involving spatial unmasking, temporal dip listening and F0 segregation.

2.3 Effects of reverberation

When listening to a sound source in an enclosed space, the acoustic signal reaching the listener's ears results from multiple paths due to the reflections of the acoustic waves on the room boundaries. Because of the different lengths of these paths and the frequency-dependent absorption of the room material, each reflection is a delayed version of the direct sound (unaltered wavesound travelling by the shortest path between the source and the receiver) with a modified spectrum. Figure I.6 presents a schematic temporal representation of the successive reflections received at a given location in a room (also called “echogram”). Roughly, it is composed of the direct sound, the early reflections (arriving within a short temporal window after the direct sound) and the late reflections (the most delayed and attenuated reflections, constituting a diffuse reverberated field). From a signal-processing point of view, a room can be approximated as a linear system having an impulse response (IR) linking an input (the source signal) to an output (the signal received at a given position in the room) by a convolution in the time domain:

$$o(t) = h(t) * i(t) = \int_{-\infty}^{+\infty} h(t - \tau) \cdot i(\tau) \cdot d\tau \quad (\text{I.2})$$

with $o(t)$, $i(t)$ and $h(t)$ representing the output, input and impulse response signals in the time domain, respectively. By definition, $h(t)$ represents the temporal response of the room to an impulse excitation [Dirac distribution, $\delta(t)$] for a given source/receiver configuration. In the frequency domain, Eq. I.2 becomes:

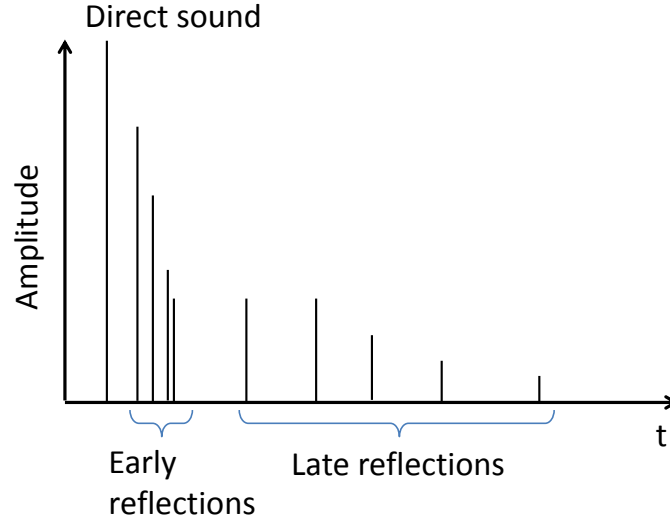


Figure I.6 – Illustration of a fictitious echogram. Each peak represents the energy of a reflection as a function of arrival time. The first peak, which is most of the time the most energetic, is called the direct sound which corresponds to the first wavefront travelling from the source to the receiver by the shortest acoustic path.

$$O(f) = H(f) \times I(f) \quad (\text{I.3})$$

with $O(f)$, $I(f)$ and $H(f)$ the respective Fourier transforms of $o(t)$, $i(t)$ and $h(t)$. $H(f)$ can be seen as a transfer function in signal-processing. Its modulus is also sometimes called “coloration” in room acoustics, psychoacoustics or hearing research. It corresponds to the alteration of the source spectrum due to the frequency-dependent absorption coefficients of the room materials and the constructive/destructive interferences between reflections.

2.3.1 Intrinsic influence of reverberation

Even in absence of masking sources in the room, reverberation influences speech intelligibility by modifying the spectro-temporal features of the speech signal received at the ears. Because of the different reflections in the room, reverberant speech is received by the listener’s ears as a succession of delayed and attenuated versions of itself. The speech signal is then temporally smeared, leading to a self-masking effect due to the temporal overlap of the successive signals emanating from the different reflections. This masking effect is called “temporal smearing of speech” and has been observed in many studies ([Lochner and Burger, 1964](#); [Bradley and](#)

Bistafa, 2002; Lavandier and Culling, 2008; Rennie *et al.*, 2011; Collin and Lavandier, 2013).

Early work from Bolt and MacDonald (1949) already proposed a statistical theory to describe the self-masking of speech due to reverberation. In this theory, they considered speech as a series of energy pulses having durations τ_1 and spaces between pulses of durations τ_2 . The sound pressure level of the pulses was uniformly distributed over a range of 30 dB. These pulses could be associated to the vowels of a speech sentence for instance. Their theory is based on the idea that one given pulse causes masking on the following pulses because of its temporal decrease of energy due to reverberation. The masking amount on a given pulse was then deduced by considering the cumulated residual energy of previous pulses as noise. By applying this approach to each pulse recurrently, the intelligibility could be estimated from the entire series of pulses by using the Articulation Index (AI, see sect. 3.1.1). But this theory has several limitations due to the number of hypothesis: speech is reduced to a series of pulses, sound pressure decrease is considered exponential and the position of the sources is not taken into account. More recent works proposed different approaches to account for the influence of reverberation on speech intelligibility.

Houtgast and Steeneken (1973) aimed to propose a new single way of quantifying the effect of the smearing of speech in reverberant environments. They focused on the fact that the smearing effect reduces the amplitude modulations in the speech signal. If this reduction could be measured and/or quantified, it could constitute a good candidate as a predictor of the reduction of speech intelligibility due to the temporal smearing of speech. By using the Modulation Transfer Function (MTF) concept, they were able to describe the modulation alterations which occurred on the source signal. MTF can be seen as the transfer function of a filter having signal envelopes as input and output (see Fig. 1.7). The way some frequency modulations are amplified or reduced in the output depends on the filter characteristics, which are related here to the room and the sources position. They showed a good correspondance (standard deviation of 4.8%) between word recognition and MTF over 68 conditions including reverberation, echoes and interfering noise making MTF a relevant factor to describe the temporal smearing of speech. This approach was implemented later in the Speech Transmission Index (STI) standard (see sect. 3.1.4).

Lochner and Burger (1964) were among the firsts to investigate the influence of early and late reflections (see Fig. 1.6) on speech intelligibility. They introduced the concept of useful-to-detrimental (U/D) ratio based on their findings showing that early reflections were integrated to the direct sound and were useful regarding speech intelligibility, whereas late reflections were responsible of the detrimental effect of the temporal smearing of speech. Eversince, many studies confirmed the useful and detrimental roles of early and late reflections, respectively (Soulodre

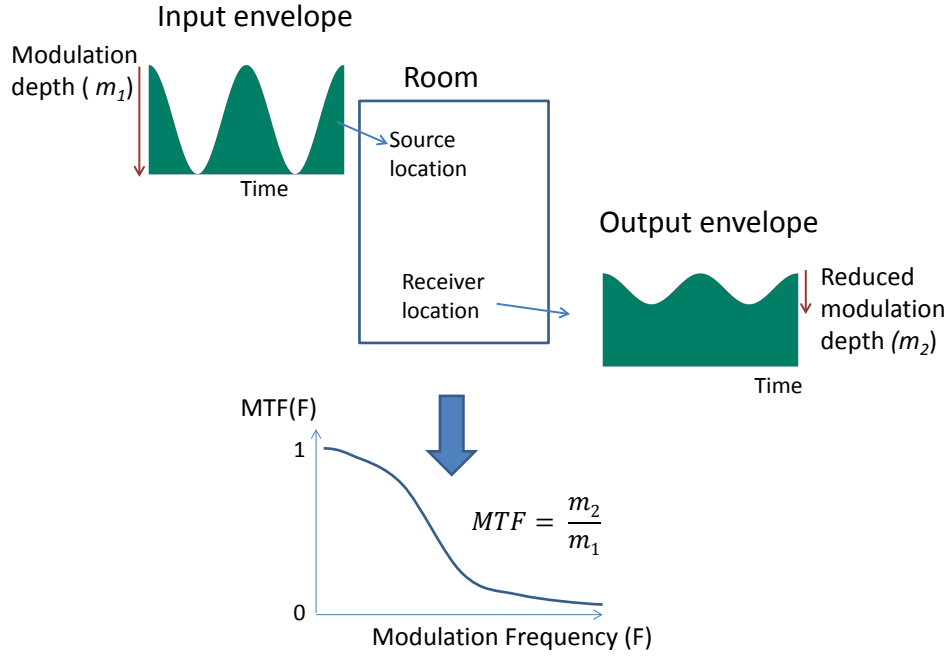


Figure I.7 – Illustration of the modulation depth reduction due to the delayed reflections in a room. The measure of this reduction for each frequency modulation constitutes the modulation transfer function (MTF) which is used as an approach to predict speech intelligibility.

et al., 1989; Bradley *et al.*, 2003; Arweiler and Buchholz, 2011; Roman and Woodruff, 2013; Warzybok *et al.*, 2013).

The listener's ears receive an acoustic signal with a modified spectrum compared to the direct sound because of 1) the frequency dependent absorption properties of the room materials and 2) the constructive/destructive interferences between reflections. This coloration reduces or amplifies the magnitude of the frequency components of both target and masking signals. This frequency weighting has direct consequences on speech intelligibility since all frequency regions do not have the same importance regarding speech intelligibility (ANSI S3.5, 1997). If important frequencies are filtered out by coloration, it would lead to a detrimental effect if the target is filtered but to a beneficial effect if the masker is filtered. Coloration can then improve or impair speech intelligibility in rooms depending on the room and the sources/listener positions.

The temporal limit defining the early reflections and the global process of how this early/late separation is operated by the auditory system remains a scientific question. The second chapter of this thesis deals with this question by proposing a binaural model which can consider the case of reverberant target by implementing a U/D approach in combination with spatially-separated sources.

2.3.2 Reverberation and unmasking mechanisms

Because the acoustic signals received at the listener's ears depend on the impulse response of the room and of the sources positions, the acoustic cues leading to the unmasking mechanisms described above could then be modified by reverberation in the time or frequency domain.

In a reverberant environment, spatial release from masking is reduced compared to anechoic conditions. [Plomp \(1976\)](#) measured intelligibility thresholds by varying the masker position (the target was kept fixed in front of the listener's head) and the reverberation time. As illustrated in figure [I.8](#), the benefit due to the spatial separation between target and masker is reduced when increasing the reverberation time, i.e. when more reflections are involved.

Both better-ear listening and binaural unmasking are affected by reverberation, resulting in less spatial masking release. The acoustic signal received at one ear arises from the direct path from the source and also from the multiple reflections on the room boundaries which have travelled around the listener's head. Because of these reflections, the interaural level difference (ILD) of both target and masker is reduced, leading to similar TMRs at each ear, which results in a poorer better-ear listening than in anechoic condition.

The interaural coherence quantifies the similarity between the left/right ear signals. Because of the multiple reflections, the source signals are modified differently when reaching the left and right ears (except if the reflection pattern is perfectly symmetrical between the ears), which reduces the interaural coherence. According to EC theory, it is more difficult to equalize (and thus cancel) an uncoherent masker ([Culling *et al.*, 2004](#)), the cancellation process would result in a residual masking signal due to a non-optimal equalization process, which causes a less effective binaural unmasking.

When listening to a speech target masked by an envelope-modulated masker in a reverberant environment, the masking release due to the envelope modulations of the masker is reduced

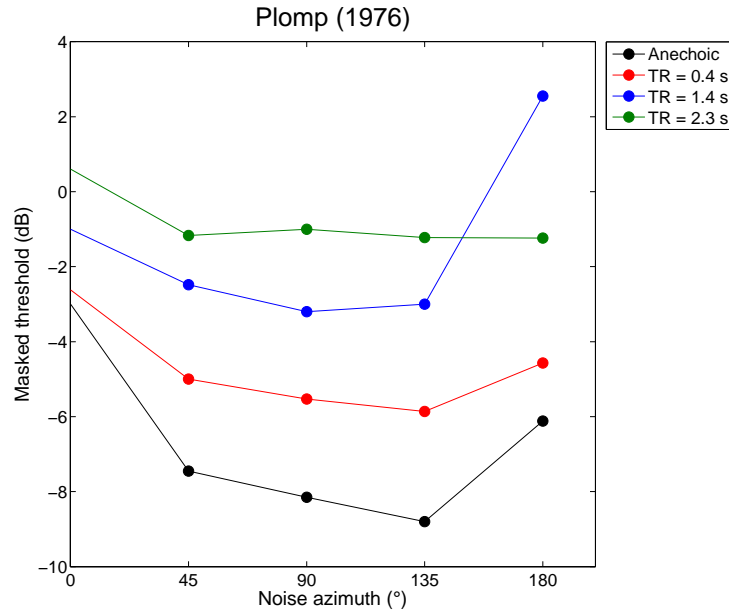


Figure I.8 – Replot from [Plomp \(1976\)](#)[*Acustica*, **34**, 200-211]. Masked threshold as a function of noise azimuth and reverberation time (RT). The speech target was in front of the listener’s head.

compared to the anechoic case ([Beutelmann *et al.*, 2010](#); [Collin and Lavandier, 2013](#)). Because of the delayed reflections, the masking signal is smeared in time, filling up the temporal dips in which the listener was able to glimpse the target signal. In addition, the frequency dependent absorption properties of the room boundaries, the constructive/destructive interferences and the different delays of the reflections would result in spectrally and temporally modified envelopes in each frequency channel. The envelope modulation across channels would then be less coherent than in the anechoic case, leading to a potential reduction of comodulation masking release.

Reverberation also influences the unmasking mechanisms based on the F0s of the sources. [Culling *et al.* \(1994\)](#) conducted vowel identification tests by presenting a vowel target against a harmonic masker with a mean F0 separated by 0 or 1 semitone from that of the target. Target and masker were both presented with either a static or sinusoidally-modulated F0 in either an anechoic or a reverberant environment. Listeners kept benefiting from a difference in F0 between the competing sources to better identify the target in reverberant conditions only for static F0s. This benefit disappeared when reverberation was introduced for target and maskers with modulated F0s. [Deroche and Culling \(2011\)](#) further investigated this interaction by conducting speech intelligibility measurements. They used sentences target and harmonic maskers separated by 2 semitones. Reverberation and F0 modulations were independently

applied on either target or masker. As [Culling *et al.* \(1994\)](#), they observed a decrease of speech intelligibility when introducing reverberation on both F0-modulated target and masker. But they could further determined that the reverberation on the masker was responsible for this detrimental effect. Because of the different delays between the acoustic reflections and the F0 fluctuations of the harmonic masker, multiple F0s would be present at a given moment which would cause an ambiguity for the harmonic cancellation process since the harmonicity of the masker is disrupted. Spectral glimpsing would also suffer from these multiple F0s at a given moment since they would filled up the spectral gaps in which listeners are able to glimpse the target signal.

3 Speech intelligibility models

Over the years, many studies tried to predict speech intelligibility by implementing the influence of acoustical factors into computational models. The existing models present different approaches to predict speech intelligibility quantitatively or qualitatively, depending on what they aim to predict (i.e. the unmasking mechanism, in anechoic or reverberant conditions, under monaural or binaural listening, etc...). A brief description of these models is presented here. They were chosen because of their close relevance to the scope of this PhD work. The intelligibility models will not be described in detail, the reader is invited to read the original publications for further details.

3.1 Monaural models

3.1.1 Articulation Index (AI)

The Articulation Index (AI) has been introduced by [French and Steinberg \(1947\)](#) and is considered to the first standard to quantify speech intelligibility ([ANSI S3.5, 1969](#)). The basic idea is that frequency bands Δf in the speech spectrum carry a certain amount ΔA related to speech intelligibility by making the assumption that the contribution of a given Δf is independent of the other frequency bands. All ΔA s can be added together across frequency bands to obtain a total amount, A , which can take values between 0 and 1. The relationship between A and speech intelligibility scores is not necessarily linear ([Kryter, 1962](#)), i.e., $A = 0.5$ does not mean that a listener would perform 50% correct at a word recognition task for example. It represents the effective proportion of speech signal conveying speech intelligibility which is available to a listener. In the case of non-optimum conditions in a given frequency band Δf , only a fraction of the maximum value of ΔA contributes to the total A value.

$$A = \sum_{n=1}^N W_n \times \Delta A_n \quad (\text{I.4})$$

with W_n representing the fraction contributing to A in the n^{th} frequency band. For convenience in the AI computations, the frequency range is divided into twenty bands (French and Steinberg, 1947) such that each ΔA_n equally contributes to A ($\Delta A_n = 0.05$). With this 20-band division, equation I.4 becomes:

$$A = \sum_{n=1}^N \frac{W_n}{20} = 0.05 \sum_{n=1}^N W_n \quad (\text{I.5})$$

French and Steinberg (1947) proposed an expression of W_n for noisy conditions derived from experimental data. It can be seen as a signal-to-noise ratio (SNR) divided by the effective dynamic range of speech (EDRS, 30 dB; Beranek, 1947) to have a W value between 0 and 1.



$$W = \frac{L_{speech} - L_{noise}}{30} \quad \text{with} \quad 0 \leq W \leq 1 \quad (\text{I.6})$$

Kryter (1962) proposed some modifications in the computation of the AI. Other frequency divisions than the twenty bands were reported: one-octave bands or third-octave bands, which correspond to a more normalized way of dividing frequencies. Kryter (1962) also provided graphical methods to compute AI taking into account the upward spread of masking (the fact that narrow-band noise can have masking effects beyond its frequency limits because of the shape of the auditory filters), and he listed different situations which can be handled by the AI:

- Fluctuating noise
- Amplitude distortion of the speech signal
- Reverberation

It should be noted that these effects are only taken into account thanks to curves and correction factors and no computational step is proposed in the original procedure. The potential masking release due to fluctuations in the masker envelope cannot be predicted by considering the long-term signals of both target and masker. Moreover, the entire speech target is used in the computation, including the detrimental part due to the late reflections of reverberation

which prevent from predicting the temporal smearing of speech. The signals from only one ear are required for the AI computation, resulting in an impossible prediction of the spatial unmasking effect. Finally, no detection of harmonicity and/or spectral dips in the masking signals are implemented in the model, which cannot account for F0-segregation.

<div style="text-align: center;">  </div> <ul style="list-style-type: none"> – Only requires target and masker signals – Simple computation 	<div style="text-align: center;">  </div> <ul style="list-style-type: none"> – No spatial unmasking – No temporal dip listening – No F0-segregation – No temporal smearing of speech
--	---

3.1.2 Speech Intelligibility Index (SII)

The Speech Intelligibility Index (SII) was adopted by the American National Standard Institute in 1997 ([ANSI S3.5, 1997](#)). The purpose was to define a computational method based on acoustical measurements which provide a metric which is highly-correlated with speech intelligibility. Like the AI, it is an index between 0 and 1 which may be interpreted as a proportion of the amount of speech information available to the listener. Its computation is based on a weighted sum of apparent SNRs (i.e. in which the hearing threshold, the upward spread of masking and other auditory features have been taken into account) across frequency bands:

$$SII = \sum_{n=1}^N I_n \cdot A_n \quad (\text{I.7})$$

where I_n is the band importance (i.e. weighting coefficients reflecting the relevant frequency bands regarding speech intelligibility) and A_n the band audibility function of the n^{th} frequency band, defined as:

$$A_n = L_n \frac{E'_n - D_n + 15}{30} \quad \text{with} \quad 0 \leq A_n \leq 1 \quad (\text{I.8})$$

with L_n the speech level distortion factor, E'_n the speech spectrum level and D_n the equivalent disturbance level (after having taken into account the upward spread of masking, the absolute hearing threshold, the internal noise and the free-field to eardrum transfer function).

Chapter I. State of the art

Four frequency-band divisions are proposed by the standard: twenty-one critical bands, seventeen equally-contributing critical bands, one-third octave bands or octave bands. The choice of the frequency bands depends on the context and the computation convenience the user is interested in. The detailed procedure for the SII computation can be found in the original standard ([ANSI S3.5, 1997](#)).

Since the input signals are the same as the AI, the SII is not able to predict the masking releases attributed to spatial unmasking, temporal dip listening or F0-segregation and, neither the masking effect of reverberation on the target speech.



- Upward spread of masking
- Hearing threshold
- Only requires target and masker spectra
- Simple computation



- No spatial unmasking
- No temporal dip listening
- No F0-segregation
- No temporal smearing of speech

3.1.3 Rhebergen and Versfeld

By considering the target and noise spectra as inputs, all temporal aspects regarding speech intelligibility are lost in the SII procedure. An extension of the SII was proposed by [Rhebergen and Versfeld \(2005\)](#) to account for the presence of fluctuating noise by calculating short-term SII values. The original SII is first determined within short time-frames, and all the SII values are then averaged across frames to result in an overall SII. [Rhebergen and Versfeld \(2005\)](#) and [Rhebergen et al. \(2006\)](#) tested this extended model on several data from the literature involving steady-state noise, speech-modulated noise, interrupted noise and sinusoidally intensity-modulated noise. For most tested data, this SII averaged over short-time frames yields good prediction of speech intelligibility in fluctuating noise. The prediction of the temporal dip listening is then achieved for this revised version of the SII. However, spatial unmasking, F0-segregation and temporal smearing of speech are still neglected by this model.



- Temporal dip listening



- No spatial unmasking
- No F0-segregation
- No temporal smearing of speech

3.1.4 Speech Transmission Index (STI)

The Speech Transmission Index (STI) was developed by [Houtgast *et al.* \(1980\)](#) who aimed to develop a relevant indicator for speech intelligibility in the context of speech transmission in rooms. Their work is based on previous studies showing that speech intelligibility in rooms is highly correlated to the ability for the listener to detect amplitude modulations in the speech signal ([Houtgast and Steeneken, 1973](#)). Figure I.7 shows a schematic representation of the concept of modulation transfer function (MTF). The temporal modulations of the target signal can be reduced because of 1) the presence of noise and 2) the delayed sound reflections in the room which interfere with the direct sound. The MTF corresponds to the reduction of the modulation depth as a function of the modulation frequency, which often behaves as a low-pass filter in real rooms, i.e. high-modulation frequencies are the most reduced compared to low-modulation frequencies. [Houtgast and Steeneken \(1985\)](#) expressed the MTF affected by noise (Eq. I.9) and reverberation (Eq. I.10) as a function of the SNR, the reverberation time (T), the audio frequency (f) and the modulation frequency (F) as follows¹:

$$MTF_{noise}(f) = \frac{1}{1 + 10^{\frac{-SNR(f)}{10}}} \quad (I.9)$$

$$MTF_{room}(F, f) = \sqrt{\frac{1}{1 + \left(2\pi F \frac{T(f)}{13.8}\right)^2}} \quad (I.10)$$

Note that only the case of stationary noise is considered by [Houtgast and Steeneken \(1985\)](#) because long-term signals are used whereas the influence of fluctuating noise would require a short-term analysis.

MTFs values are then converted into an apparent SNR, i.e. the reduction of modulation is interpreted as an increase of SNR which would have the same effect on speech intelligibility. This apparent SNR is calculated as a function of the modulation frequency (third-octave intervals from 0.63 Hz to 12.5 Hz) and the audio frequency (third-octave intervals from 125 Hz to 8 kHz):

1. To account for the MTF due to both noise and reverberation, the equations I.9 and I.10 are combined by multiplying them together.

$$SNR_{app}(F, f) = 10 \log \left(\frac{MTF(F, f)}{1 - MTF(F, f)} \right) \quad \text{with} \quad -15 \leq SNR_{app} \leq 15 \quad (\text{I.11})$$

From this apparent SNR, the STI is computed as follows:

$$STI = \frac{1}{30} \left(15 + \frac{1}{14} \sum_{i=1}^{14} \sum_{j=1}^7 w_j \cdot SNR_{app}(F_i, f_j) \right) \quad (\text{I.12})$$

where i and j designate the indexes of the modulation frequency and audio frequency bands, respectively. The values w_j refer to weightings coefficients derived from those used in the SII calculation, accounting for the importance of a given frequency band regarding speech intelligibility.

Like the AI and SII, the STI is ranged between 0 and 1. It is based on a modulation approach, using the MTF as a predictor of speech intelligibility in rooms. It is then able to account for the effects of temporal smearing of target speech by reverberation as well as stationary interfering noises using a limited number of physical parameters: speech and noise spectra and reverberation time. However, as the AI and SII, the STI is a monaural index which cannot account for spatial unmasking, temporal dip listening or F0-segregation.



- Temporal smearing of speech
- Backward compatibility for anechoic environments



- No spatial unmasking
- No temporal dip listening
- No F0-segregation

3.1.5 Copenhagen model

The STI appeared to yield accurate predictions concerning speech intelligibility in rooms with or without the presence of masking noise. However, the STI cannot handle situations where speech is subjected to non-linear processings such as deterministic envelope reduction, envelope compression or spectral subtraction (which could occur in hearing-aids processing for example). To account for both reverberation and spectral subtraction, [Jørgensen and Dau \(2011\)](#) proposed a speech-based envelope power-spectrum model (sEPSM) to predict speech intelligibility by considering signal-to-noise ratios in the modulation domain. It is an extension of the models of [Dau *et al.* \(1999\)](#) and [Ewert and Dau \(2000\)](#), who used an envelope power

spectrum model (EPSM) to predict amplitude modulation detection. The sEPSM takes the noisy speech and noise signals as inputs and predicts percent correctly-recognized items as illustrated in Figure I.9. First, signals are passed through a gammatone filterbank covering the range from 63 Hz to 8 kHz. The envelope is then extracted in each frequency band using the Hilbert transform and passed through a modulation filterbank covering the range from 0 Hz to 64 Hz. The envelope power of both the noisy speech and the noise are calculated at the output of each modulation filter and the envelope power signal-to-noise ratio (SNR_{env}) is determined. An overall SNR_{env} is calculated by integrating all individual SNR_{env} across modulation frequency and audio frequency channels. This overall SNR is then converted into a score performance with statistical and probabilistic methods. SNR_{env} is first converted into a d' value which is used as a parameter of a m -alternatives forced-choice (mAFC) model which determines the probability of an ideal observer for selecting the correct speech item from a set of m alternatives.

Jørgensen and Dau (2011) did not test their model on reverberant speech in the absence of noise but the authors expect that purely reverberant conditions could be accounted for by the sEPSM. Since predictions are based on long-term integrated SNR_{env} , this model cannot take into account the effect due to temporal fluctuations in the masker envelope. Such modulations would increase the power of the noise at the output of a given modulation filter and then reduce SNR_{env} . This model version would predict a decrease of speech intelligibility, which would not be in agreement with the “listening in the dips” ability.

To account for this mechanism, Jørgensen *et al.* (2013) further extended their model by considering, like the model of Rhebergen and Versfeld (2005), SNR_{env} within short-term time-frames instead of using the long-term signals. The two first stages (envelope extraction and modulation filtering) of the model are roughly the same as in Jørgensen and Dau (2011). The key novelty is that the envelope at the output of each modulation filter is segmented into rectangular time-frames without overlap between frames. The length of the time-frame varies across modulation channels and is chosen as the inverse of the center frequency of the considered filter, e.g. 250 ms for the 4-Hz filter and 125 ms for the 8-Hz filter. Because of this time-length dependence, the model was renamed as multi-resolution envelope power spectrum model (mr-sEPSM). Within each time-frame i , $\text{SNR}_{\text{env},i}$ is calculated with the same process as in Jørgensen and Dau (2011). All $\text{SNR}_{\text{env},i}$ are then averaged across time-frames, resulting in a single SNR_{env} in each modulation channel. It differs from the SNR_{env} obtained in the previous version of the model (Jørgensen and Dau, 2011) because the temporal variations of both the noisy target and noise envelopes are taken into account within the time-frames.

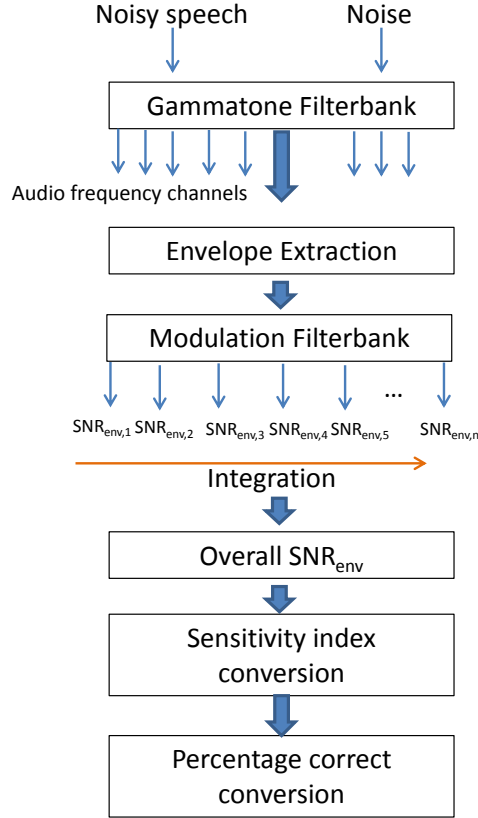


Figure I.9 – Bloc diagram of the Speech-based Envelope Power Spectrum Model (sEPSM) developed by Jørgensen and Dau (2011). Noisy speech and noise signals are passed through a gammatone filterbank and their temporal envelope is extracted in each frequency band. Each envelope is then passed through another filterbank and the SNR is evaluated in the modulation domain before being converted into score performance.

The sEPSM and its extension mr-sEPSM were tested and validated on experimental data involving speech distorted by spectral subtraction and reverberation (Jørgensen and Dau, 2011) and also by fluctuating maskers (Jørgensen *et al.*, 2013). In all conditions with reverberation and spectral subtraction, predictions from both models were very close to the score performance reached by human listeners. With envelope-modulated maskers, mr-sEPSM kept predicting accurately the experimental data, while sEPSM failed by overestimating the measured SRT. By considering long-term envelopes, sEPSM did not take into account the potential masking release due to the variations of SNR_{env} over time. However, the predictive power of the models on these experimental dataset was built on a fitting process of three parameters until the best possible fit between predictions and measurements was achieved. The model then requires a calibration step, related to the speech material used in the experiment aimed to be predicted.

Like sEPSM, mr-EPSM is still a monaural model which only consider the acoustic signal at one ear, which does not allow the prediction of spatial unmasking.



- Temporal dip listening
- Temporal smearing of speech
- Spectral subtraction
- Predictions can be made from noisy speech signal



- Does not account for spatial unmasking
- Does not account for F0-segregation

3.2 Binaural models

3.2.1 Van Wijngaarden and Drullman (2008)

A binaural version of the STI was proposed by [van Wijngaarden and Drullman \(2008\)](#) in order to account for spatial unmasking in addition to the temporal smearing effect already predicted by the STI. Like in the STI procedure, target and masker signals from left and right ears are first filtered into octave bands (centered from 125 Hz to 8 kHz). In the bands centered at 500, 1000 and 2000 Hz, MTF is determined based on interaural cross-correlograms. For the other frequency bands (125 and 250 Hz and 4000 and 8000 Hz), MTF is calculated in each ear channel separately and the highest is simply chosen in each frequency band. These processes result in seven MTF values, which can be combined to calculate an overall STI.

This new STI version was compared to the answers of four listeners who performed consonant-vowel-consonant (CVC) tests over 39 conditions involving spatial unmasking and temporal smearing in four rooms (anechoic, classroom, listening room and cathedral). Spatial unmasking was quite well predicted by the “binaural MTF” approach, while some discrepancies were more noticeable in reverberant environments (especially for the cathedral). Temporal dip listening and F0-segregation cannot be handled by this model.



- Accounts for spatial unmasking
- Accounts for temporal smearing of speech



- No temporal dip listening
- No F0-segregation
- Tested on only one dataset (4 listeners)

3.2.2 Madison model

Jones and Litovsky (2011) proposed a revised version of the model proposed by Bronkhorst (2000) in order to predict spatial release from masking from both speech and noise maskers. From a multiple regression fit on experimental data, Bronkhorst (2000) proposed a mathematical formula to predict spatial unmasking of a frontal speech target disrupted by multiple masking noises located anywhere in the horizontal plane (Eq. I.13). Instead of considering better-ear-listening and binaural unmasking contributions, Bronkhorst (2000) expressed the SRM as a function of the angular separation between target and maskers and of the asymmetry of the maskers positions.

$$SRM = C \left[\underbrace{\alpha \left(1 - \frac{1}{N} \sum_{i=1}^N \cos \theta_i \right)}_{\text{angular separation}} + \underbrace{\beta \frac{1}{N} \left| \sum_{i=1}^N \sin \theta_i \right|}_{\text{asymmetry}} \right] \quad (\text{I.13})$$

where N is the number of masking sources, θ_i is the azimuth of the i^{th} masker, and α , β and C are scaling coefficients.

Jones and Litovsky (2011) conducted speech intelligibility tests where a frontal speech target was presented in competition with two maskers of three types: speech, stationary SSN, and envelope-modulated SSN. A various number of masker positions were tested in order to get both spatial separation and asymmetry involved. Experimental results were compared to the predictions of the Bronkhorst's model to determine the modifications needed to reach better prediction performance. Because of the accurate match of the asymmetry component of the model with the data, the revised version kept this component as formulated by Bronkhorst (2000). They, however, revised the spatial separation component to get a better fit to their experimental data, including a distinct front/back component. The final revised model is expressed as follows:

$$SRM = D \left[\underbrace{\alpha \frac{1}{N} \sum_{i=1}^N \tanh(3\theta_i^*)}_{\text{angular separation}} + \underbrace{\beta \frac{1}{N} \left| \sum_{i=1}^N \sin \theta_i \right|}_{\text{asymmetry}} + \underbrace{\gamma \frac{1}{N} \sum_{i=1}^N \frac{1}{1 + e^{-0.5(|\theta_i| - 110)}}}_{\text{front/back}} \right] \quad (\text{I.14})$$

with

$$\theta_i^* = \begin{cases} |\theta_i| & \text{if } -90 \leq \theta_i \leq 90 \\ |\theta_i - 180| & \text{if } 90 \leq \theta_i \leq 180 \\ |\theta_i + 180| & \text{if } -180 \leq \theta_i \leq -90 \end{cases} \quad (\text{I.15})$$

By adjusting fitting coefficients, this revised model proposed by [Jones and Litovsky \(2011\)](#) is then able to accurately predict spatial unmasking of an anechoic frontal target from multiple noises or concurrent speech sources. An important advantage of the model is that it only requires the masker positions as inputs. But despite the original approach and the simplicity of this model, no other unmasking effect can be predicted since only the source positions are required as input and the acoustical properties of the received signals are totally disregarded, which prevent from any prediction of temporal dip listening, F0-segregation or temporal smearing of speech. The authors also suggest that the influence of the room can be taken into account with a global scaling factor (D). This implementation is arguable since it seems hardly conceivable to model the effect of temporal smearing, coloration and the influence of reverberation on the different unmasking mechanisms with a single scaling factor.



- Spatial unmasking for noise and speech maskers (spatial release from informational masking)
- Only requires the masker positions



- No temporal dip listening
- No F0-segregation
- No temporal smearing of speech
- Target must be located at 0°
- No influence of the room (effect of reverberation reduced to a scaling factor)
- Scaling factors (α , β , γ) must be fitted differently whether the masker is noise or speech so that the difference between speech and noise masker is not predicted by the model.

3.2.3 Oldenburg model

Beutelmann and Brand (2006) developed a binaural model to account for spatial unmasking in rooms. They combined an implementation of the EC theory with the SII. A bloc diagram of their model is represented in Figure I.10. Speech and noise signals from each ear are required as input of the model. In each ear channel, signals are split into 30 frequency bands using a gammatone filterbank. An EC stage is then directly implemented by attenuating and delaying the signals in each frequency band (equalization). The right channel is then subtracted from the left (cancellation). Gain and delay parameters are obtained using an optimization process until the SNR is maximal after cancellation in the frequency band. As originally suggested by Durlach (1963), artificial variance was added to gain and delay parameters in order to model human inaccuracy using a Monte Carlo method. Without it, the EC stage could result in a perfect cancellation of noise (see Fig. 2 in Beutelmann and Brand, 2006). In parallel to the EC stage, monaural SNRs are determined in each frequency band at each ear. SNRs from each ear and from the EC stage are compared and the resulting signals yielding the best SNR are then resynthesized through a gammatone filterbank and used as input for the calculation of the SII. The SII value is then converted into SRT using an algorithm which iteratively fits a psychometric function linking the intelligibility score to the SII.

Beutelmann and Brand (2006) compared model predictions to measured SRTs for nine masker azimuths (with the speech target in front) in three listening environments (anechoic, cafeteria, office). The model accurately predicted the spatial release from masking in reverberation illustrated by good correlations between experimental data and model predictions.

Beutelmann *et al.* (2010) proposed an extension of the model of Beutelmann and Brand (2006), aiming at a simpler implementation, an increased computational efficiency and a prediction of the temporal dip listening mechanism. They adapted the frequency weightings of the SII procedure to allow the SII to be computed directly from the gammatone frequency bands and determined the processing errors with analytical methods instead of using a Monte Carlo procedure. The main modification in this new version is the extension allowing to consider amplitude-modulated noise maskers. To account for the masking release due to temporal fluctuations in the masker envelope, they used the method originally proposed by Rhebergen and Versfeld (2005) and applied their binaural model to short-time frames of the input signals instead of considering the entire signals. The performance of this new version was compared to measured SRTs in the presence of modulated noises in four listening environments (anechoic, listening room, classroom, church). Good correlations were obtained between predicted

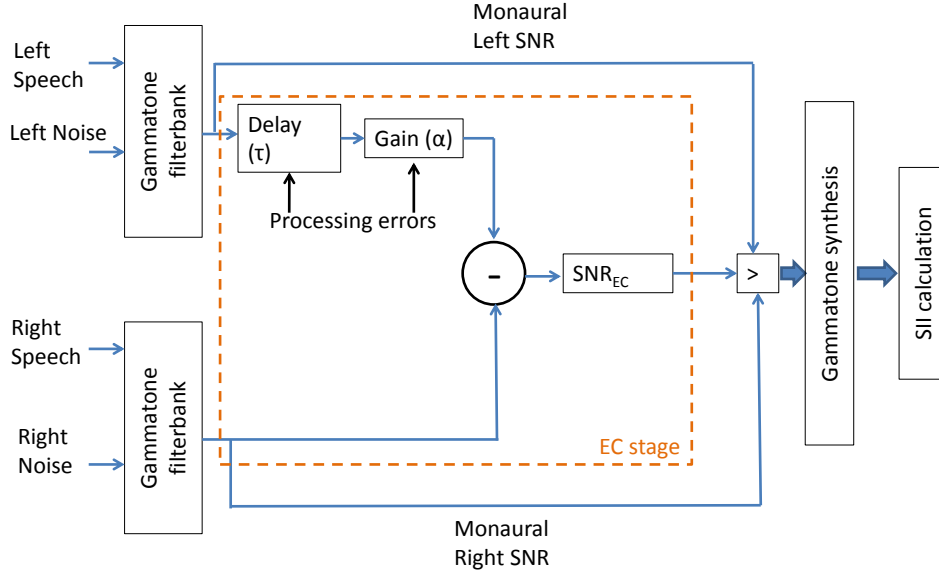


Figure I.10 – Schematic representation of the model developed by [Beutelmann and Brand \(2006\)](#). It combines an EC stage by frequency-bands to account for binaural unmasking, a comparison of monaural SNRs to account for better-ear listening and a SII computation. Optimal gains and delays are obtained iteratively by maximizing the SNR after cancellation within each frequency band. This binaural SNR is compared to monaural SNR and the highest is retained, making the assumption that binaural processing can only improve monaural SNR.

and measured SRTs while maintaining prediction performance equivalent to the original model for stationary noise. The model of [Beutelmann and Brand \(2006\)](#) and its revised version by [Beutelmann et al. \(2010\)](#) only holds for a near-field target and cannot predict the temporal smearing of speech in reverberant environments.

To account for the temporal smearing of target speech, [Rennies et al. \(2011\)](#) extended the models of [Beutelmann and Brand \(2006\)](#)² by testing three different approaches: modulation transfer function (MTF), definition factor (D_{te}) and useful to detrimental ratio (U/D). The definition factor is an architectural acoustic indicator which quantifies the proportion of early-reflection energy within the overall energy of an impulse response ([ISO 3382, 1997](#)). With the instantaneous acoustic pressure noted $p(t)$ and te being the temporal limit until the reflections

2. but the computational improvements brought by [Beutelmann et al. \(2010\)](#) were kept by [Rennies et al. \(2011\)](#).

are considered as “early”, it is defined as:



$$D_{te} = \frac{\int_0^{te} p^2(t)dt}{\int_0^{\infty} p^2(t)dt} \quad (\text{I.16})$$

For the MTF and the D_{te} versions of the model, the main structure of the model of [Beutelmann and Brand \(2006\)](#) is preserved, and the temporal smearing of speech is computed in a parallel path: a correction factor computed from MTF or D_{te} measurements calculated from the target BRIR is applied to the SNRs obtained after the EC stage. The contribution of temporal smearing and the spatial release from masking are then implemented independently. Conversely, in the third alternative using the U/D approach, the temporal smearing of speech is implemented at the very beginning of the model by splitting the target BRIR into early and late parts according to a temporal early/late limit (ELL). Each part is then convolved with the speech signal to create “early speech” and “late speech”. The same model as [Beutelmann and Brand \(2006\)](#) is then used but only the “early speech” is considered as the target while the “late speech” signal is combined with the noise according to the U/D approach which attributes different roles to early and late reflections regarding speech intelligibility ([Lochner and Burger, 1964](#)). [Rennies et al. \(2011\)](#) tested these three approaches by comparing the model predictions to experimental SRTs, obtained for reverberant and spatially-separated target/masker configurations. With the D_{te} and U/D approaches, three ELLs were tested: 50, 80 and 100 ms. These two approaches with $ELL = 100$ ms appeared to yield the most accurate predictions compared to the approach based on MTF.

[Rennies et al. \(2014\)](#) further investigated the results obtained by [Rennies et al. \(2011\)](#) to determine the best approach to account for the temporal smearing of speech. They used the models presented in [Rennies et al. \(2011\)](#) to predict the data from [Warzybok et al. \(2013\)](#), who examined the spatial and temporal influences of a single reflection on a speech target. The conditions tested by [Warzybok et al. \(2013\)](#) were chosen to determine 1) which approach between D_{te} and U/D yields the best predictions for reverberant speech intelligibility and 2) how should the impulse response be splitted into early and late reflections. They concluded that the U/D approach gave to most accurate prediction on this dataset.

In the different versions proposed by the Oldenburg teams, the predicted effects were handled individually (or in pairs) by separate models: spatial unmasking in [Beutelmann and Brand \(2006\)](#), spatial unmasking and temporal dip listening in [Beutelmann et al. \(2010\)](#) and spatial

unmasking and temporal smearing in [Rennies *et al.* \(2011, 2014\)](#). No unique model was yet developed to combine all these effects.

<div style="text-align: center;">  </div> <ul style="list-style-type: none"> – Spatial unmasking in rooms – Temporal dip listening – Temporal smearing of speech 	<div style="text-align: center;">  </div> <ul style="list-style-type: none"> – No F0-segregation – Complexity of the computations – No unified model
--	--

3.2.4 Boston model

A binaural EC-based model with time-varying jitters was developed by [Wan *et al.* \(2010\)](#). A schematic representation of the model structure is presented in Figure I.11. It is based on the EC theory ([Durlach, 1963](#)) by taking, in each frequency band, the best SNR between the SNR at each ear and the one resulting from an EC process, which is directly implemented by looking for the optimal interaural amplitude and time equalization parameters which minimize the residual energy of the masker after cancellation. The retained SNRs are then weighted across frequency using the SII weightings, and converted into SRTs using the comparison to a reference curve deduced from an SII calculation for a colocated condition. The new aspect of this model is the implementation of time-varying jitters at the input of the EC stage. Instead of using fixed values to account for equalization errors due to human inaccuracy (as it was the case in [Beutelmann *et al.*, 2010](#); [Lavandier and Culling, 2010](#)), amplitude and time jitters are implemented as a function of time and may vary across frequencies. According to the notation used in Fig. I.11, the jittered waveforms at the output of a given bandpass filter (denoted by the index i) are given by:

$$\begin{aligned} L_i(t) &= (1 + \varepsilon_{Li}(t)) \cdot X_{Li}(t - \delta_{Li}(t)) \\ R_i(t) &= (1 + \varepsilon_{Ri}(t)) \cdot X_{Ri}(t - \delta_{Ri}(t)) \end{aligned} \tag{I.17}$$

where ε and δ represent the amplitude and time jitters, respectively, which are characterized as zero-mean Gaussian random variables with standard deviations independent of frequency.

[Wan *et al.* \(2010\)](#) compared their model predictions with experimental measurements already reported in the literature ([Hawley *et al.*, 2004](#); [Marrone *et al.*, 2008](#)). The model yielded a good performance in predicting the intelligibility of a speech target masked by multiple spatially-separated SSNs and envelope-modulated SSNs. It is important to note that predic-

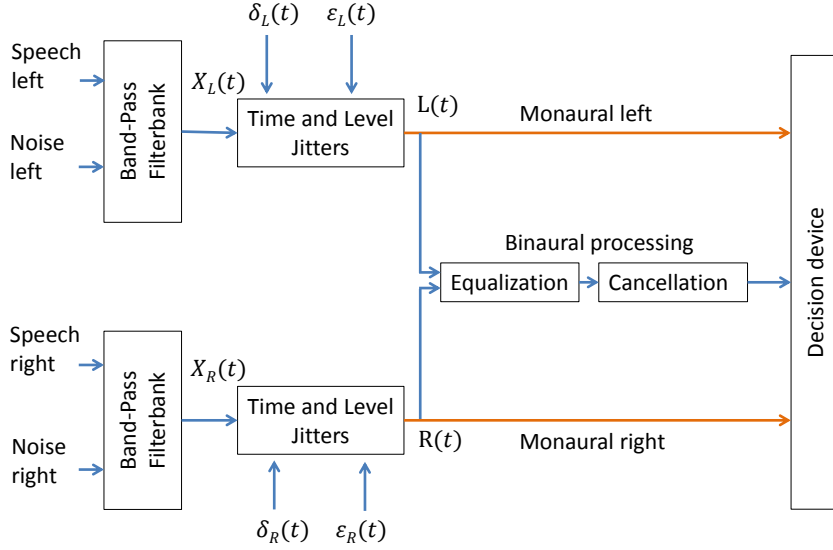
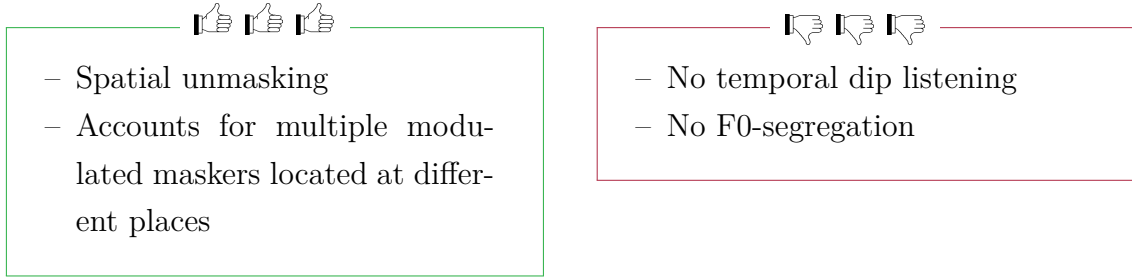


Figure I.11 – Schematic structure of the model of [Wan et al. \(2010\)](#). Speech and noise signals from each ear are filtered into a third-octave filterbank with the center frequencies ranging from 160 Hz to 8kHz. Filtered signals are then jittered in amplitude and time independently for each ear and frequency band. For each frequency band, the SNR on each ear is computed as well as the SNR resulting from an EC process of the jittered signals. A decision device takes the maximum SNR among the three calculated SNRs and converts it into a SRT.

tions have been fitted to the experimental data independently for each masker type. The model needs a calibration step to reach good performance level. Some discrepancies between predicted and measured SRTs remained for speech and reverse-speech maskers.

This model was further developed by [Wan et al. \(2014\)](#) to improve predictions of spatial unmasking when multiple maskers are modulated, like speech or reverse speech. In the presence of multiple modulated maskers located at different places, speech intelligibility could be influenced by short-term variations in the direction of the most energetic masker. In this new model version, a short-time approach is implemented which is why it is called “short-time equalization cancellation” (STEC) model in opposition to the previous “steady-state” version (SSEC). In addition to be filtered, input signals are segmented into short-time frames within each frequency band and the EC process is applied within each time-frequency units. In each frequency band, the signals resulting from the equalization-cancellation in each short-time frame are summed

across frames and the SNR is computed from the resulting waveform. The temporal variations in SNR are then disregarded and the “listening in the dips” mechanism is not considered in this framework. The decision device remains the same as in the SSEC model, comparing the SNRs provided by the two monaural pathways and by the short-time EC process, and selecting the best ratio.



3.2.5 Cardiff/Lyon model

Figure I.12 schematically represents the binaural model proposed by Lavandier and Culling (2010) which aims to predict spatial release from masking in rooms. It takes the target and masker BRIRs convolved with noise as inputs, and computes an effective broadband target-to-masker ratio (TMR) which can be compared to differences in SRT by inverting the TMR. It cannot predict absolute intelligibility but differences between conditions. All input signals are first passed through a gammatone filterbank. In each frequency band, two components are computed to account for spatial unmasking: better-ear listening and binaural unmasking. Better-ear listening is computed by comparing the SNRs at each ear and retaining the highest. SNRs are calculated from the excitation patterns of both target and masker (Moore and Glasberg, 1983; Glasberg and Moore, 1990). Instead of directly implementing the EC process and search for the gains and delays which maximize SNRs, binaural unmasking is estimated from the computation of a binaural masking level difference (BMLD) using an analytical formula proposed by Culling *et al.* (2004, 2005) derived from the EC theory (Durlach, 1963):

$$BMLD = 10. \log \left(\frac{k - \cos(\phi_{target} - \phi_{masker})}{k - \rho_{masker}} \right) \quad (I.18)$$

with

$$k = (1 + \sigma_\epsilon^2).e^{\omega^2 \sigma_\delta^2} \quad (I.19)$$

where $\sigma_\varepsilon = 0.25$ and $\sigma_\delta = 105 \mu\text{s}$ are the standard deviations of the zero-mean amplitude and time jitters, respectively, proposed by [Durlach \(1963\)](#) to account for human inaccuracy. All the other parameters needed to compute the BMLD (ρ_{masker} , ϕ_{target} and ϕ_{masker}) are obtained by cross-correlating the signals of each source between the ears. The interaural correlation (ρ) corresponds to the maximum value of the cross-correlation function and the delay of this maximum gives the interaural phase difference for a specific frequency band. Better-ear SNRs and BMLDs are then weighted across frequency using the SII-weighting coefficients ([ANSI S3.5, 1997](#)), and finally added together to yield the effective TMR representing the binaural advantage in a given condition.

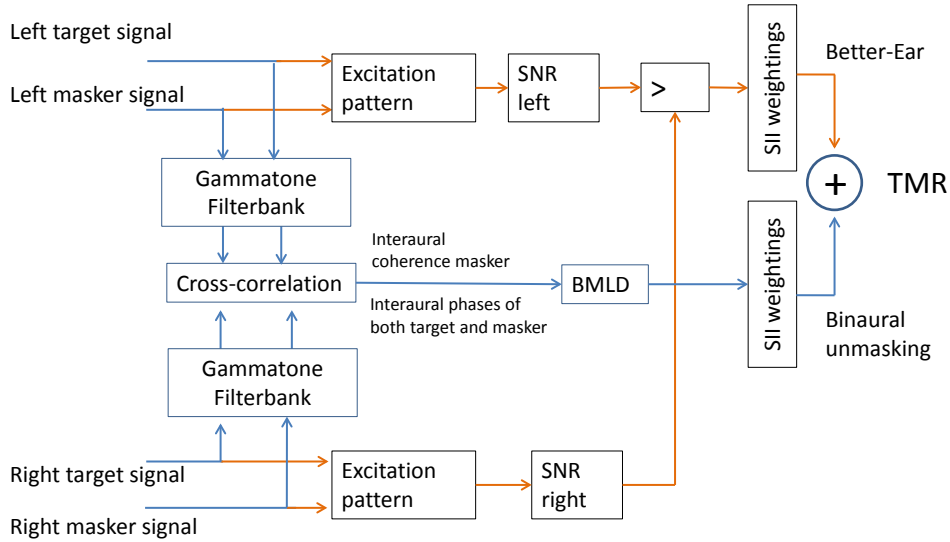


Figure I.12 – Structure of the model of [Lavandier and Culling \(2010\)](#). It combines in each frequency band, a better-ear SNR (orange path) based on excitation patterns with binaural unmasking (blue path) based on interaural coherence of the masker and interaural phase differences of both target and masker obtained by cross-correlating the signals of each source between the ears. Better-ear SNR and BMLD are then weighted across frequency using SII weightings, and summed together to yield one single effective target-to-masker ratio (TMR).

This model correctly predicts the masking release due to spatial separation between target and masking noises by showing a 0.95 correlation between predicted and measured SRTs ([Lavandier and Culling, 2010](#)). It also takes into account the disruption of spatial unmasking by reverberation which is mainly due to the decrease of the masker coherence ([Lavandier and](#)

Culling, 2008). The influence of the room on the sources' spectra (coloration) is directly implemented in the computation of the better-ear listening component. However, like the model of Beutelmann and Brand (2006) and Beutelmann *et al.* (2010), it is limited to near-field targets and cannot account for the temporal smearing of speech since the model considers the whole target signal as useful and does not involve any target envelope-modulation approach. It is also limited to stationary noises and cannot predict any masking release due to temporal gaps in the masker envelope.

Two revisions of the model of Lavandier and Culling (2010) were achieved by Jelfs *et al.* (2011). First, the inputs signals of the model have been replaced by the BRIRs of each source (without requiring the convolution by noises). It yields an improved computational efficiency by saving a convolution operation, but it also produces non-stochastic results which do not need to be averaged over several noise samples anymore. Second, the separate excitation patterns needed to compute the better-ear TMR have been replaced by energy ratios computed on the gammatone-filtered BRIRs for target and masker. The BMLD computation remains unchanged since all the binaural cues needed for binaural unmasking (interaural phase and coherence) are contained in the BRIR.

This new version has been tested on several experimental data from the literature including spatial unmasking with binaural unmasking and better-ear listening both in isolation and in combination, with a single or up to three stationary noise maskers in anechoic conditions. It has been validated by Lavandier *et al.* (2012) who further evaluated the model for more realistic situations involving head-shadow, multiple stationary maskers in reverberation from real rooms.

Collin and Lavandier (2013) proposed to extend the model of Lavandier and Culling (2010) to be able to consider envelope-modulated maskers. They presented a “proof of concept” with the will to keep the same structure as Lavandier and Culling (2010). Inspired by Rhebergen and Versfeld (2005) and Beutelmann *et al.* (2010), the stationary model is successively applied to short-time frames along the target and masking temporal signals to account for the variations of SNR across time. The resulting predictions are then averaged across frames. Some modifications on the model were needed to ensure backward compatibility. In the presence of modulated noise, the energy of the noise is expected to drop down near zero within a time-frame, which may affect the computation of both better-ear listening and binaural unmasking. Better-ear listening can yield huge values of SNRs in the absence of noise, and BMLD is impossible to compute without the interaural coherence and phase of the masker. More conceptually, spatial unmasking does not make any sense in quiet conditions. To avoid these computation artefacts, Collin and Lavandier (2013) implemented a ceiling value for the better-ear SNR and BMLD

was set to zero for such frames presenting very low noise energy. Measured SRTs from three experiments were best predicted using a 10-dB ceiling value (correlation between 0.84 and 0.9, and mean absolute error below 0.8 dB), validating this time-frame approach. This version still needs to be optimized regarding the precise value to be chosen as ceiling parameter.

None of these versions is able to predict the temporal smearing of target speech occurring in reverberant environments because the entire target signal is considered by the model, without any way to detect the detrimental effect of late reflections. No harmonicity or periodicity analysis is implemented in the model, which prevent any prediction of F0-segregation or speech intelligibility in the presence of harmonic maskers.



- Spatial unmasking
- Temporal dip listening
- Handles multiple maskers
- Only requires BRIR of each source



- No temporal smearing of speech
- No F0-segregation

3.3 Summary

Many years of research have conducted to identify several unmasking mechanisms on which the auditory system can rely on. These mechanisms have been described above (spatial unmasking, temporal dip listening and F0-segregation) and rely on the acoustic properties of both the target and masking sources. Reverberation can alter these acoustic cues due to the multiple reflections arriving with a temporal delay and a modified spectrum, leading to some interactions between the perceptual mechanisms and the room. In addition, reverberation intrinsically influences speech intelligibility in both the temporal and frequency domains, even in the absence of masking sources. Some of these features regarding speech intelligibility in noise have been partially modelled in many studies with different approaches. Tables [I.1](#) and [I.2](#) present a comparative summary of all the models described in this chapter.

4 Aims of the PhD

In this PhD work, the chosen approach to deal with the cocktail-party problem is to focus on energetic masking occurring between target and masking sources at a peripheral level. This PhD aims to extend the model of Lavandier and Culling towards a binaural model for predicting speech intelligibility among competing voices in rooms. The influence of the room on the speech target was first investigated and implemented into the model (chapter II). In a second study (chapter III), the case of sources with different spectra is considered and, based on experimental results, the model parameters have been modified. Chapter IV presents an experimental work examining the potential interactions between some of the unmasking mechanisms described above. This work is still at an experimental stage and modelling the data is a potential perspective of this PhD work. General conclusions are summarized in a final chapter and some perspectives for this research work are suggested.

Models		<i>Spatial Unmasking</i>	<i>Temp. Dip Listening</i>	<i>F0 segregation</i>	<i>Temp. smearing of speech</i>
Monaural	AI	-	-	-	-
	SII	-	-	-	-
	STI	-	-	-	✓
	R&V	-	✓	-	-
	Copenhagen	-	✓	-	✓
Binaural	vW&D	✓	-	-	✓
	Madison	✓	-	-	-
	Oldenburg	✓	✓	-	✓
	Boston	✓	-	-	-
	Cardiff/Lyon	✓	✓	-	-

Table I.1 – Comparative table of different unmasking effects handled by each model approach. These models were tested and validated on different experimental data. The prediction performances are not discussed here since only a few of them tested the same effects on the same data.

Models		Inputs	Approach
Monaural	AI	Target and masker signals	Sum of weighted SNRs across frequency bands
	SII	Target and masker signals	Sum of weighted SNRs across frequency bands
	STI	Target and masker signals	Modulation Transfer Function (MTF)
	R&V	Target and masker signals	SII within short-time frames
	Copenhagen	2011: Noisy speech and noise signals	SNR based on the power of the envelope
		2013: Noisy speech and noise signals	SNR based on the power of the envelope within short-time frames of variable length
Binaural	vW&D	Target and masker signals	MTF from interaural correlograms
	Madison	Masker azimuths	Mathematical determination of spatial unmasking depending on angular separation, asymmetry and front/back configuration.
	Oldenburg	2010: Target and masker signals	EC theory applied on short-time frames
		2014: Target and masker BRIRs	U/D with EC theory
	Boston	Target and masker signals	EC theory with time-varying jitters applied to short-time frames
	Cardiff/Lyon	2010: Target and masker signals	EC theory implemented analytically
		2011-2012: Target and masker BRIRs	EC theory implemented analytically
		2013: Target and masker signals	EC theory implemented analytically and applied to short-time frames

Table I.2 – Comparative table of the different models presented above highlighting the different inputs needed and the approach used for the predictions.

Chapter II

Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

This chapter deals with the extension of the model of [Lavandier and Culling \(2010\)](#) in its revised version by [Jelfs *et al.* \(2011\)](#), by taking into account the temporal smearing of speech with a U/D approach. It has been published in the Journal of Acoustical Society of America, Leclère, T. , Lavandier, M. and Culling, J. F. [(2015). J. Acoust. Soc. Am. **137**, 3335-3345].

1 Introduction

Speech intelligibility is impaired in noisy rooms by both noise and reverberation. The speech signal is mixed with delayed versions of itself reflected by room boundaries: the speech can be smeared and self-masked ([Bradley, 1986](#); [Houtgast and Steeneken, 1985](#)). In the presence of discrete noise sources, a listener is able to partly separate target speech from masking noise using the binaural system. This ability is impaired by reverberation ([Beutelmann and Brand, 2006](#); [Culling *et al.*, 2003](#); [Plomp, 1976](#)). The corresponding loss of intelligibility appears at lower levels of reverberation, and thus occurs more readily, than the loss of intelligibility associated with the smearing of speech ([Lavandier and Culling, 2008](#)). The aim of the present study was to propose and validate a model predicting these multiple effects.

Architectural acoustic indicators of intelligibility have focused on the effects of temporal smearing of speech and masking by diffuse ambient noise. The speech transmission index (STI) measures the reduction of amplitude modulation in the speech signal due to reverberation and noise ([Houtgast and Steeneken, 1985](#)). The useful-to-detrimental (U/D) ratio computes a signal-to-noise ratio (SNR), in which the early reflections of the target are regarded as useful and as the “signal” because they reinforce the direct sound ([Bradley *et al.*, 2003](#)), while the late reflections are regarded as detrimental and effectively a part of the noise ([Bradley *et al.*, 1999](#); [Bradley, 1986](#); [Lochner and Burger, 1964](#)). These monaural indicators neglect the listener’s ability to separate target speech from interfering sounds using the binaural system, as well as the susceptibility of this ability to reverberation.

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

In the presence of discrete noise sources, masking is less efficient when the target and noise sources are on different bearings (Hawley *et al.*, 2004; Plomp, 1976). This spatial release from masking is based on two mechanisms (Bronkhorst and Plomp, 1988): better-ear listening and binaural unmasking, which rely on interaural level and time differences (ILDs and ITDs) respectively. Target and interferers at different locations often produce different ILDs so that one ear usually offers a better SNR than the other, and listeners can attend to the ear offering the better ratio. Differences in the ITD generated by target and interferer facilitate binaural unmasking, in which the auditory system is able to “cancel” to some extent the noise generated by the interferer (equalization-cancellation (EC) theory; Durlach, 1972), thus improving the internal SNR. Both processes are affected by reverberation. Sound reflections traveling around the listener reduce the acoustic shadowing by the head (Plomp, 1976) and impair binaural unmasking mainly by decorrelating the interfering noise at the two ears (Lavandier and Culling, 2008).

Beutelmann and Brand (2006) implemented this binaural ability into a model of speech intelligibility. Simulated stimuli at the ears are processed through a gammatone filterbank, an EC stage, then re-synthesized, and the speech intelligibility index (SII) method is used to evaluate intelligibility (ANSI S3.5, 1997). For each frequency band of the gammatone filterbank, the EC stage directly implements a mechanism based on EC theory, testing different delays and attenuations for the signals at the ears and choosing those maximizing the SNR. Lavandier and Culling (2010) developed a prediction model also based on EC theory but the better-ear listening and binaural unmasking are computed separately. The direct implementation of cancellation is replaced by a predictive equation, extending the models of Levitt and Rabiner (1967) and Zurek (1993). Binaural unmasking prediction and better-ear target-to-interferer ratio are added and weighted across frequency with the SII-importance band coefficients. Like in the model of Beutelmann and Brand (2006), the prediction method is based on the signals in the room, requiring averaging across signals to produce reliable predictions. Beutelmann *et al.* (2010) revised their original model by improving the computational EC stage with an analytical expression instead of using probabilistic methods. The model of Lavandier and Culling (2010) was also revised by directly applying the model to binaural room impulse responses (BRIRs) instead of signals, thus producing non-stochastic predictions (Lavandier *et al.*, 2012; Jelfs *et al.*, 2011). Like the model of Beutelmann and Brand (2006), the model of Wan *et al.* (2010) uses a direct implementation of an EC process, but with time-varying jitters in time and amplitude and monaural pathways in addition to the binaural pathway. All these binaural models neglect the temporal smearing of speech by reverberation, so their predictions only hold for near-field targets with a high direct-to-reverberant (D/R) ratio.

Van Wijngaarden and Drullman (2008) introduced a binaural version of the STI. This approach makes the assumption that the target is the only source of modulation at the listener’s ears, so that it does not offer any opportunity for extension to modulated noise (Collin and Lavandier, 2013; Beutelmann *et al.*, 2010) or speech interferers. In these cases, the modulation is coming from both target and interferer. Rennies *et al.* (2011) extended the model of Beutelmann *et al.* (2010) to take the smearing effect of reverberation into account using three alternatives: the modulation transfer function (MTF), the definition factor (D_{te} , ISO 3382, 1997) and the U/D ratio. In the first two approaches, spatial unmasking and temporal smearing are processed separately: the SNRs obtained with their binaural model applied to the entire speech and noise signals are corrected a posteriori by either measuring the MTF or D_{te} of the target room impulse response. In the third approach, this impulse response is split into early and late parts which are convolved with the speech signal to create an “early speech” signal and a “late speech” signal. The prediction process is then similar to that of Beutelmann *et al.* (2010) except that the original target signal is replaced by the early speech and the late speech is added to the interferer, so that the detrimental influence of late reflections is taken into account before the binaural process. Rennies *et al.* (2014) tested these three approaches on the data of Warzybok *et al.* (2013) which involved a frontal target smeared by a single reflection. They introduced a weighting function to separate early and late reflections within the impulse response (with the D_{te} and U/D extensions). These modelings allowed them to retain the U/D approach as the most suitable to account for the temporal smearing of speech.

The present study aimed to test the U/D approach to extend the validity of a different binaural model framework (Lavandier and Culling, 2010). In the literature, U/D models are based on a wide range of values/methods to separate early and late reflections (Rennies *et al.*, 2014, 2011; Bradley *et al.*, 2003; Soulodre *et al.*, 1989; Bradley, 1986; Lochner and Burger, 1964). So, this study further investigated the influence of the early/late separation (see sect. 2.2), using realistic reverberation from different rooms.

None of the binaural models presented above have ever been shown to predict the “squelching” effect of binaural hearing. In the literature, the term “binaural squelch” has been used to describe the general advantage of binaural hearing over monaural hearing (Koenig, 1950) or the binaural advantage when better-ear listening has been taken into account (Bronkhorst and Plomp, 1988). However, this last advantage is also sometimes referred to as “binaural unmasking” or “binaural interaction”. To avoid any ambiguity, the term “binaural de-reverberation” will be preferred to “binaural squelch” here. It will refer hereafter to the benefit from binaural listening compared to diotic/monaural listening in reverberation even in the absence of an interfering source. This benefit has been shown to slightly improve intelligibility for reverberant

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

speech in quiet (Nábělek and Robinson, 1982; Moncur and Dirks, 1967). Such a small but significant binaural advantage was also measured by Lavandier and Culling (2008) in the presence of a noise interferer. Binaural speech led to lower thresholds than diotic speech. Because binaural unmasking from the noise was probably not affected by the target listening mode in this configuration, the authors concluded that the result could be explained by the binaural de-reverberation observed in quiet.

An integrated model is proposed here to account for speech transmission (and temporal smearing), spatial unmasking from noise interferers and binaural de-reverberation as defined above. The predictions were compared with Speech Reception Thresholds (SRTs, level of the target compared to that of the interferer for 50% intelligibility) measured in three experiments from the literature (Rennies *et al.*, 2011; Lavandier and Culling, 2008), in which spatial unmasking and target smearing were both simultaneously involved. Two versions of the model were tested: a room-dependent (RD) model whose parameters were adjusted in each room, and a room-independent (RI) model with fixed parameters across rooms. The RI model was tested on a fourth dataset which involved several rooms not used to define its parameters (van Wijngaarden and Drullman, 2008).

2 The integrated model

2.1 Model structure

Since the U/D approach requires the target BRIR as input, the present study extends the model of Lavandier and Culling (2010) in its implementation based on the BRIRs measured between the sources and listener positions (referred to as “old model” in this paper; Jelfs *et al.*, 2011; Lavandier *et al.*, 2012) rather than the last version proposed by Collin and Lavandier (2013) which is not applied to BRIRs but to the signals within short-time frames. The target BRIR is first separated into an early and a late part (see section 2.2 for details). The early part constitutes the useful component. The late part is combined with the BRIRs of the interferers to form the detrimental component. These BRIRs are concatenated rather than added to preserve phase information and avoid constructive/destructive interference (Jelfs *et al.*, 2011). The binaural model is then applied to the useful and detrimental components in the same way as it was previously applied to the target and interferer BRIRs. The detailed implementation of the old model is not described here, but it can be summarized by three steps: (1) gammatone filtering, (2) computation of the better-ear listening and binaural unmasking, (3) SII weightings (ANSI S3.5, 1997). Better-ear listening is estimated from the U/D energy ratios computed as

a function of frequency at each ear, selecting the ear for which the ratio is higher. Binaural unmasking is estimated from the binaural masking level difference (BMLD) computed using the interaural phase differences of the useful and detrimental parts and the interaural coherence of the detrimental part (Lavandier *et al.*, 2012, Eq. 1,2). The resulting better-ear U/D ratios and BMLDs (in dB) are SII weighted, integrated across frequency and summed to provide a broadband binaural U/D ratio.

To be compared with SRTs, which are by definition signal-to-noise ratios, binaural U/D ratios are inverted, so that high ratios correspond to low thresholds. Differences in inverted ratios can be directly compared to SRT differences, or a reference is chosen for the comparison. The reference here was the averaged SRT across conditions for each experiment.

2.2 Early/Late separation parameters

Useful and detrimental signals are obtained by splitting the target BRIR into early and late parts. This separation uses two temporal weighting windows: the early and the late windows which isolate the early and late parts, respectively, by multiplying the original impulse response by the window in the time domain. Here, early and late windows are always defined to be complementary, such that their sum is always 1 (Fig. II.1).

Before the early/late separation, the direct sound was defined as the earliest sound at the ears. A recursive algorithm was applied to each BRIR channel (left and right) to locate the direct sound, and then, the earliest of the two was taken as the unique direct arrival time of the BRIR. The algorithm found the first sample which is at least 25% greater than all previous samples in the BRIR channel. This algorithm was used because taking the maximum value or the first non-zero sample in the BRIR could induce biases in the direct sound arrival time (if a combination of reflections is stronger than the direct sound or if some ambient noise is recorded before the impulse).

The rectangular window is the most usual way to split an impulse response into early and late parts. The early part is defined as the original impulse response until a temporal limit, beyond this limit, the samples of the window are set to zero. This early/late limit (ELL) is relative to the direct sound and is the only parameter required for the rectangular window. Despite the simplicity of this window, the frontier between useful and detrimental is very sharp and thus two reflections can be considered very differently even if they are separated with only few samples. Warzybok *et al.* (2013) highlighted this limitation in the presence of a single reflection. The work of Lochner and Burger (1964) showed that only a part of the energy of early reflections can be considered as “useful” regarding speech intelligibility. Rennies *et al.* (2014)

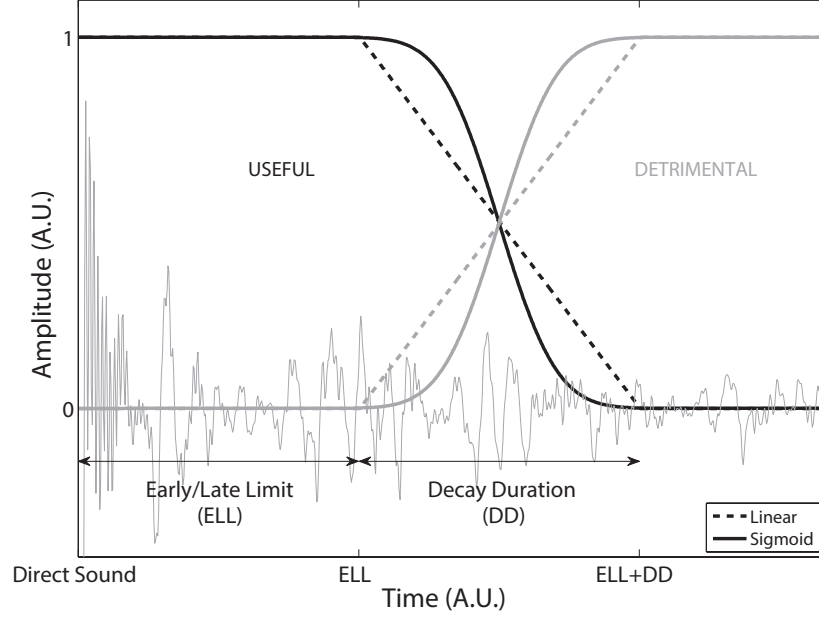


Figure II.1 – Illustration of the temporal weighting windows tested in the present study. Black curves represent the early windows whereas the grey curves represent the late windows. Samples in the impulse response are either considered as fully useful (before the early/late limit [ELL]), fully detrimental (beyond ELL + decay duration [DD]) or partially useful (during DD). The rectangular window is a linear window with a null DD and ELL as a unique parameter.

Leclère *et al.* [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

also tested a linearly decaying window to separate early and late reflections. Two window shapes with a progressive weighting of reflections across time were thus tested: the “linear” window and the “sigmoid”¹ window (see Fig. II.1), which both have a decay duration (DD) parameter in addition to ELL. These temporal parameters are here defined differently than in Rennies *et al.* (2014). ELL defines the duration of the flat part of the window, whereas DD is the duration of the decrease starting from one at ELL and ending at zero at ELL+DD (Fig. II.1). With these definitions, a rectangular window is a linear window with DD = 0 ms.

Three parameters were thus tested concerning the separation of early and late parts of the target BRIR: ELL, DD and window shape.

1. The sigmoid window was defined as $\Phi(t) = 0.5 \times (1 + \operatorname{erf}(\frac{t-\mu}{\sigma\sqrt{2}}))$ with σ and μ defined such that $\Phi(ELL) = 0.999$ and $\Phi(ELL + DD) = 0.001$.

3 Validation of the room-dependent model and definition of the room-independent model

3.1 Data from the literature

The model predictions were compared to SRTs measured using headphones in three experiments (Rennies *et al.*, 2011; Lavandier and Culling, 2008), with one target source (connected speech) in competition with one interferer source (speech-spectrum noise). The three modeled experiments are briefly presented to describe the effects which need to be predicted by the proposed model: spatial unmasking, temporal smearing and binaural de-reverberation. More details are available in the original publications.

3.1.1 Temporal smearing and spatial unmasking

In their experiment 1 (referred to as RBK in the following), Rennies *et al.* (2011) measured SRTs across 12 conditions in a virtual room. The reverberation level was varied by moving the listener away from the fixed frontal target (0.5 m, 1.5 m, 3.5 m and 13 m). For each distance, the single interferer source was placed either frontally, at 22.5° or at 90° to the right of the listener. The distances between listener and each source were generally the same for all listener positions. Since both the azimuth of the interferer source and the reverberation level on the target varied across conditions, both spatial unmasking and temporal smearing were observed in the results.

In the experiment 3 (referred to as LC3 in the following) of Lavandier and Culling (2008), the listener was facing a target and an interferer source spatially separated at fixed positions (65° to the left and right of the listener's head) in a virtual room whose absorption coefficients were set to four values: 1 (anechoic room), 0.7, 0.5 and 0.2. The reverberation level was varied across conditions, independently for target and interferer, such that intelligibility was disrupted by both the smearing of target speech and the reduction of binaural unmasking due to reverberation on the interferer.

3.1.2 Binaural de-reverberation

In their experiment 4 (referred to as LC4 in the following), Lavandier and Culling (2008) simulated the sources and listener at fixed positions in a virtual room (slightly wider than in LC3). The interferer source was located at 65° on the right of the listener's head while the target was straight ahead. The absorption coefficient of the room boundaries was fixed to 0.5

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

for the interferer, while two coefficients (1 and 0.2) were tested for the target. The interferer was always binaural, whereas the target was either binaural or diotic. SRTs increased when the target was reverberant rather than anechoic (temporal smearing), but this deleterious effect of reverberation was reduced when the target was binaural rather than diotic (see Fig. II.6). This reduction illustrates binaural de-reverberation as it is defined in this paper: in the presence of a reverberant target, SRTs are lower under binaural listening conditions compared to diotic conditions.

3.2 Model parameters and performance criteria

As discussed in section 2.2, three model parameters have to be defined for a given early/late separation: ELL, DD and window shape. In the literature, this separation process has often used the equivalent of a rectangular window with ELL as the unique parameter and its value changed quite significantly across studies. An early/late limit of 50 ms (“Rect₅₀”) has been used very commonly (Roman and Woodruff, 2013; Arweiler and Buchholz, 2011; Bradley *et al.*, 2003; Soulodre *et al.*, 1989) but other studies also used a limit of 35 ms (Bradley, 1986), 80 ms (Bradley, 1986) or 100 ms (Rennies *et al.*, 2011; Lochner and Burger, 1964). Because of the wide range of ELLs reported in the literature, the present study carried out a systematic test on the three model parameters in order to determine their role in reverberant speech recognition. Twenty-one ELL values were tested (from 0 ms to 100 ms each 5 ms), along with twenty-one DD values (from 0 ms to 100 ms each 5 ms) and two window shapes (linear and sigmoid). The rectangular window predictions were obtained from those of the linear window with $DD = 0$ ms.

Model predictions and experimental data were compared for each model setup. Prediction performance was assessed using the correlation coefficient (r), the mean absolute error ($\bar{\varepsilon}$) and the largest error (ε_{max}) across conditions between data and predictions, for each of the three experiments mentioned above.

3.3 Results

Figure II.2 presents the mean absolute prediction error across conditions as a function of ELL for the rectangular window. For the three experiments, the prediction error is first reduced with increasing ELL, it reaches a minimum and then increases for longer ELLs (even if not plotted here, the error increased for ELLs above 100 ms for the data of RBK). For RBK and LC3, involving temporal smearing and spatial unmasking, the prediction error is small over a broad range of ELLs. For an ELL between 40 ms and 200 ms for RBK and between

II.3 Validation of the room-dependent model and definition of the room-independent model

25 ms and 95 ms for LC3, the mean error is less than 1 dB. For LC4 involving binaural de-reverberation, the same mean error is reached for ELLs between 20 ms and 60 ms. Since the de-reverberation effect is only about 1 dB, the range of ELLs leading to good predictions of binaural de-reverberation is much narrower (30-40 ms) for LC4.

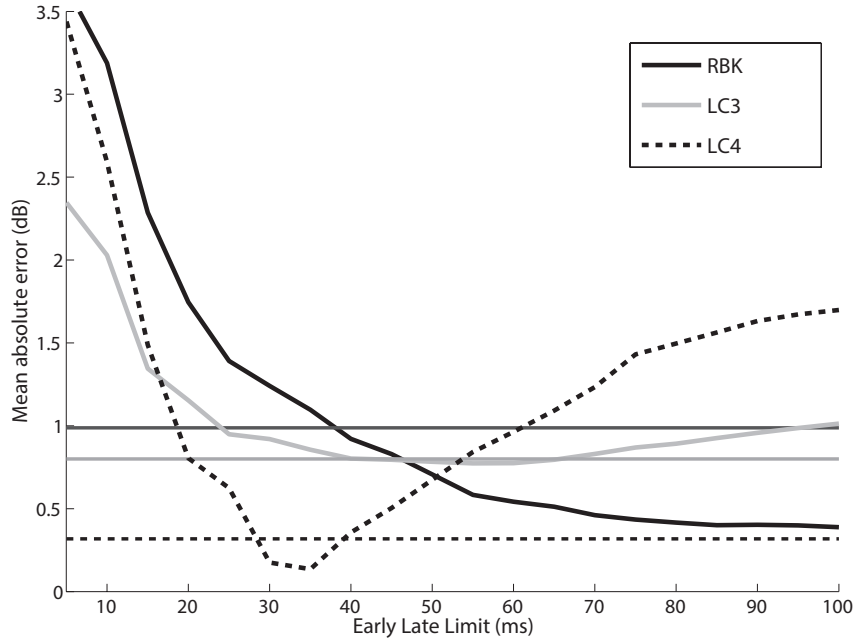


Figure II.2 – Mean absolute error between measurements and model predictions for each experiment as a function of early/late limit (ELL) for the rectangular window. The mean absolute errors of the room-independent (RI) model are plotted as horizontal lines.

Leclère *et al.* [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

Figure II.3 presents contour plots for RBK, LC3 and LC4 showing the prediction error as a function of ELL and DD with a linear window. In addition to the contour lines, a “best” point (black cross) and a minimum area (grey zone) are plotted. The “best” point represents the pair of parameters which leads to the smallest mean absolute error² (ε_{min}). The minimum area is the zone between the levels ε_{min} and $\varepsilon_{min} + 0.05 \times (\varepsilon_{max} - \varepsilon_{min})$. In this area, the prediction error is close to its minimum, within 5% of the spread of prediction errors. For each experiment, the influence of ELL on the prediction error follows the same pattern as in Fig. II.2

2. Since RBK and LC3 presented best performance for border values, the systematic tests have been purchased further than 100 ms for ELL (for RBK) and DD (for LC3). The best performance for RBK was still reached at ELL = 100 ms, while it was reached again at DD = 145 ms for LC3 with the same mean error as with DD = 100 ms ($\bar{\varepsilon} = 0.6$).

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

for the rectangular window. The differences across experiments mainly concern the gradient (along ELL and DD) of the mean error and consequently, the size of the area where the mean absolute error was minimized.

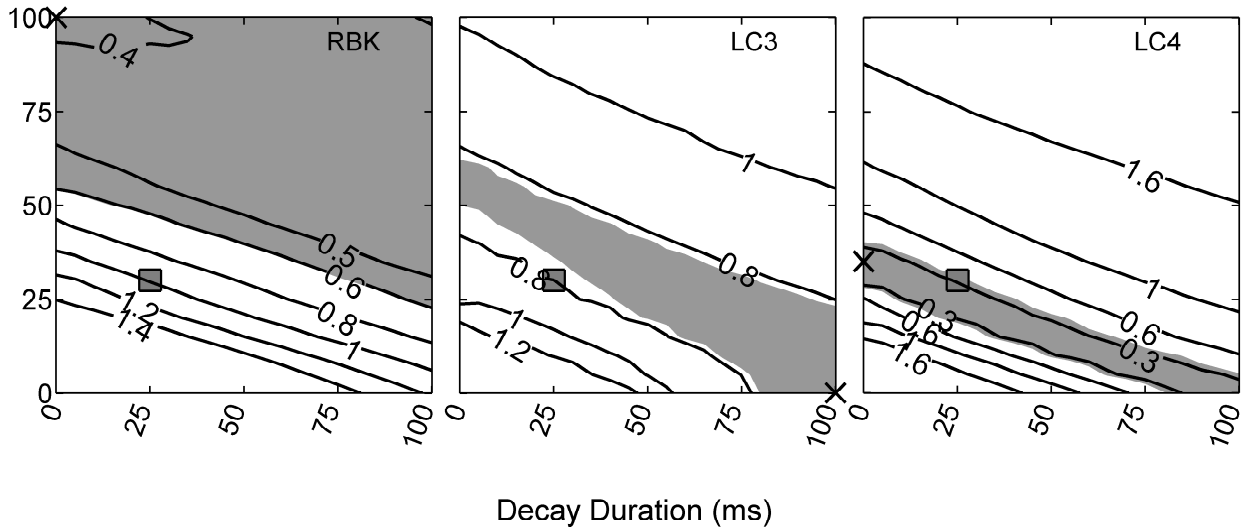


Figure II.3 – Contour plots of the mean absolute error between measurements and model predictions as a function of early/late limit (ELL) and decay duration (DD) for each experiment. The grey area represents the minimal error zone. The black cross indicates the smallest prediction error among all predictions. The grey square represents the error of the room-independent (RI) model ($ELL = 30$ ms and $DD = 25$ ms). Leclère *et al.* [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

The results obtained with the sigmoid window were very similar to those obtained with the linear window. On average across experiments, the correlation coefficient between the mean absolute error obtained with the linear and sigmoid windows was 0.99. On average across ELL and DD values, the differences of mean absolute errors were 0.05 dB (RBK), 0.01 dB (LC3), and 0.07 dB (LC4). The present study thus focused on the linear window (which is simpler to implement).

The three minimum areas obtained with RBK, LC3 and LC4 did not clearly overlap and the best performances were obtained for very different values of ELL and DD across experiments. These three sets of data did not lead to a unique and optimal value of the window parameters suggesting that the best performance of the model could be room-dependent (RD): the window

II.3 Validation of the room-dependent model and definition of the room-independent model

parameters of the proposed model have to be adjusted differently in each experiment to yield the best performance. In order to propose a room-independent (RI) model with a fixed window, a pair of parameters was chosen with a will to keep the binaural de-reverberation well predicted since it presents the smallest minimum area (the two other experiments should be more robust to the compromise). This pair of RI parameters is presented as a grey square on each contour plot ($ELL = 30$ ms, $DD = 25$ ms).

Figures II.4, II.5 and II.6 compare the measured SRTs to the RD and RI model predictions for RBK, LC3 and LC4, respectively. The predictions of the old model (Lavandier *et al.*, 2012; Jelfs *et al.*, 2011), without splitting the target BRIR, are also plotted. The predictions obtained with the RD model accurately fit experimental data, especially for RBK and LC4. A recurrent discrepancy occurred for the anechoic target in LC3. The RI model is less accurate than the RD model even though it does predict the trends associated with temporal smearing, spatial unmasking and binaural de-reverberation. The old model led to very poor performances by considering the entire reverberant target speech as useful.

The performances of three model configurations are compared in Table II.1: RD, RI ($ELL = 30$ ms, $DD = 25$ ms) and “Rect₅₀” (rectangular window with $ELL = 50$ ms commonly used in the literature). The best performance is achieved by the RD model according to r , $\bar{\varepsilon}$ and ε_{max} in the three experiments. The RI and Rect₅₀ models predict well the trends of temporal smearing and spatial unmasking in reverberation, as indicated by the high correlations obtained, but with less accuracy than the RD model (larger errors). Prediction accuracy is improved when the early/late parameters are adjusted to each room and only the trends are predicted with fixed parameters.

Experiment	RD			RI			Rect ₅₀		
	r	$\bar{\varepsilon}$	ε_{max}	r	$\bar{\varepsilon}$	ε_{max}	r	$\bar{\varepsilon}$	ε_{max}
RBK	0.98	0.4	1	0.97	1	2.1	0.98	0.7	1.7
LC3	0.90	0.7	1.2	0.86	0.8	1.3	0.83	0.8	1.5
LC4	0.99	0.1	0.3	0.99	0.3	0.6	0.99	0.7	1

Table II.1 – Prediction performance for each experiment for different model setups: room-dependent (different model parameters for each experiment, see Fig. II.4 to II.6), room-independent ($ELL = 30$ ms, $DD = 25$ ms, linear window) and “Rect₅₀” ($ELL = 50$ ms, rectangular window). Performance is assessed using the correlation coefficient (r), the mean absolute error ($\bar{\varepsilon}$ in dB) and the largest absolute error (ε_{max} in dB) between data and predictions.

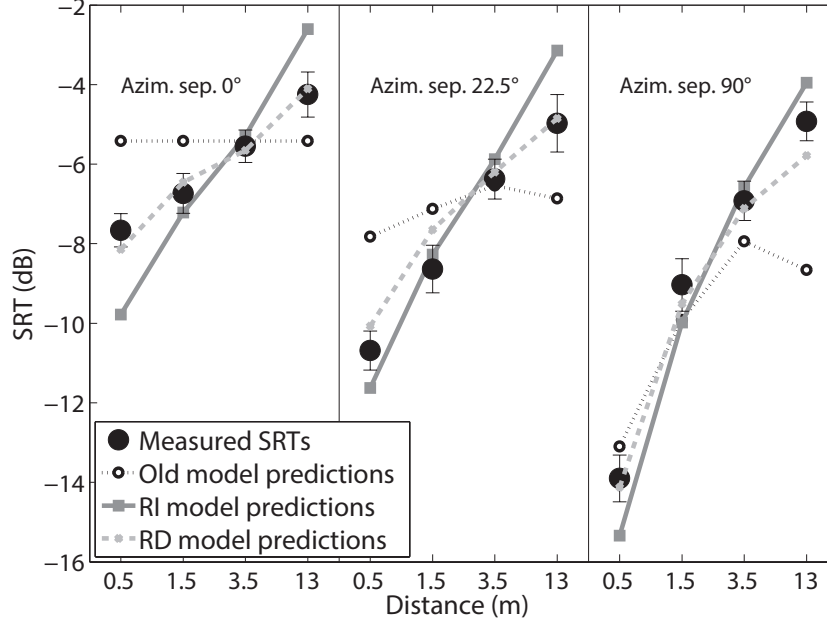


Figure II.4 – Mean SRTs (black circles) with standard errors across listeners measured by [Rennies et al. \(2011, RBK\)](#) as a function of target-to-listener distance and azimuth separation (Azim. sep.). Predictions are plotted for the room-dependent model (crosses; early/late limit [ELL] is 100 ms and decay duration [DD] is 0 ms), the room-independent model (squares; $ELL = 30$ ms, $DD = 25$ ms) and the old model (dotted line; without splitting the target BRIR into early and late parts).

Leclère et al. [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

4 Discussion

For each dataset, the model performance initially improved as soon as ELL or DD increased. This result confirms the usefulness of early reflections for speech intelligibility in rooms ([Arweiler and Buchholz, 2011](#); [Bradley et al., 2003](#); [Lochner and Burger, 1964](#)). Performance decreased when ELL or DD became too long, highlighting the detrimental effect of the late reflections on speech intelligibility.

In RBK and LC3, reverberation disrupted intelligibility by reducing the spatial masking release and by temporally smearing the target speech. The RD model accurately predicted these two effects with a similar level of performance as previous models in the literature ([Rennies et al., 2014](#); [Lavandier et al., 2012](#); [Jelfs et al., 2011](#); [Rennies et al., 2011](#); [Beutelmann and Brand, 2006](#)): $r > 0.9$, $\bar{\varepsilon} < 1$ dB and $\varepsilon_{max} < 1.5$ dB. A noticeable discrepancy of about 1 dB recurrently occurred for the anechoic target condition in LC3. It only concerned one BRIR which

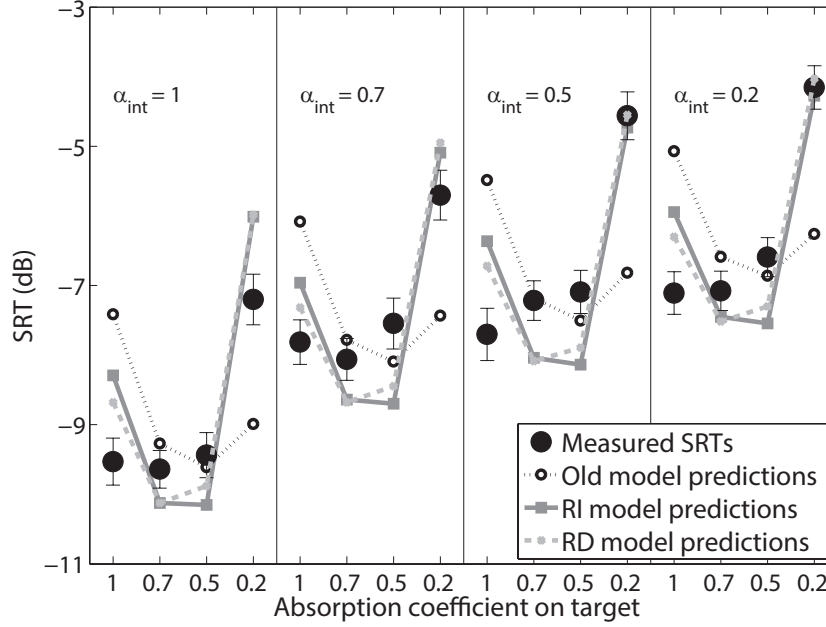


Figure II.5 – Mean SRTs (black circles) with standard errors across listeners measured by Lavandier and Culling (2008, LC3) as a function of the absorption coefficient used for the target and interferer (α_{int}). Predictions are plotted for the room-dependent model (crosses; early/late limit [ELL] is 0 ms and decay duration [DD] is 100 ms), room-independent model (squares; $ELL = 30$ ms, $DD = 25$ ms) and the old model (dotted line; without splitting the target BRIR into early and late parts).

Leclère *et al.* [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

was tested against four different maskers. According to the model predictions (even in its old version), the SRT decrease between the anechoic target and the moderately reverberant ones would be due to coloration which influenced the better-ear component of the model. Listeners did not seem to have taken any advantage of this coloration.

The monaural STI and U/D ratio cannot predict spatial unmasking nor the reduction of spatial unmasking caused by reverberation. The binaural model of Lavandier and Culling (2010) can predict these two effects: it predicts the decrease of SRT with increasing azimuth separation of sources (at fixed distances) and also the reduction of this spatial unmasking advantage with increasing source distance in Fig. II.4 (see also the prediction of the SRT increase with increasing reverberation for the interferer, at fixed reverberation levels for the target in Fig. II.5). However, this old model does not predict the temporal smearing of speech, as represented by the predicted SRTs remaining constant with increasing target distance in the first panel of Fig. II.4 (colocated source condition). Splitting the target BRIR into a useful

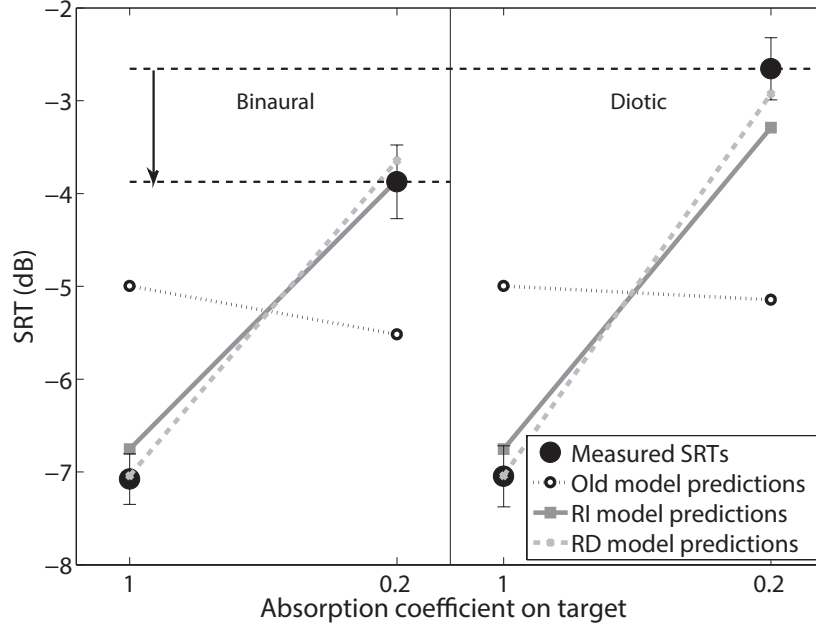


Figure II.6 – Mean SRTs (black circles) with standard errors across listeners measured by Lavandier and Culling (2008, LC4) as a function of the absorption coefficient and listening mode (binaural/diotic) used for the target. Predictions are plotted for the room-dependent model (crosses; early/late limit [ELL] is 35 ms and decay duration [DD] is 0 ms), room-independent model (squares; $ELL = 30$ ms, $DD = 25$ ms) and the old model (dotted line; without splitting the target BRIR into early and late parts). In the presence of a reverberated target, the benefit between binaural and diotic conditions illustrated by an arrow corresponds to the binaural de-reverberation effect.

Leclère *et al.* [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

and detrimental parts facilitated an extension of the model prediction ability to reverberant targets, while keeping accurate predictions for spatial unmasking.

Rennies *et al.* (2011) modeled their data by extending their binaural speech intelligibility model (BSIM) in three different ways: MTF, D_{te} and U/D. In the models using MTF or D_{te} , the binaural model is applied to the entire speech signal including the late reverberant part and the binaurally improved SNRs are corrected afterwards to take into account the temporal smearing of the target speech. As in the model proposed here, the U/D extension computes the early and late parts of the target before applying the binaural model to the useful (early target) and detrimental (late target + interferer) components. They observed similar levels of performance with the U/D and D_{te} models whereas the MTF approach induced a larger bias. Three ELL values (50, 80 and 100 ms) were tested with a rectangular window for the U/D and D_{te} extensions. The ELL of 100 ms gave the best predictions for both models: $r = 0.98$,

$\rho_S = 0.95$ (Spearman’s rank correlation) and $RMSE = 1.4$ dB (root mean square error) for U/D and $r = 0.98$, $\rho_S = 0.97$ and $RMSE = 1.1$ dB for D_{te} . The model proposed here yielded its best predictions ($r = 0.98$, $\rho_S = 0.97$ and $RMSE = 0.48$ dB) on the same data with the same window (rectangular with ELL of 100 ms). Rennie *et al.* (2014) tested the three approaches proposed by Rennie *et al.* (2011) on the data of Warzybok *et al.* (2013) in which reverberation was limited to a single reflection. In addition, they tested two temporal window shapes for the U/D and D_{te} versions: a rectangular window with $ELL = 100$ ms and a window equivalent to our linear window with $ELL = 0$ ms and $DD = 200$ ms. They observed the best performance with the U/D approach and a linear window, reaching a similar level of performance as the model proposed here: $r = 0.97$ and an $RMSE = 0.9$ dB across three noise conditions (diffuse, located at 0° or at 135°). They also tested six ELLs, four DDs and four window shapes in the case of a frontal reflection with a colocated or separated noise source.

The present study focused on the U/D approach and investigated the influence of each model parameter, extending the tests conducted by Rennie *et al.* (2014): all combinations of window parameters have been tested, and this was done in three different rooms with realistic reverberation. The conclusions of the present study were consistent with those of Rennie *et al.* (2014) concerning the shape of the window, indicating a minor influence of using either a linear or a sigmoid window. A clearer understanding of the influence of ELL and DD is also provided by Fig. II.3 which revealed ELL and DD can be adjusted to reach a given level of performance. Predictions obtained with a rectangular window ($DD = 0$ ms) can also be of the same accuracy as those obtained with a linear window ($DD > 0$ ms) as long as a different ELL is used. Thus, the parameter values required to reach a given prediction error are not unique, several window configurations can provide the same performance.

Previous studies (Roman and Woodruff, 2013; Arweiler and Buchholz, 2011; Bradley *et al.*, 2003; Soulodre *et al.*, 1989) often used a “Rect₅₀” window to separate early from late reflections in an impulse response. Early-to-late energy ratios (or clarity) are usually computed using a 50 ms limit for speech and an 80 ms limit for music (ISO 3382, 1997). Warzybok *et al.* (2013) highlighted the limitation of a rectangular window in presence of a single reflection. In this extreme case, such a window is clearly not suitable since the reflection is considered either as fully useful or fully detrimental. Conversely, in the presence of more realistic reflection patterns, the present study showed that the “Rect₅₀” window yielded similar correlations to the RI or RD models but with larger errors (Table II.1). The present work does not question previous uses of this window but it is pointed out here that the prediction is limited to an approximation of the temporal smearing effect. The “Rect₅₀” window does not appear suitable to predict binaural de-reverberation (LC4). An ELL of 35 ms (previously used by Bradley, 1986) rather than 50

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

ms led to a better performance for predicting this effect.

The systematic tests of the model parameters on RBK, LC3 and LC4 highlighted that the parameters giving the best prediction are room-dependent. This dependence could partially explain the wide range of ELL reported in the literature. Fixing the window parameters across experiments did not lead to satisfactory predictions: the RI model defined by these three experiments could predict the trends of temporal smearing and binaural de-reverberation but less accurately than the RD model. This would suggest that the U/D approach might not be sufficient to describe speech perception in rooms.

The validity of the RI model and its ability to describe the trends of speech transmission independently from the room was further tested on a fourth dataset which was not used to define its parameters. It involved temporal smearing and spatial unmasking in different rooms.

5 Room-Independent model validity

5.1 Experimental data

Van Wijngaarden and Drullman (2008) measured consonant-vowel-consonant (CVC) scores (which uses simple nonsense words embedded in carrier sentences) instead of SRTs to measure speech intelligibility in thirty-nine conditions. Among these thirty-nine conditions, only twenty-four were modeled here ([1-5; 8-12; 15-18; 22-24; 27-31; 35; 38]), excluding the conditions in quiet (they present an infinite SNR and CVC scores conversion into SRTs is possible only with finite SNRs; see section below) and the conditions in which noise was not convolved by a BRIR (since the proposed model requires BRIRs as inputs). Intelligibility scores were measured at different SNRs (-6, -3, 0, 3 and 6 dB) using headphones by simulating a target masked by a discrete speech-shaped noise in four listening environments: anechoic room, listening room, classroom and cathedral (see Table 1 of [van Wijngaarden and Drullman \(2008\)](#) for a detailed description of the conditions).

5.2 Scores transformation

To be compared to the model predictions, the experimental CVC scores were first transformed into SRTs according to the psychometric function proposed by ([Brand and Kollmeier, 2002](#), Eq. 1) which has the SRT and its slope at SRT as parameters. The slope can be deduced from the conditions which only differ in SNR. Such conditions should share the same psychometric function and SRT. Eight pairs of such conditions were identified (1/8, 2/9, 3/10,

4/11, 5/12, 15/17, 16/18 and 35/38). For each pair, the two SRTs obtained by transforming the CVC score with the psychometric function should be equal. It was not the case in practice since experimental errors occurred during the measurement. A unique slope value (9.68%/dB) was then determined with a least-square method such that it minimized this experimental error across the eight pairs. The score-to-SRT transformation was then applied to all modeled conditions using the same slope value.

Sixteen transformed SRTs (averages of each eight pairs and eight singles) were compared to the predictions obtained with the RI model (linear window, $ELL = 30$ ms, $DD = 25$ ms).

5.3 Results

Figure II.7 presents the transformed SRTs and the predictions from both the RI model and the old model (without splitting the target BRIR) for the sixteen conditions considered. The different panels refer to the tested rooms (anechoic room, cathedral, classroom and listening room). The abscissa refers to the condition index taken from Table 1 of [van Wijngaarden and Drullman \(2008\)](#). According to this table, spatial unmasking occurred in the anechoic conditions³ (between conditions 1, 2, 3, 4 and 5) as well as in the classroom (between conditions 27, 28, 29, 30 and 31). Temporal smearing of speech occurred in the cathedral conditions (between condition 15 and 16) as well as in the classroom (between condition 23 and 24). No binaural de-reverberation was highlighted in any condition.

For each room, the RI model defined in sect. 3 only described the trends of the transformed SRTs with a limited accuracy. By first averaging the transformed SRTs of each of the eight pairs, the correlation coefficient between experimental data and model predictions was $r = 0.96$, the mean absolute error over the sixteen conditions was $\bar{\varepsilon} = 1.77$ dB and the largest error was $\varepsilon_{max} = 4.87$ dB. The old model predicted less accurately this experimental dataset ($r = 0.65$, $\bar{\varepsilon} = 2.27$ dB and $\varepsilon_{max} = 7.43$ dB). The prediction errors were even larger than with the RI model, in some conditions. In particular, the old model did not predict the deleterious effect of temporal smearing (conditions 15/16 and 23/24).

3. Based on previous data on spatial unmasking in anechoic conditions ([Beutelmann and Brand, 2006](#); [Plomp, 1976](#); [Hawley et al., 2004](#)), we strongly suspect that labels have been switched among the anechoic conditions. This would explain some odd results: for instance, the target is more unmasked when the masker is located at 30° rather than 60° (conditions 3 and 4 or 10 and 11). We then decided to re-assign the labels of the conditions by conserving logical scores regarding spatial unmasking (except for 0° , the azimuth labels have just been shifted one rank upward such that the conditions 4/9, 3/10, 2/11 and 5/12 correspond to the azimuth 30° , 60° , 90° and 150° , respectively).

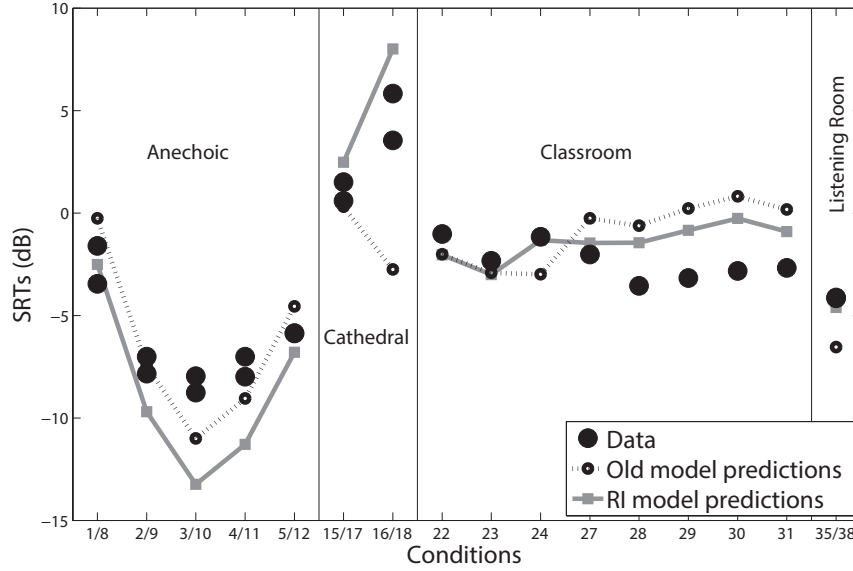


Figure II.7 – Transformed SRTs (black circles) from CVC scores measured by [van Wijngaarden and Drullman \(2008\)](#) in four rooms: anechoic room, cathedral, classroom and listening room. Predictions are plotted for the room-independent model (squares; early/late limit [ELL] is 30 ms, decay duration [DD] is 25 ms) and for the old model (dotted line; without splitting the target BRIR into early and late parts). The condition numbers are labelled as they appear in the Table 1 from [van Wijngaarden and Drullman \(2008\)](#), except for the re-assigned conditions³.

Leclère *et al.* [(2015). *J. Acoust. Soc. Am.* **137**, 3335-3345]

© *J. Acoust. Soc. Am.*

5.4 Discussion

The performance of the RI model for the experimental data from [van Wijngaarden and Drullman \(2008\)](#) was less accurate than the modeling of the other three experiments even though the trends of the different effects are described (resulting in a good correlation). Four rooms were tested in this experiment, that is the reason why it appeared suitable to test the RI model. Even if this model described the main trends in the data, it failed to accurately predict intelligibility in all conditions. This might indicate an inherent limitation of this model. The observed discrepancies across rooms confirm the room-dependence of the window parameters. Some sources of variability in the experimental and modeling processes might have also affected the model performance. First, only seven listeners participated in the experiment which contained 39 conditions and the variability in the experimental data was not presented in the results. The transformation of the CVC scores into SRTs implied a fitting of the psychometric function slope

(s_{50}), assuming it only depends on speech material. This fitting process prevents any direct comparison between data and predictions as performed with the three other experiments.

The predictions obtained with the old model did not fit to the experimental data. The largest errors occurred in presence of temporal smearing, while predictions were similar to the RI model for high D/R ratios. For instance, very accurate predictions were reached in the anechoic conditions (the offset between the two models being only due to the fact the predictions are compared to the data by fitting the averaged SRT across all 16 conditions, this average being different for the two models). The entire target BRIR is considered as useful and the detrimental part only consists of the noise BRIR, so that the two models are identical.

Van Wijngaarden and Drullman (2008) modeled their data by applying a binaural STI model using interaural correlograms from modulation transfer functions on the left and right ears. Since they compared their model to the STI reference curve instead of measuring its goodness of fit to the data, a direct comparison of performance is not possible.

6 General Discussion

6.1 Limitations of the U/D approach

In the four experimental datasets used in the present study, the predictions of the RI model were always limited to the trends of the effects. Adjustments on the early/late separation parameters were needed to yield accurate predictions. Unlike previous studies (Rennies *et al.*, 2014, 2011), the U/D approach was tested here in different rooms. It was thus able to highlight this room-dependence which might constitute a fundamental limitation to the U/D approach to predict speech intelligibility in rooms. The current version of the model cannot be used to make a priori predictions in different rooms. The early-late separation might depend on other parameters which are not taken into account in the current version of the model proposed here. To obtain both prediction accuracy and room-independence for the model, the early/late separation could be determined by modeling other perceptual mechanisms. For instance, previous studies showed that listeners are able to adapt to room acoustics thanks to prior exposure (Brandewie and Zahorik, 2010; Watkins, 2005). The proposed RI model would be improved by including this adaptation ability which might be related to room acoustics parameters: do listeners adapt to the particular BRIR or to the room as a whole? The separation between early/useful and late/detrimental parts of speech might also depend on the speech rate. The direct sound of a pronounced word can overlap with the reflection of the previous word depending on how fast the words are spoken, illustrating how a reflection can be regarded as

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

useful or detrimental depending on the speech rate. To account for this effect, the early-late separation could be made dependent on the frequency modulation in each frequency band, but this implementation would not be easy in the present model framework.

Room-dependence appears to be a relevant aspect of speech intelligibility modeling. This suggests that other approaches (MTF, D_{te}) should be considered as potential candidates to account for temporal smearing and tested across different rooms. Even if [Rennies *et al.* \(2011\)](#) implemented and compared the performance of these approaches, their room-independence should be investigated.

6.2 Unified interpretation of spatial unmasking, temporal smearing and binaural de-reverberation

By adding the late target to the interferer to constitute the detrimental input of the binaural process, the proposed model provides an interpretation of temporal smearing in terms of self-masking of the target induced by late reflections in the room. The late target is an additional masker, treated like any other interfering source by the model. Its effect appears at high levels of reverberation ([Lavandier and Culling, 2008](#)) because the late target needs to be sufficiently energetic to become a non-negligible new source of interference.

The RD model predicted correctly the effect of binaural de-reverberation in a narrow range of ELLs. According to the model, this ability to benefit from binaural listening in reverberant environments can be understood simply in terms of binaural unmasking of the early target against the late target. This interpretation is compatible with both the EC theory ([Durlach, 1972](#)) and the U/D ratio concept ([Lochner and Burger, 1964](#)). In diotic listening, early and late targets do not have any interaural phase differences, so cancellation is impossible and there is no binaural unmasking. For binaural targets, reverberation spreads part of the late energy to different interaural phases from that of the early target, so that the EC mechanism can eliminate a part of this late target (its coherence determining the level of cancellation). It should be noted that early and late targets might have different ILDs so that better-ear listening could also contribute to de-reverberation which would then involve the two components of spatial unmasking.

The interpretation of de-reverberation in terms of binaural unmasking is also consistent with the signal-processing technique proposed by [Allen *et al.* \(1977\)](#) to remove reverberation from speech signals. It consists in decomposing in frequency bands the signals from two microphones placed in the room and weighting the different frequency bands according to the cross-correlation of the two signals in each band, before synthesizing the composite de-reverberated

signal. Based on the hypothesis that the early signal is more correlated than the late signal at the two microphones, the weighting process aims at re-synthesizing only the coherent early part of the signal. The binaural system processes a similar cancellation of the late signal but this cancellation is based on differences of interaural phase difference between early and late targets rather than on coherence. The low coherence of the late reverberated target is a limitation for the binaural system which prevents the EC mechanism from cancelling the late target perfectly. This limitation could explain why Allen’s signal-processing technique was found to perform better than the binaural system.

Libbey and Rogers (2004) interpreted binaural de-reverberation as binaural overlap-masking release with reverberation acting as masking noise. They compared the ability to unmask reverberation and reverberation-like noise. The benefit of binaural listening was reduced with reverberation-like noise compared to reverberation. This could be explained by the fact that reverberation-like noises were constructed by randomizing the reverberation phases leading to uncorrelated noise. In contrast, reverberation is not totally uncorrelated and it is its correlated part which can be unmasked by the binaural system. Thus, the difference of performance did not necessarily reveal that two mechanisms were involved, but rather than a unique mechanism (spatial unmasking) behaved differently to different levels of correlation (as predicted by the proposed model).

Warzybok *et al.* (2013) investigated the influence of a single delayed reflection on frontal target speech masked by discrete noise. Their main findings are in good agreement with the conceptual interpretation of the proposed model. First, they observed no influence of the delay of a frontal speech reflection on spatial unmasking. Such a reflection cannot be unmasked since it has the same interaural phase as the target whatever the delay is, resulting in no BMLD. Second, the detrimental effect of long delays on a frontal reflection was reduced by separating the reflection from the target direction. Since a late reflection is regarded as a masker, unmasking is easier as soon as target and reflection are spatially separated. Third, in the presence of a discrete noise, the late reflection was less detrimental when it arrived from the same hemisphere as the noise than when it arrived from the opposite hemisphere. The binaural unmasking process in the present model is applied to the detrimental component (late speech + noise sources) which could be more coherent (so easier to cancel) when the masking sources come from the same spatial region.

Arweiler *et al.* (2013) investigated the integration of early reflections for improving speech intelligibility. Participants listened (monaurally or binaurally) to a frontal target (in anechoic or with early reflections) masked by a speech-shaped noise (diffuse or located at 90° on the right). Since no advantage was observed between the monaural and binaural conditions, the authors

Chapter II. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation

concluded that the integration process of early reflections with the direct sound “appears to be monaural for both the directional and the diffuse masker”, which, at first, does not seem in agreement with the concept of the binaural model proposed here. This model might however explain why no binaural effect was observed concerning the early/late integration process. First, late reflections were not involved, so that their binaural effect on interferer coherence could not be observed. Then, early reflections influence target interaural phase difference, but when the difference in interaural phase difference between target and interferer is large (which was the case in this study), the interaural phase difference of each source has little effect if any on binaural unmasking (Lavandier and Culling, 2010). So, the early/late integration was reduced to its monaural component in the particular conditions tested, and this study fits in the framework of the proposed binaural model.

7 Conclusion

A model computing binaural U/D ratios was proposed to simultaneously account for temporal smearing, spatial unmasking and binaural de-reverberation in reverberant environments. It combines a binaural model predicting spatial unmasking of a near-field target from multiple discrete noise interferers and a U/D decomposition taking into account the temporal smearing effect of reverberation on speech transmission. The early/late limit and decay duration used in the U/D separation both contribute to the model accuracy, but, it has been shown that these two parameters can be adjusted to reach a given prediction error, so that there is no unique way of defining early and late parts. The best model performance was achieved by adjusting the early/late separation for each experiment, leading to a room-dependent model. A room-independent model with fixed parameters was proposed, but it always predicted the trends of the temporal smearing with less accuracy than the room-dependent model. This result suggests that a fixed early/late separation might not be sufficient to predict speech intelligibility in rooms jeopardizing the generalization of the U/D approach to any room. However, the present modeling showed a unified interpretation of temporal smearing, spatial unmasking and binaural de-reverberation in terms of masking of early target (useful) by late target (detrimental) combined with unmasking by the binaural system. Temporal smearing during speech transmission is just masking from a particular interferer: the late target. Binaural de-reverberation is simply spatial unmasking of this particular interferer (or spatial un-self-masking of the target).

Chapter III

Speech intelligibility for a target and masker with different spectra

The experimental part of this study was mostly conducted by David Théry, a MSc student Mathieu Lavandier and I co-supervised during his internship. He conducted all the experimental work presented here and analyzed the outcome results. I worked on the modelling part presented in the last sections of this chapter. This work was done in collaboration with John F. Culling and the experimental part was presented at the International Symposium on Hearing (ISH, [Leclère *et al.*, 2015b](#)) in June 2015 (Groningen, The Netherlands).

1 Introduction

Speech intelligibility in noise is strongly influenced by the relative level of the target compared to that of the noise ([French and Steinberg, 1947](#); [Pollack, 1948](#)), referred to as signal-to-noise ratio (SNR). High SNRs lead to better speech recognition than low SNRs. But, although SNRs can take infinite values, it is not the case for speech intelligibility. Some SNR floors and ceilings must exist, such that intelligibility would not be influenced by varying the SNR above or below these limits. To account for this, the AI and SII proposed that speech intelligibility is only influenced by a limited range of SNRs [-15 dB; +15 dB] (see Fig. [III.1](#)). In these standards, variations of SNR below -15 dB or above +15 dB in any frequency band would not have any impact on speech intelligibility. This range has its origins in the work of [Beranek \(1947\)](#) who reported that the effective dynamic range of speech (EDRS, i.e. the speech level distribution) is about 30 dB by interpreting the short-term speech spectrum measurements reported by [Dunn and White \(1940\)](#).

Recent studies ([Studebaker *et al.*, 1999](#); [Studebaker and Sherbecoe, 2002](#)) suggested that the SNR range influencing speech intelligibility would be larger than the 30 dB reported by [Beranek \(1947\)](#). For instance, [Studebaker and Sherbecoe \(2002\)](#) measured intelligibility scores of monosyllabic words and derived the results into importance functions (IFs) which represent how much a given SNR in a specific frequency band contribute to the total amount of intelligibility. By comparing the obtained IFs to those adopted by the SII (Fig. [III.1](#)), they highlighted

that 1) the EDRS is larger than the 30 dB proposed by [Beranek \(1947\)](#) and 2) the EDRS varies with frequency. Note that it is difficult to accurately determine the EDRS from the IFs obtained by [Studebaker and Sherbecoe \(2002\)](#) because they are frequency-dependent and the importance declines for high SNRs. However, [Studebaker and Sherbecoe \(2002\)](#) reported that the widths of their IFs (between the minimum and maximum values) ranged between 36 and 44 dB with an average just over 40 dB.

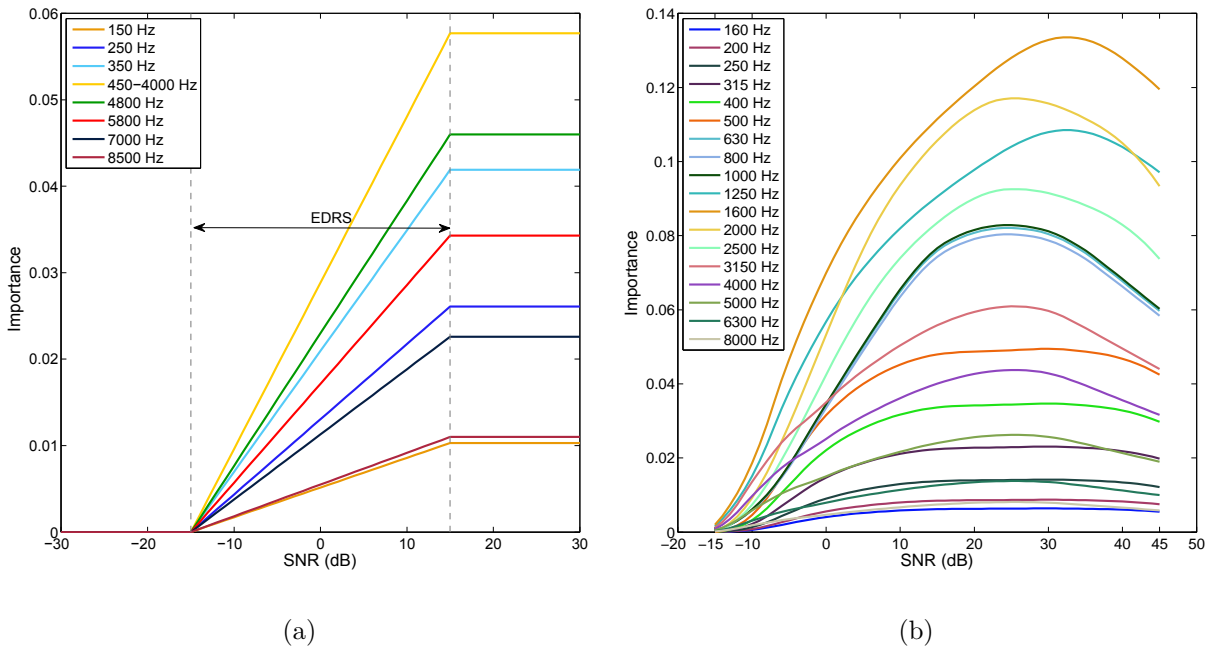


Figure III.1 – Importance (i.e. the effective proportion of speech signal conveying speech intelligibility which is available to a listener) as a function of SNR. (a): Importance functions adopted by the SII standard for each frequency band. Contribution to intelligibility linearly increases with SNR in the range $[-15 \text{ dB}; +15 \text{ dB}]$. This range is also called the effective dynamic range of speech (EDRS). Out of this range, importance is no longer influenced by the SNR. Also note that if SNRs above 15 dB are present in all frequency bands, the sum of the resulting importances across frequency gives 1. (b): Importance functions derived by [Studebaker and Sherbecoe \(2002\)](#). EDRS varies with frequency and importance varies non-linearly with the SNR.

These differences could have significant practical implications since the SII standard is widely used in many speech intelligibility models ([Beutelmman and Brand, 2006](#); [Beutelmman et al., 2010](#); [Rennies et al., 2011](#); [Rhebergen and Versfeld, 2005](#)) and because other models also implemented some aspects of the SII such as the band importance coefficients ([Houtgast and Steeneken, 1985](#); [Lavandier and Culling, 2010](#); [Wan et al., 2010](#); [Lavandier et al., 2012](#); [Wan et al., 2014](#); [Leclère et al., 2015a](#)). The binaural model of [Lavandier and Culling \(2010\)](#) and its different extensions ([Lavandier et al., 2012](#); [Leclère et al., 2015a](#)) do not use any SNR

limitation in the computations, meaning that the use of these models is limited to target and masking sources presenting similar spectra because if large spectral differences were present, it would result in infinite SNRs and then infinite predicted intelligibility. Collin and Lavandier (2013) extended the model of Lavandier *et al.* (2012) to account for maskers with a modulated temporal envelope by applying the model within short-time frames. They needed to implement a ceiling value for the SNR to prevent infinite SNRs when the masking level was near zero in a given time frame. After having tested a few ceiling values, they retained 10/15 dB as the SNR ceiling which best fitted the experimental data.

The aim of the present study was to determine floor and ceiling values based on four speech intelligibility experiments and to compare these values to those proposed by the SII. In contrast to the studies conducted by Studebaker and colleagues, the present study would determine these SNR limits by using sentences rather than monosyllabic words and by measuring Speech Reception Thresholds (SRTs) rather than scores in percent correct. This choice was also motivated by the fact that our model framework is particularly suitable to predict differences in SRTs.

SRTs were measured in the presence of a speech target and a speech-shaped noise (SSN). Target or noise was attenuated above or below 1400 Hz at different levels of attenuation in order to vary the SNR in the low or high frequency regions. The SRT was expected to increase along with the attenuation level in the case of a filtered target. Conversely, when the noise was filtered, the SRT was expected to decrease while the attenuation level was increased. In both cases, SRTs were expected to remain unchanged and reach an asymptote beyond a certain attenuation level, above which variations of SNR should not influence speech intelligibility any longer.

General methods of the experiments are presented first, detailing the conditions and stimuli tested in this study, then followed by the results of each experiment. These results are then discussed, and a model is finally proposed to describe the data, comparing different implementations.

2 General methods

2.1 Design of the stimuli

2.1.1 Target sentences

The speech material used for the target sentences was designed by [Raake and Katz \(2006\)](#) and consisted of lists of 12 anechoic recordings of the same male voice digitized and down-sampled here at 44.1 kHz with 16-bit quantization. These recordings were semantically unpredictable sentences in French and contained four key words (nouns, adjectives, and verbs). For instance, one sentence was “la LOI BRILLE par la CHANCE CREUSE” (the LAW SHINES by the HOLLOW CHANCE).

2.1.2 Maskers

Maskers were 3.8-s excerpts (to make sure that all maskers were longer than the longest target sentence) of a long stationary speech-shaped noise obtained by concatenating several lists of sentences, taking the Fourier transform of the resulting signal, randomizing its phase, and finally taking its inverse Fourier transform. Broadband levels of target and masker signals were first equalized before being filtered. All stimuli were heard in diotic conditions by the participants.

2.1.3 Filters

Digital finite impulse response filters of 512 coefficients were designed using the host-windowing technique ([Abed and Cain, 1984](#)). High-pass (HP) and low-pass (LP) filters were used on the target or the masker with different attenuations (0 to 65 dB) depending on the experiment. The cut-off frequency was set to 1400 Hz for both HP and LP filters to achieve equal contribution from the pass and stop bands according to the SII band importance function ([ANSI S3.5, 1997](#); [Dubno *et al.*, 2005](#)).

2.2 Tested conditions

Each type of filter (HP or LP) was tested on each source (target or masker), resulting in four experiments: HP target, LP target, HP masker and LP masker. Within one experiment, one source was filtered at different attenuation levels across conditions while the other source was kept unprocessed. Table [III.1](#) presents the different attenuation levels tested in each experiment. Except for experiment 1 (HP target) where only eight conditions were tested, each experiment

was composed of two sub-experiments of eight conditions because no asymptote was reached with the first set of eight conditions. Instead of testing only consecutive attenuation levels to sub-experiment 1, it was chosen to test a few attenuations inside the linear increase/decrease of SRTs observed in sub-experiment 1 (to allow for test/re-test comparison) in addition to further attenuation levels which would generate the expected asymptote. For each experiment, the two sub-experiments involved the same target sentences/speech material but different listeners.

Experiment	Attenuation Level (dB)	
	Sub-experiment 1	Sub-experiment 2
HP target	0 5 10 15 20 25 30 35	-
LP target	0 5 10 15 20 25 30 35	0 13 23 28 33 39 43 45
HP masker	0 5 10 15 20 25 30 35	0 18 40 45 50 55 59 64
LP masker	0 10 20 30 38 42 43 43	15 25 34 47 51 55 60 65

Table III.1 – Presentation of the different attenuation levels measured in the signals tested in each experiment. Because the first set of attenuations (sub-experiment 1) did not yield an asymptote of SRTs, a second set of higher attenuation levels (sub-experiment 2) was tested with the same sentences but different listeners.

2.3 Measurement of SRTs

Although the procedure was the same for all SRTs measured in this PhD (chapter III and IV), it is presented in each study to allow for an independent reading each chapter.

In each sub-experiment, each SRT was measured using a list of twelve target sentences and an adaptive method (Brand and Kollmeier, 2002). The twelve sentences were presented one after another against a different noise excerpt corresponding to the same condition. Listeners were instructed to type the words they heard on a computer keyboard after each presentation. The correct transcript was then displayed on a monitor with the key words highlighted in capital letters. Listeners identified and self-reported their score (number of correct key words they perceived). For the first sentence of the list, listeners had the possibility to replay the stimuli, producing a 3-dB increase in the broadband SNR, which was initially very low (-25 dB). Listeners were asked to attempt a transcript as soon as they believed that they could hear half of the words in the sentence. No replay was possible for the following sentences, for which the broadband SNR was varied across sentences by modifying the target level while the masker level was kept constant at 80 dB SPL (74 dBA). For a given sentence, the broadband SNR was increased if the score obtained at the previous sentence was greater than 2, it was decreased

if the score was less than 2 and it remained unchanged if the score was 2. The sound level of the k^{th} ($2 < k < 12$) sentence of the list (L_k , expressed in dB SPL) was determined by Eq. 1 (Brand and Kollmeier, 2002):

$$L_k = L_{k-1} - 10 \times 1.41^{-i} \times ((\text{SCORE}_{k-1}/4) - 0.5) \quad (\text{III.1})$$

where SCORE_{k-1} is the number of correct key words between 1 and 4 for the sentence $k - 1$ and i is the number of times $(\text{SCORE}_{k-1}/4) - 0.5$ changed sign since the beginning of the sentence list. The SRT was taken as the SNR in the passband averaged across the last eight sentences. The measured SRT then corresponded to the SNR in the passband for 50% intelligibility.

In each sub-experiment, the eight conditions were presented in a pseudorandom order and rotated for successive listeners to counterbalance the effects of condition order and sentence lists, which were presented in a fixed sequence. Within one sub-experiment, every listener heard only once each target sentence in the same order and, across the group of eight listeners, a complete rotation of conditions was achieved. In each experiment, listeners began the session with two practice runs, to be familiarized with the task, followed by eight runs with a break after four runs.

2.4 Determination of floor and ceiling values

The floor/ceiling value of each experiment was determined objectively by fitting a function to the experimental data and minimizing the mean absolute error between the data and the function (`fminsearch` function in MATLAB®). According to our hypothesis, the fitting function, $f(x)$, was chosen to be a line equation for low attenuation levels and a constant for high attenuation levels (Eq. III.2). The attenuation level at which SRTs stopped increasing/decreasing linearly to become constant corresponded to the researched floor/ceiling value. Three parameters needed to be adjusted to fit the data and then determine the floor/ceiling value ($\text{attenuation}_{limit}$ in Eq. III.2): a, b and SRT_{limit} which represent the slope of the linear function, its origin ordinate and the constant SRT when intelligibility plateaus, respectively. The variable x referred to the filter attenuation. The fitted curve is plotted in red solid line in each result figure presenting the measured SRTs of each experiment (see sect. 3).

$$f(x) = \begin{cases} ax + b & \text{if } x \leq \text{attenuation}_{limit} \\ \text{SRT}_{limit} & \text{if } x > \text{attenuation}_{limit} \end{cases} \quad (\text{III.2})$$

2.5 Equipment

Although the equipment was the same for all the experimental work conducted in this PhD (chapter III and IV), it is presented in each study to allow for an independent reading of each chapter.

Signals were presented to listeners over Sennheiser HD 650 headphones in a double walled soundproof booth after having been digitally mixed, D/A converted, and amplified using a Lynx TWO sound card. A graphical interface was displayed on a computer screen outside the booth window. A keyboard and a computer mouse were inside the booth to interact with the interface and gather the transcripts.

2.6 Listeners

Listeners self-reported normal hearing and French as their native language and were paid for their participation. Eight listeners took part in each sub-experiment. Within each experiment, no listener participated in both sub-experiments since the same target sentences were used in each sub-experiment.

3 Results

3.1 HP target

Figure III.2 presents the SRTs measured in the presence of a HP-filtered target as a function of the filter attenuation in the low-frequency region. SRTs first increased linearly from 0-dB to about 15-dB attenuation and then remained constant for further attenuations, i.e. intelligibility was not disrupted any longer after filtering out the target by about 15-dB attenuation. A one-factor repeated-measure analysis of variance (ANOVA) was performed on the experimental data, showing a main effect of the filter attenuation [$F(7, 7) = 10.58$; $p < 0.001$]. Tukey pairwise comparisons were performed on the data: SRTs obtained for attenuations of 0, 5 and 10 dB were significantly different while none of the SRTs from 10-dB to 35-dB attenuation were significantly different from each other.

For the HP target experiment, the floor value was determined at 13-dB attenuation and the slope of the linear increase of SRT was 0.53 dB SRT/dB attenuation. SRT was about 0 dB for higher attenuation levels than the SNR floor.

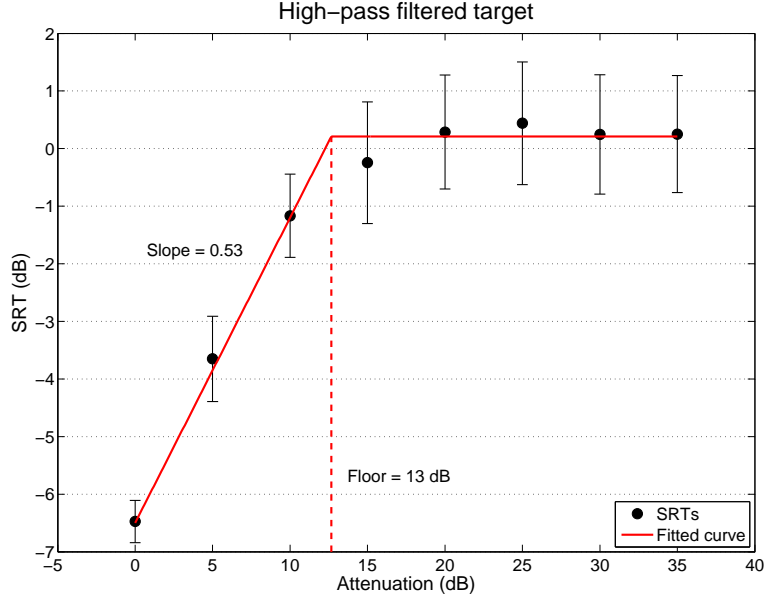


Figure III.2 – SRT measurements for a high-pass filtered target as a function of the filter attenuation. The red solid line corresponds to the fitted function used to determine the floor value.

3.2 LP target

SRT measurements in the presence of a LP-filtered target are plotted in Fig. III.3 as a function of the filter attenuation. As in the HP target case, SRTs increased linearly with a slope of 0.48 dB SRT/dB attenuation, but unlike the HP case, SRTs kept increasing until a floor value of 37-dB attenuation. Intelligibility remained at a SRT of about 10 dB for further attenuations.

A one-factor repeated-measure ANOVA was performed on each sub-experiment independently. In both sub-experiments, a main effect of attenuation was observed [$F(7, 7) > 16.2$; $p < 0.01$]. Post-hoc Tukey pairwise comparisons were performed on the data of each sub-experiment. In sub-experiment A (filled circles), the four SRTs at the lowest levels of attenuation (0, 5, 10 and 15 dB) were significantly different from all the other SRTs while the four SRTs at the highest levels of attenuation (20, 25, 30 and 35 dB) were not significantly different from each other. In sub-experiment B (open circles), the two SRTs at the lowest attenuation levels (0 and 13 dB) were significantly different from all the other SRTs while the six SRTs at the highest attenuation levels (23, 28, 33, 39, 42 and 45 dB) were not significantly different from each other.

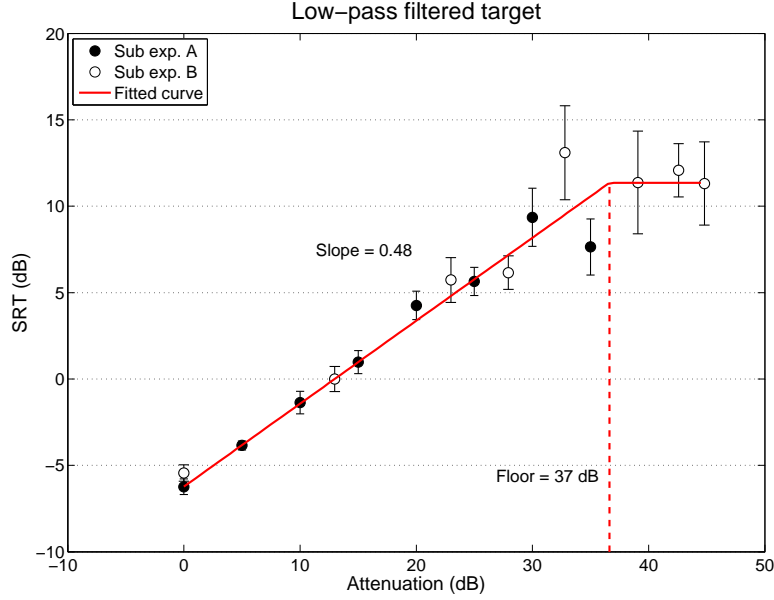


Figure III.3 – SRT measurements for a low-pass filtered target as a function of the filter attenuation. Filled circles correspond to the first sub-experiment while open circles correspond to the second. The red solid line corresponds to the fitted function used to determine the floor value.

3.3 HP masker

SRTs measured with a HP filtered masker are presented as a function of the filter attenuation in Fig. III.4. SRTs decreased linearly with a slope of -0.65 dB SRT/dB attenuation indicating an improvement of speech intelligibility by filtering out the low frequencies in the masker signal. At 43-dB attenuation (ceiling value), SRTs stopped decreasing and presented an asymptote at about -35 dB. Two one-factor repeated-measure ANOVAs indicated a significant main effect of the filter attenuation on speech intelligibility in each sub-experiment [$F(7, 7) > 41.8$; $p < 0.01$]. Tukey pairwise comparisons were performed on the dataset of each sub-experiment. In sub-experiment A (filled circles), only the pairs of SRTs at the attenuations 30/25, 10/15 and 0/5 were not significantly different. All the other pairs of SRTs were significantly different from each other. In sub-experiment B (open circles), SRTs for 0 and 18-dB attenuation were significantly different from each other and from all the other SRTs. None of the other SRTs were significantly different from each other.

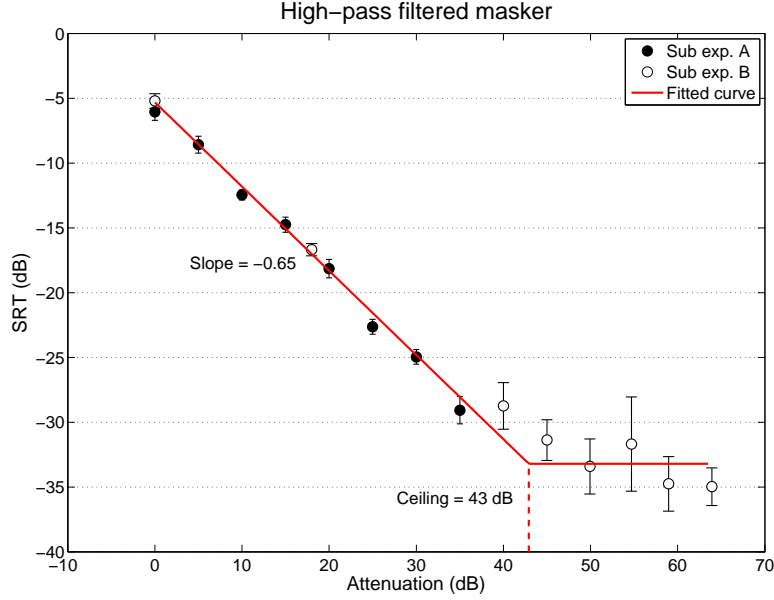


Figure III.4 – SRT measurements for a high-pass filtered masker as a function of the filter attenuation. Filled circles correspond to the first sub-experiment while open circles correspond to the second. The red solid line corresponds to the fitted function used to determine the ceiling value.

3.4 LP masker

Figure III.5 presents the SRTs measurements obtained in the presence of a LP filtered masker as a function of the filter attenuation. As in the HP masker case, SRTs linearly decreased with attenuation until the ceiling value of 36-dB attenuation. For further attenuations, SRTs were constant at about -35 dB. The slope of the linear decrease of SRTs was -0.76 dB SRT/dB attenuation. A one-factor repeated measures ANOVA was performed on the data from each sub-experiment independently. A main effect of the filter attenuation was found in each case [$F(7, 7) > 44.5$; $p < 0.01$], which was further investigated by performing Tukey pairwise comparisons. In sub-experiment A (filled circles), all SRTs were different from each other except for those corresponding to an attenuation of 38 dB and above. In sub-experiment B (open circles), none of the SRTs between 34-dB and 65-dB attenuation were significantly different. SRTs obtained for lower attenuations were all significantly different from each other and from all SRTs obtained at 34-dB attenuation and above.

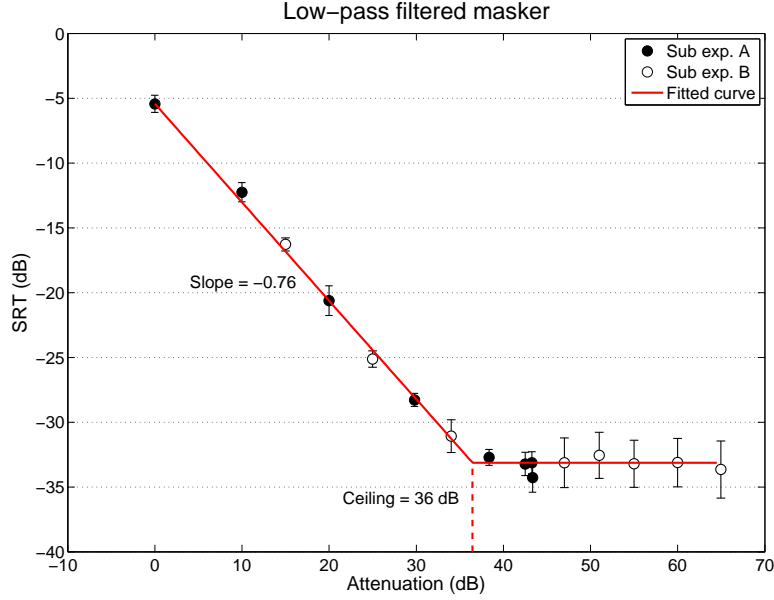


Figure III.5 – SRT measurements for a low-pass filtered masker as a function of the filter attenuation. Filled circles correspond to the first sub-experiment while open circles correspond to the second. The red solid line corresponds to the fitted function used to determine the ceiling value.

4 First discussion

Between the two sub-experiments of each experiment, SRTs measurements were in very good agreement at low attenuation levels since the corresponding SRTs shared the same linear increase/decrease (Fig. III.3-III.5). This was obtained using the same sentences with different listeners. In addition to illustrate very consistent results, this highlights the fact that eight listeners were sufficient to get reliable results in the present study.

In all experiments, the measured SRTs were in agreement with the experimental hypothesis mentioned in introduction: speech intelligibility was first influenced by the attenuation in a given band and, in a second time, remained constant for higher attenuation levels. The floor/ceiling values obtained in these experiments were not in agreement with the SII assumptions [-15 dB; +15 dB], except when the target was HP-filtered. In this specific case, the results suggested a floor value at 13-dB attenuation. The results from the other experiments suggested a larger range of SNRs contributing to speech intelligibility [-37 dB; 43 dB]. In addition, the present data showed that this range was frequency-dependent, in opposition to the SII standard which proposed the same floor/ceiling values for every frequency bands.

The obtained SNR range influencing speech intelligibility was even larger than previous

ranges reported by [Studebaker *et al.* \(1999\)](#) and [Studebaker and Sherbecoe \(2002\)](#) who already suggested that the 30-dB value proposed by [Beranek \(1947\)](#) and adopted by the SII standard needed to be revised. The increase of SNR range observed here could emanate from the difference in procedures with the study of [Studebaker and Sherbecoe \(2002\)](#). They used monosyllabic words filtered into octave bands. Roughly, two octave bands were presented to the listener. One band had a fixed sound level (pedestal band) while the sound level of the other band (the remote band, which was separated by at least one octave from the pedestal band) was varied between 19 and 91 dB SPL. Within the passband of the remote band, the noise was always presented at 44 dB SPL, resulting in tested SNRs in the remote band ranging between -25 and 47 dB (against -45 dB to +65 dB in the present study). However, the main discrepancy between the SNR ranges reported by [Studebaker and Sherbecoe \(2002\)](#) and those obtained here concerned the low limit of SNR, i.e. the floor value, below which speech intelligibility could not be further impaired. Since [Studebaker and Sherbecoe \(2002\)](#) tested this floor value in the presence of low speech levels in a limited frequency band (19 dB SPL in a one-octave band), it is likely that hearing threshold could have limited the speech detection rather than a masking effect due to the presence of the noise. By testing low SNRs at high sound levels (masker at 80 dB SPL) in a wider frequency band, the present study showed that speech intelligibility could be further impaired when speech was attenuated below -25 dB at high frequencies. When low frequencies were attenuated, a similar threshold was observed as in the study of [Studebaker and Sherbecoe \(2002\)](#), about -15 dB). The IFs proposed by [Studebaker and Sherbecoe \(2002\)](#) also attribute a negative contribution to very high SNRs (> 30 dB on average across frequency bands) regarding speech intelligibility. This negative contribution might be due to high sound levels used in their study which could have led to poorer word recognition scores ([Studebaker *et al.*, 1999](#); [Dubno *et al.*, 2005](#)). Their derived importance functions would then take into account both the effects of SNR and of absolute speech level. It could be preferable to separate the influence of these two effects.

Changing the filter type (HP or LP) had a small influence on the slope of the linear increase (or decrease) of SRTs before reaching the asymptote, indicating that the chosen low and high frequency regions would have contributed equally to speech intelligibility. However, the floor value depended on which frequency region of the target spectrum was attenuated (floors of -13 dB at low frequencies and -36 at high frequencies), i.e. for high attenuation levels, speech intelligibility was different depending on high or low frequencies were filtered out from the target signal. This seemed to indicate that, even though low attenuation levels led to similar changes in SRT in both frequency regions, the maximal contribution of low frequencies to intelligibility is greater than the contribution of high frequencies when the target source is filtered.

The results also showed that attenuating the masker had a larger impact on speech intelligibility than attenuating the target. The change of SRT was not the same as the SNR was varied up or down with the same amount. The benefit was greater when the noise was filtered compared to the detrimental effect when the target was filtered similarly. Like a previous study (Studebaker and Sherbecoe, 2002), this result questions the uniformly distributed importance over the $[-15; +15]$ interval adopted by the SII and suggests that a greater importance could be attributed to positive SNRs compared to negative SNRs. In other words, the SII standard considers that in a given frequency band, each increase of SNR would lead to the same benefit. In contrast, the present study suggests that the amount of benefit due to a SNR increment would be larger for positive SNRs than for negative SNRs (within a frequency band).

A high-level band limited noise generates masking on a target signal located at the upper and lower side of the noise spectrum limits. This is referred to as upward and downward spread of masking, respectively (USM and DSM, Egan and Hake, 1950). In the present study, these effects could have been highlighted in the LP masker experiment (USM) and HP masker (DSM): even though the high frequencies of the masker were filtered out, the low frequencies could still generate some masking effect on the higher part of the target spectrum and conversely when the low frequency region was attenuated. If USM was effective in the LP masker experiment (or DSM in the HP masker experiment), this should have resulted in a non-linear decrease of SRT because of the masking due to the noise located in the passband region, which is independent from the filter attenuation but increases as the target level decreases. When the filter attenuation increases, the target level decreases (due to the SRT procedure), the USM/DSM should then become less and less negligible and SRTs should not decrease linearly with attenuation. The SRTs did decrease linearly in this study. This seems to indicate that either no USM/DSM have disturbed the listeners in listening the target speech in the rejected band or the width of the spread of masking caused by the passband noise was narrower than the transition band of the filter, leading to a very small influence of USM/DSM in the rejected band.

5 Modelling

A first attempt at modelling the collected data is presented here. In the spirit of this PhD, the aim was to improve the model of Lavandier and Culling (2010) in order to predict the experimental while keeping the model as simple as possible. Indeed, in its current version, the model would compute SNRs in each frequency band, weight them using the SII band importance

coefficients, and sum them across frequency. No ceiling or floor limitation is present, resulting in an impossibility to predict the asymptote of SRT beyond a certain SNR limit.

5.1 New implementations

The model of [Lavandier and Culling \(2010\)](#) was first simplified into a monaural version since neither binaural unmasking nor better-ear listening was involved in the present study. The target-to-masker ratio (TMR) computed by the model is obtained from the sum of individual SNR_i weighted by band importance coefficients (I_i) in each frequency band (Eq. [III.3](#)).

$$TMR = \begin{bmatrix} SNR_1 & SNR_2 & \cdots & SNR_N \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix} = \sum_{n=1}^N SNR_n \cdot I_n \quad (\text{III.3})$$

SNRs were computed within classical third-octave bands instead of using a gammatone filterbank which presents similar attenuations in the rejected bands to the attenuations tested in the present study (Fig. [III.6](#)). This means that the frequency bands which were supposed to be rejected by the filter presented a similar level than the band where the filter response is the strongest, resulting in a biased computation of SNRs.

The first modelling step was to implement the limitation of the SNR range influencing the final TMR. Compare to the ceiling/floor values adopted by the SII, the results of the present study suggest larger SNR ranges which also depend on the frequency (Fig. [III.7](#)): [-13 dB; 43 dB] at low frequencies ($f < 1400$ Hz) and [-37 dB; 36 dB] at high frequencies ($f \geq 1400$ Hz).

It has also been observed that, for low attenuations, attenuating the target or the masker at the same level did not result in a similar absolute change of SRT. This was illustrated by a different slope where SRT linearly increases/decreases as a function of attenuation. It is then necessary to account for this difference in slopes in the linear region, depending on whether the masker or the target was filtered. A simple way to achieve this is to introduce a weighting coefficient which would reflect the larger impact that positive SNRs have regarding speech intelligibility compared to negative SNRs. This proposed non-linear implementation was inspired from the IFs suggested by [Studebaker and Sherbecoe \(2002\)](#) and would lead to the following modification on Eq. [III.3](#):

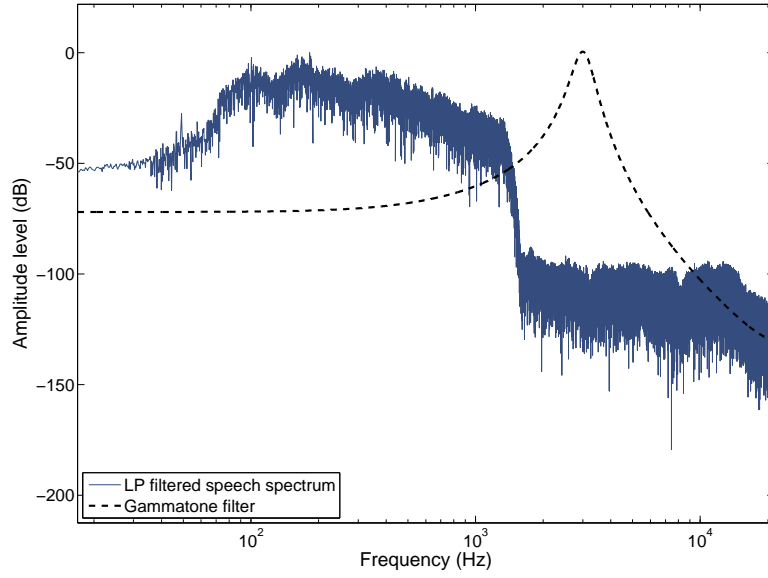


Figure III.6 – Response of a gammatone filter at 3 kHz. The attenuation in the rejected band is less important than the attenuation used in the stimuli, resulting in a biased consideration of the energy contained into a given frequency band.

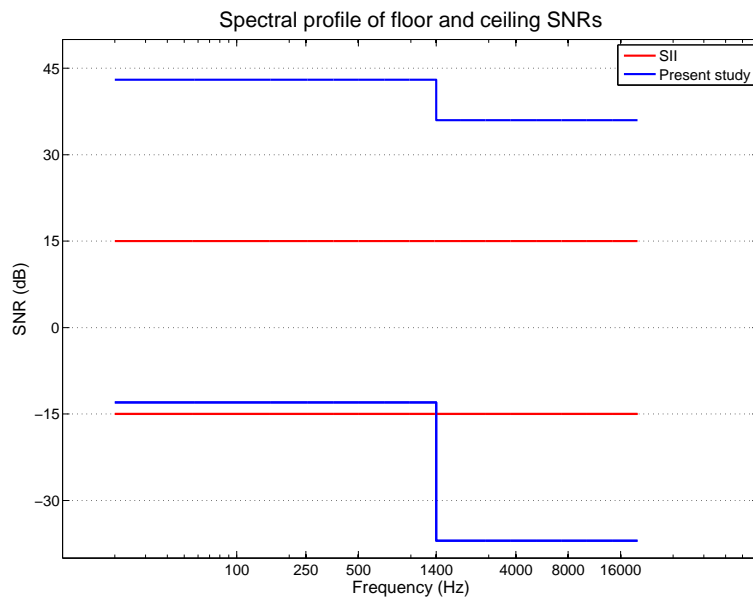


Figure III.7 – Spectral profile of the ceiling and floor SNRs adopted in the SII (in red) and derived in the present study (in blue).

$$TMR = \begin{cases} \sum_{n=1}^N \alpha \cdot SNR_n \cdot I_n & \text{if } SNR \geq \beta \\ \sum_{n=1}^N SNR_n \cdot I_n & \text{if } SNR < \beta \end{cases} \quad (\text{III.4})$$

To determine α , the absolute ratios between the slopes of the filtered masker and filtered target experiments for both HP and LP filtering were computed. It yielded ratios of 1.22 and 1.58 for the LP and HP case, respectively, which leads to an averaged ratio of 1.4. Therefore, by testing several values of α between 1.3 and 2 by 0.1, the mean absolute error between predictions and measurements was the smallest averaged across the four experiments for $\alpha = 1.6$ ($\bar{\varepsilon} = 0.58$ dB). In the experimental data, the difference in slope of the linear region could be observed as soon as the masking source was filtered instead of the target, which seemed to indicate that SNRs were more important when they were positive compared to when they were negative. Consequently, the β parameter was set to 0 dB.

The modifications brought to the importance functions (IFs) used in [Lavandier and Culling \(2010\)](#) are illustrated in Fig. III.8 for the frequency band centered at 2500 Hz. In the previous version of the model (red dashed line), an increase of SNR corresponded to a proportional increase of TMR in the frequency band. No floor or ceiling limitation was implemented, which could lead to infinite TMRs in the case of sources with very large differences between spectra. The proposed modified IFs (blue solid line) include a floor/ceiling limitation such that each band can carry a finite maximal contribution to the total intelligibility. They also present a non-linear behavior (illustrated by the change of slope), such that positive SNRs have a larger impact on speech intelligibility than negative SNRs.

5.2 Model performance

The original version of the SII standard was first used to predict the experimental data. It was not expected that the SII could accurately predict the results obtained here since 1) they are not in agreement with the SII assumptions concerning the ceiling/floor SNR values and their dependence with frequency and 2) the SII standard was not designed to account for sharply-filtered sources. However, it was of interest to take a look at the SII predictions as a reference point for this study.

Figure III.9(a) presents the computations of the SII as a function of attenuation level in the stimuli. The target and masker signals were equalized at the same level before filtering, i.e. at a 0-dB SNR in the passband, in order to highlight the influence of the attenuation level on the SII. The computed predictions indicate a linear increase of the index when the masker is

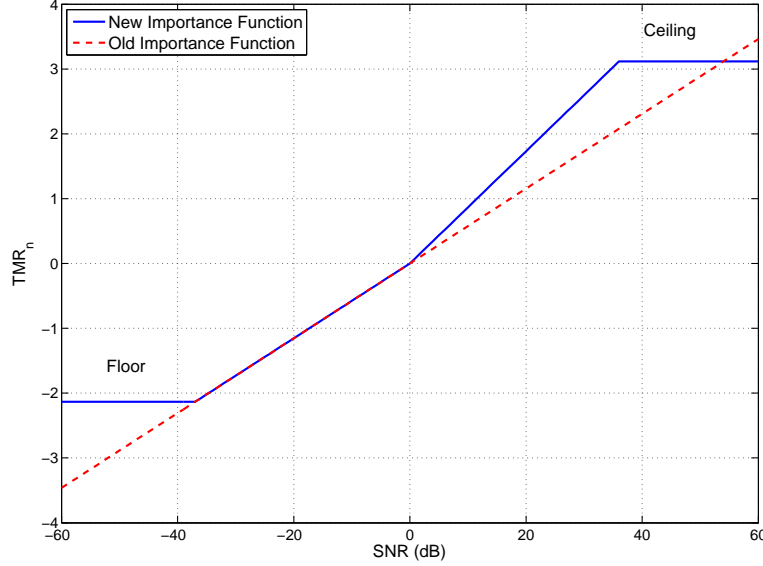


Figure III.8 – Example of the modifications brought to an importance function in the model of Lavandier and Culling (2010), for a frequency band centered at 2500 Hz. Compared to the IF used in the previous version of the model (red dashed line), the new IF (blue solid line) includes a SNR limitation (floor and ceiling) and a different slope for positive SNRs which reflects the larger impact they have on speech intelligibility.

attenuated and a linear decrease when the target is attenuated. In all experiments, the SII is constant for attenuations above 15 dB. To be compared to the measured SRTs, SII has been computed for target and masker equalized such that the SNR (before filtering) corresponded to the measured SRT in each condition. This way, the SII should indicate a constant intelligibility because all measured SRTs correspond, by definition, to the same amount of intelligibility (50% of identified key words). This was the case for the HP target experiment. This is certainly due to the close agreement between the observed floor value and the one used in the SII standard (-13 dB against -15 dB). For the other experiments, even though the SII only fluctuates between 0.2 and 0.4, it is not as constant as in the HP target experiment, reflecting some discrepancies between the SII assumptions and the observed results.

Since the SII failed to describe the experimental data obtained in the present study, three other models were tested. Figure III.10 compares the measured SRTs to the predictions obtained with a sum of SNRs across frequency bands (Eq. III.3) using three different importance functions, i.e. the implementation of I_i differed between models. To be compared to the measured SRTs the computed TMRs were first transformed into their opposite ($-TMR$) and,

independently for each experiment modelled, translated such that the mean values of $-TMR$ and that of the measured SRTs across conditions matched (Jelfs *et al.*, 2011; Lavandier *et al.*, 2012; Leclère *et al.*, 2015a).

First, the proposed model (Fig. III.7 and Eq. III.3) described above was tested on the experimental data of the present study. A very good correspondence between the measured SRTs and the computed TMRs was obtained in all experiments yielding a mean absolute error ($\bar{\varepsilon}$) of 0.9 dB for LP target, 0.2 dB for HP target, 0.40 dB for LP masker and 0.9 dB for HP masker with respective largest errors (ε_{max}) of 3.4 dB, 0.5 dB, 1.5 dB and 2.5 dB. Across experiments, $\bar{\varepsilon}$ was 0.6 dB on average and the largest ε_{max} was 3.4 dB (in the LP target experiment). By implementing new frequency-dependent ceiling and floor values and attributing more importance to positive SNRs, the experimental SRTs can be simply described with a sum of SNRs computed in third octave bands. However, the proposed model was only tested on the experimental data which has been used to define its parameters, so that its predictive power remains to be tested on external data.

A second weighted sum of SNRs was investigated, which relied on the SII parameters (dashed lines in Fig. III.10): the SNR computed in each frequency band was limited to the range $[-15$ dB

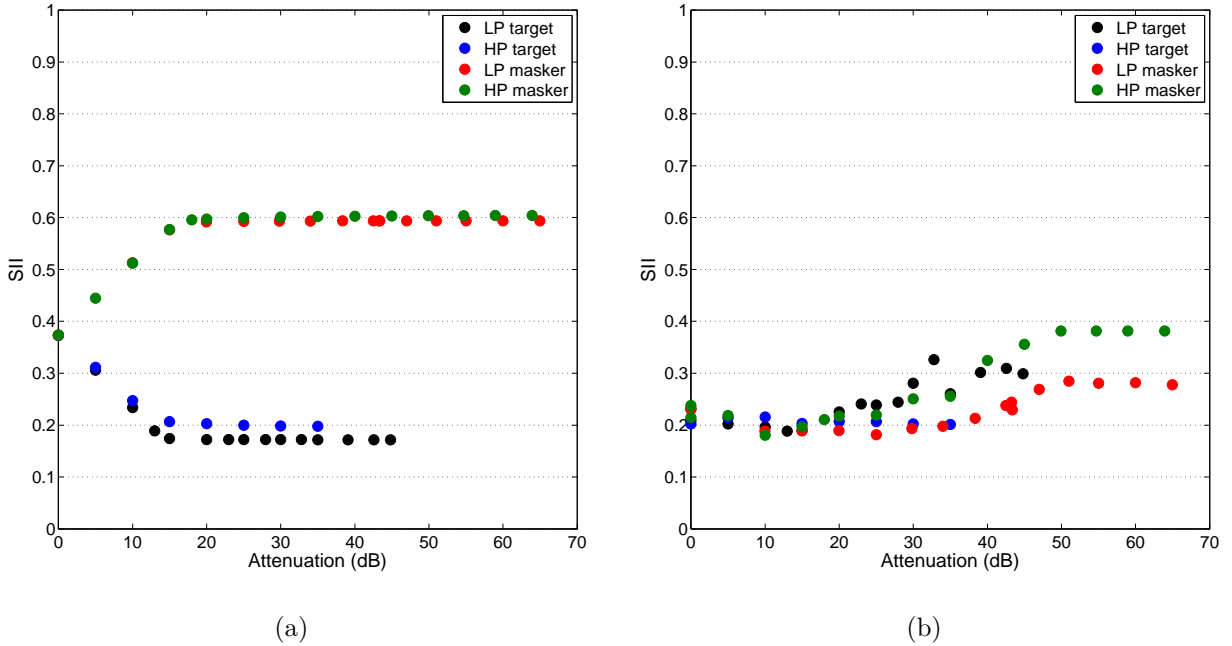


Figure III.9 – Speech Intelligibility Index computed according to the standard recommendations (ANSI S3.5, 1997) for two types of SNR equalization in the passband (a): SNR = 0 dB; (b): SNR = SRT measured in the present study.

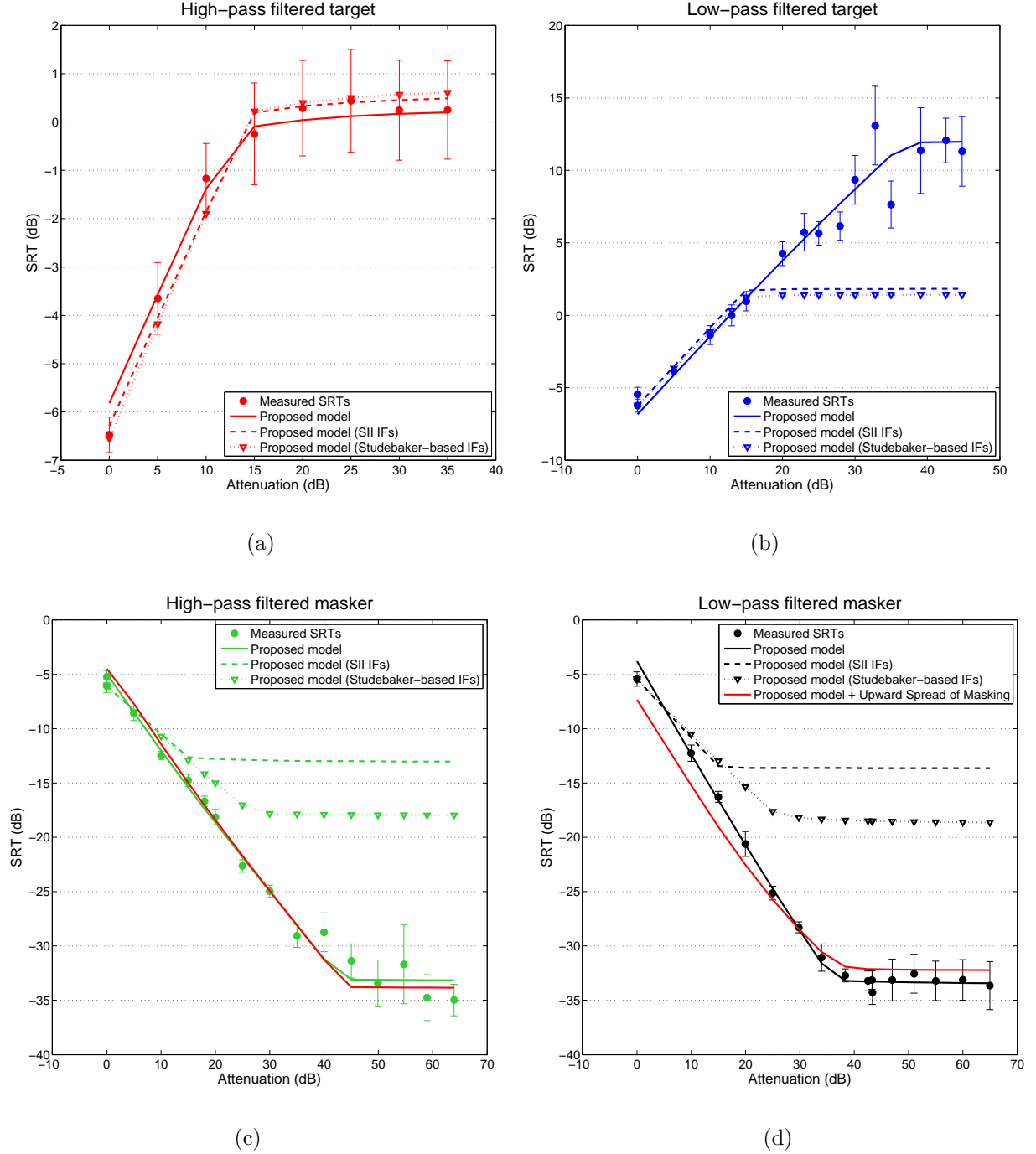


Figure III.10 – Measured (circle markers) and modelled SRTs of the four experiments of this study: HP target (a), LP target (b), HP masker (c) and LP masker (d). Three model implementations have been tested for each experiment: the SII parameters (dashed line), a modified version (see main text) of the importance functions proposed by Studebaker and Sherbecoe (2002, dotted line with triangle markers) and the importance functions derived from the present results. The red solid line in the LP masker (bottom-right, panel d) experiment corresponds to modelled SRTs using an additional implementation of the upward spread of masking (ANSI S3.5, 1997) in the model proposed in the present study.

; +15 dB] and weighted according to band importance functions proposed in the standard (, see Fig. III.1 ANSI S3.5, 1997). In the HP target case (Fig. III.10(a)), this implementation led to an accurate description of the experimental data by matching both the slope and the asymptote observed in the data ($\bar{\varepsilon} = 0.3$ dB and $\varepsilon_{max} = 0.8$ dB). Note that, compared to the other experiments, this experiment yielded a floor value which was the closest to the SII assumptions (-13 dB). When filtering out the high frequencies of the target (LP target, Fig. III.10(b)), the slope was still well described by using the SII parameters while the SRTs asymptote was completely underestimated: measured SRTs kept increasing beyond an attenuation of -15 dB in the rejected band, and this was ignored by the importance function of the SII which considered that all the contributions of the high frequency bands were null below -15 dB. This resulted in poor prediction performances ($\bar{\varepsilon} = 4.3$ dB and $\varepsilon_{max} = 11.1$ dB). The same discrepancy occurred for the HP and LP masker experiments (Fig. III.10(c) and (d)): importance functions of the SII predicted that intelligibility would remain unchanged by increasing the SNR beyond +15 dB in a specific frequency band. The experimental results were not in agreement with this assumption and speech intelligibility kept being improved when SNR was increased above the +15-dB limit. In addition, by using the SII parameters, the same (absolute) slope was predicted at low attenuation levels in all experiments which is not in agreement with the observations made in the present study: filtering the masker resulted in a larger absolute difference in SRTs than filtering the target with similar attenuation. The prediction performances obtained with these SII parameters were $\bar{\varepsilon} = 14.6$ dB and 10.7 dB with $\varepsilon_{max} = 20.8$ dB and 22.1 dB for the LP masker and HP masker experiments, respectively.

A third approach was tested by summing weighted SNRs based on the importance functions proposed by Studebaker and Sherbecoe (2002). They were implemented in the same way as the importance functions of the SII (SNR computation in each band, limitation by ceiling/floor values and weighting across frequency bands). The ceiling value was determined by taking the highest SNR having a positive importance in the density importance functions proposed by Studebaker and Sherbecoe (2002, Table II), i.e. the SNRs presenting negative importance in this table were ignored. The floor value was the same as in Studebaker and Sherbecoe (2002), i.e. it was set to -15 dB for every frequency band. The weighting coefficients of each band (I_i) were determined by selecting the maximum of the cumulative IFs reported by Studebaker and Sherbecoe (2002) (see Fig. III.1). Since the proposed model framework is based on a sum of weighted SNRs across frequency bands (Eq. III.3), the contribution of a given band to the total TMR only varies with the SNR value in the band and the band importance is kept fixed. In contrast, Studebaker and Sherbecoe (2002) and the SII standard are based on a sum of importance values across frequency bands to yield a final index. Contrarily to the SII,

Studebaker and Sherbecoe (2002) proposed that the importance did not vary linearly with SNR within a band, as if each increment of SNR did not provide the same amount of importance. They presented a look-up table to find the importance corresponding to a given SNR in a given frequency band. Because of the difference in frameworks, this characteristic was not implemented in the proposed model but was simplified by only considering the maximum of the IFs reported by Studebaker and Sherbecoe (2002) as the band importance (I_i).

This other implementation of IFs did not lead to more accurate prediction performances. As with the SII parameters, the experimental data from the HP target experiment were the best described by this model version ($\bar{\varepsilon} = 0.4$ dB and $\varepsilon_{max} = 0.8$ dB). The same discrepancies as already encountered with the SII parameters occurred with this model version: the SRT asymptote was reached at too low attenuation levels and the slope of the linear SRT decrease was underestimated. By presenting higher ceiling values than the SII, the predictions were a little closer to the data but it still resulted in poor model performances in the LP target, LP masker and HP masker experiments: $\bar{\varepsilon} = 4.5$ dB, 10.9 dB and 7.7 dB with $\varepsilon_{max} = 11.5$ dB, 15.9 dB and 17.2 dB, respectively.

5.3 Preliminary test of backward compatibility

The simple model proposed in the previous sections described the experimental data of the present study with good accuracy. A first attempt to implement its parameters (floor, ceiling, α , β values) into a more complex framework (Jelfs *et al.*, 2011; Lavandier *et al.*, 2012)¹ is carried out here by investigating the ability of the new model implementations to predict previous experimental data with a similar accuracy as the original version of the model (backward compatibility).

The framework of Lavandier *et al.* (2012) was first simplified by removing the computation of binaural unmasking since the way the parameters could be implemented into a binaural model was not in the scope of the present study. By only considering better-ear listening, the framework is slightly more complex than the proposed model in the previous sections and constitutes a good candidate to test the parameters proposed above. Indeed, better-ear listening consists in computing SNRs at each ear and retaining the highest SNR in each frequency band. Ceiling/floor and α/β parameters were then implemented once the left/right SNR comparison was made in the model.

1. The model developed by Jelfs *et al.* (2011) and Lavandier *et al.* (2012) is an extension of the one developed by Lavandier and Culling (2010) which only presents a few differences in the implementation. The main structure is still based on better-ear listening and binaural unmasking components (see sect. 3.2.5 in chapter I).

This new implementation was tested on some experimental data from [Lavandier et al. \(2012\)](#) and compared to the original predictions (made without the parameters implementation). The choice of the data from [Lavandier et al. \(2012\)](#) was motivated by the fact that, in some conditions, they used processed BRIRs such that the ITDs were removed while preserving ILDs (see the original paper for more details, only the conditions with processed BRIRs in experiment 1 are considered here). This way, only better-ear listening allowed for spatial unmasking, there was no binaural unmasking involved in these conditions.

[Lavandier et al. \(2012\)](#) conducted speech intelligibility tests with a target speech disturbed by speech-shaped noise maskers located at different positions in a room. The masker was tested at two distances, 0.65 m (near) and 5 m (far), and three azimuths, 25° (left), 0° (front), and 25° (right) whereas the target was always at near-right (0.65 m, 25°).

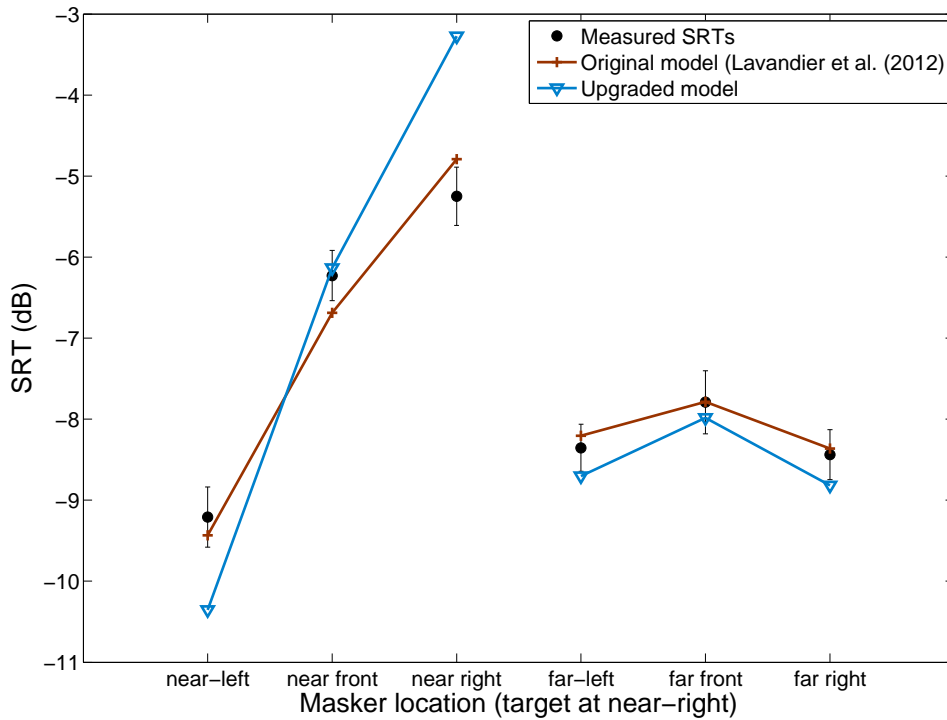


Figure III.11 – Predictions of the SRTs measured by [Lavandier et al. \(2012\)](#). The original predictions from the model version proposed by [Lavandier et al. \(2012\)](#), (brown line) are compared to the upgraded version where the proposed modifications have been implemented (blue line).

Figure III.11 shows the predictions from both the original model ([Lavandier et al., 2012](#)) and its upgraded version with the new implementations proposed in the present study ($\alpha = 1.6$ and $\beta = 0$ dB). As described previously, the computed TMRs from each version of the model were

first transformed into their opposite (-TMR) and translated such that the mean values of -TMR and that of the experimental SRTs across conditions matched. On average across conditions, the original model yielded a mean absolute error between predictions and measurements of $\bar{\varepsilon} = 0.2$ dB, while the new implementations led to a poorer prediction accuracy ($\bar{\varepsilon} = 0.7$ dB). Further analysis revealed that the implemented ceiling or floor SNR limits were never solicited in this experiment. The observed discrepancies between the two versions is then due to the α and β parameters. Even if it is not plotted here, the backward compatibility was ensured (i.e. the same predictions were made for both versions) when either $\alpha = 1$ or when $\beta = 15$ dB (or more). When $\alpha = 1$, the upgraded model is the same as the original version since the same importance is attributed to positive and negative SNRs. The same thing occurred when $\beta = 15$ dB since the computed SNRs at the better ear were all less than 15 dB, i.e. all SNRs have the same importance and the SNR region where α switches from 1 to 1.6 was never solicited.

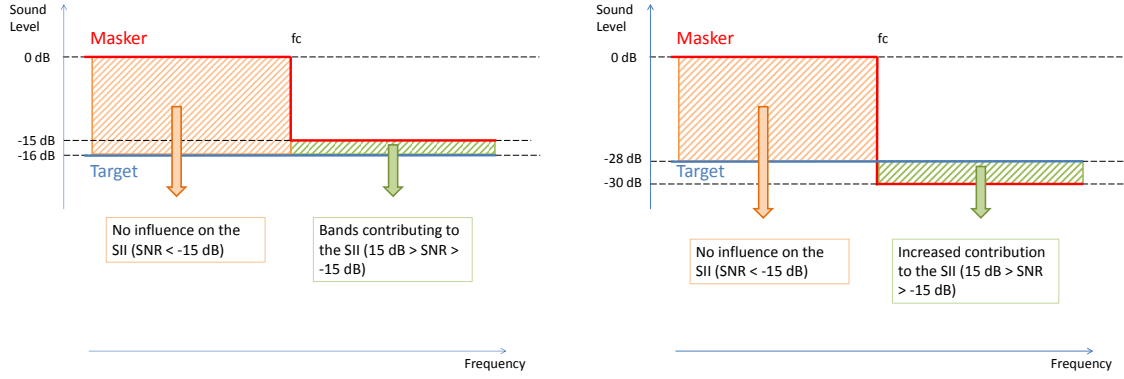
6 Second discussion

The computed SII with target and masker being equalized at the SRT (Fig. III.9(b)) were not constant as they should have been since all measured SRTs present a constant intelligibility. Except in the HP target experiment, the SII increased with attenuation level. This observed increase of SII between the attenuations 15, 30 and 50 dB is illustrated in Fig. III.12. When the target level is at the SRT, only the SNRs in the rejected band contribute to the total SII because SNRs in the passband are below the -15-dB floor. While increasing the attenuation, SNRs also increase in the rejected band, and this increase is not counterbalanced in any way by the SNRs in the passband which are still below the floor limit. Therefore, the computed SII only reflects the information available in the rejected band and ignores the variations of SNRs in the passband which seemed to have influenced speech intelligibility in the present study.

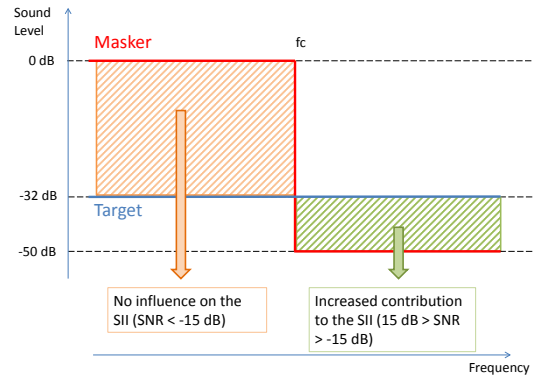
The different implementations tested in the same model framework yielded different performances in the predictions, highlighting the influence of the IFs and of the ceiling/floor SNRs profiles used in the model. The widely adopted [-15 dB; +15 dB] SNRs limits led to an underestimation of the SRT asymptote: speech intelligibility could be further improved above +15-dB SNR and impaired below -15-dB SNR (except when the target was high-pass filtered). As suggested by Studebaker and Sherbecoe (2002), a non-uniform distribution of importance along this range of SNR could account for the difference in slope observed depending on whether the target or the masker was filtered.

The proposed importance functions do not take into account the detrimental effect of the

Chapter III. Speech intelligibility for a target and masker with different spectra



(a) Masker attenuated by 15 dB at high frequencies. The measured SRT was about -16 dB. (b) Masker attenuated by 30 dB at high frequencies. The measured SRT was about -28 dB.



(c) Masker attenuated by 50 dB at high frequencies. The measured SRT was about -32 dB.

Figure III.12 – Schematic illustration of the SII increase between the 15-dB, 30-dB and 50-dB attenuations in the LP masker experiment.

target presented at high sound levels on speech intelligibility (Dubno *et al.*, 2005; Studebaker *et al.*, 1999). In the importance functions proposed by Studebaker and Sherbecoe (2002), the importance decreases above a certain SNR (about 30 dB on average across frequency), which creates an ambiguity because a high SNR does not necessarily mean a high target level. Studebaker and Sherbecoe (2002) combined the influence of the absolute sound pressure level and of the SNR in their importance functions. The implementation of such functions should be easier if these two aspects were separated as in the SII standard which includes a distortion factor (noted L_i in the standard) to describe how speech intelligibility is impacted by high sound levels of the target speech. In the present study, the importance functions derived from the

SRT measurements did not present any decrease at high SNRs. The present model framework disregard the absolute level of the sources and only focus on the target-to-masker ratios in each frequency band.

Figure III.10(d) compares the predictions from two versions of the model proposed above, including or not the upward spread of masking. The USM effect was implemented as presented in the SII standard (ANSI S3.5, 1997). With the USM version, the predicted SRTs decrease with a smaller slope because of the additional masking operating in the rejected band. This masking is more and more effective as the attenuation level increases as indicated by the slight change of slope before the SRT asymptote. The version without USM yields a better correspondence between data and predictions, which raises two questions: was USM really not effective in the present experiment? Is the computation of USM valid for the considered experiment? According to the experimental data, intelligibility did not seem affected by USM, but the two models (with and without USM computations) were not in agreement with this observation since they yielded different predictions. If USM was really negligible in the present experiment, the model including USM would have reach similar predictions than the model without USM. Either the way USM was computed was not suitable to describe the data of the present study, or another perceptual effect interacted with the USM effect and was not taken into account in the model including USM. These questions/interpretations are still opened and constitute a potential research perspective of this work.

The new implementations proposed in the present study were inserted in the original model of Lavandier *et al.* (2012) in order to test the backward compatibility of the upgraded model. This latter failed to predict the measured data of Lavandier *et al.* (2012) with the same accuracy as the original model. This was due to the α and β parameters which were not suitable for this dataset. It is important to note that α and β parameters were derived from the experimental results of the present study. Although the proposed model with these parameters yielded good prediction performances, it could be seen as a first and simple attempt to describe the collected results. Further investigations should be conducted to ensure the backward compatibility of this model. The α and β should be reconsidered by either search for optimized values which can lead to good prediction performances across several external datasets. Or, the brutal change in SNR-importance at 0 dB could also be considered differently with a smoother transition between “important” and “less-important” SNRs regarding speech intelligibility.

7 Conclusions and perspectives

Speech reception thresholds were measured in four speech intelligibility experiments. Target or noise was either high-pass filtered or low-pass filtered, creating different SNRs in the rejected band by varying the attenuation level of the filter. As expected, it was observed in each experiment that SRTs remained constant beyond a certain attenuation level (i.e. a certain SNR in the rejected band). In general, the SNR value from which SRT was not influenced any longer differed from previous values reported in the literature, especially in the SII standard. These results provide ceiling and floor values of SNR for wide frequency bands based on experimental measurements. They do not question the validity of the SII (which was not designed for sharply-filtered sources) but rather point out the need of non-linear SNR-importance functions in speech intelligibility models based on SNR weightings to predict SRTs, especially if these models aim to account for sources with very different spectra.

A simple model was proposed to account for such target and masker presenting different spectra. It is based on a weighted sum of SNRs across third-octave bands. It includes a floor and ceiling limitation of the SNR, which restricts the influence of a given SNR to the total intelligibility and avoids infinite predicted SRTs. As observed in the experimental results, the different influence of positive SNRs compared to negative ones on speech intelligibility has been taken into account by introducing different weightings depending on the sign of the SNR in a given frequency band.

Further work needs to be done to more precisely determine ceiling and floor values in narrower frequency bands, especially for the transition between low and high frequencies which is discontinuous here. The predictive power of the model outside the dataset used to define its parameters has not been tested yet. It is then necessary to test the model on external data. Finally, the model could be extended to binaural listening since the SII or SII parameters are widely used in binaural speech intelligibility models.

Chapter IV

Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

This work has been conducted in collaboration with Mickael L.D. Deroche. It is currently under review for a publication in the Journal of the Acoustical Society of America ([Leclère et al., 2016](#)).

1 Introduction

To unmask a target voice among masking sources, a listener can rely on several auditory mechanisms such as F0-segregation, spatial unmasking and temporal dip listening. Speech intelligibility is improved when the F0 difference ($\Delta F0$) between a speech target and a harmonic masker is increased ([Brokx and Nootboom, 1982](#); [Summerfield and Assmann, 1991](#); [Culling and Darwin, 1993](#); [Deroche and Culling, 2013](#)), when target and masker are spatially separated ([Plomp, 1976](#); [Hawley et al., 2004](#); [Beutelmann and Brand, 2006](#)) or when the broadband envelope of the masker presents modulations or temporal gaps ([Festen and Plomp, 1990](#); [Gustafsson and Arlinger, 1994](#); [Beutelmann et al., 2010](#); [Collin and Lavandier, 2013](#)). In the literature, these cues and their influence on speech intelligibility have been tested independently. But it remains unclear how would each auditory mechanism behave when several cues are present at the same time, as mostly encountered in realistic situations. The present study aims to investigate the resulting benefit when these three mechanisms operate in pairs.

A few studies examined the influence of the presence of both modulations in the masker envelope and spatial separation between competing sources on speech intelligibility. [Carhart et al. \(1966\)](#) observed that listeners benefited from both the envelope modulation and the interaural phase difference of a white noise masker allowing spatial unmasking. More recent studies ([Hawley et al., 2004](#); [Jones and Litovsky, 2011](#); [Collin and Lavandier, 2013](#)) confirmed these findings by separating speech target and noise masker using head-related transfer functions (HRTFs). It has been repeatedly observed that the benefit from each cue added up as if

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

listeners independently relied on both cues to improve speech intelligibility. Thus, it can be assumed that spatial unmasking does not interact with temporal dip listening. The present study investigated the potential interactions between F0-segregation and spatial unmasking and between F0-segregation and temporal dip listening.

Hawley *et al.* (2004) could have examined these interactions by presenting a target voice against four types of masker: speech, reverse-speech, stationary or speech-modulated noise. The spatial location of the masker varied while the target remained in front. Similar spatial masking releases (SMRs) were observed for each type of masker. Compared with stationary noise, SRTs for the three other masker types were lower, indicating a benefit presumably due to temporal fluctuations in the masker envelope. Furthermore, SRTs were lower for speech and reversed-speech masker than for speech-modulated noise, suggesting that in addition to the role played by temporal envelope modulations, F0 differences between the harmonic sources could have been used as well. But using a speech or reverse-speech masker generates informational masking. Their masking stimuli were then not suitable to highlight the interactions investigated here and the masking release (MR) observed by Hawley *et al.* (2004) might not be attributed to F0 differences only. The present study aims to further investigate whether the benefit solely due to $\Delta F0$ linearly adds up with the spatial unmasking benefit or with the benefit due to modulations in the temporal envelope of the masker by using a non-linguistic type of harmonic masker as a way to limit informational masking.

It could be that these mechanisms are independent one another, so the total MR will simply be the sum of the individual benefits. However, there are a number of reasons why this may not be the case. First, the amount of MR caused by a given cue is largely dependent on how much masking there is to start with. Take an extreme example where two sources may have been filtered in different frequency bands, the MR provided by any acoustic cue would be very limited since there is not so much to unmask. In realistic cocktail-party situations, it may well be that one cue provides a substantial amount of MR, leaving little room for another cue to clear the auditory scene further. This sort of interaction can generally fall under the category of ceiling effect of MR (or floor effect of masking), and refers to the fact that there may be times during a target sentence where target audibility is perfect because masking would have been already eliminated, before the cue under investigation could be helpful. It is a rather simple form of interaction where each mechanism still operates in the same way whether another cue is present or not. Second, there may be a more genuine interaction when the mechanism with which a given cue provides MR actually facilitates or on the contrary impairs the action of a second mechanism based on a different cue. This sort of interaction is a lot more interesting as it could inform about the way these mechanisms operate together.

In realistic situations, competing voices have natural F0 fluctuations over time. Previous studies investigated the influence of these F0 variations by either using deterministic F0 patterns (Deroche and Culling, 2011) or actual speech (Hawley *et al.*, 2004; Jackson and Moore, 2013) as the masking source. Deroche and Culling (2011) observed a detrimental effect of the masker F0 modulation using harmonic complexes sinusoidally modulated by ± 2 semitones at 5 Hz. Jackson and Moore (2013) measured speech reception thresholds (SRTs; levels of the target compared to those of the masker for 50% intelligibility) by varying the mean F0 difference ($\overline{\Delta F0}$) between the speech target and a background speech masker. Target and masker were either both monotonized (steady F0 across time) or both intonated (fluctuating F0 over time). They did not observe any significant improvement of speech intelligibility by increasing $\overline{\Delta F0}$ when both sources were intonated, whereas it was the case when both sources were monotonized. Unfortunately, Jackson and Moore (2013) manipulated the F0 of both sources together, leaving it uncertain whether the effect of F0 variations was mediated by those of the target or those of the masker, a gap that the present study fills by holding the F0 of one source fixed while letting the other fluctuating naturally.

Two experiments were conducted to investigate how listeners benefited from two simultaneous cues. Experiment 1 focused on spatial separation and $\Delta F0$, while experiment 2 dealt with $\Delta F0$ and modulations in the masker envelope. In both experiments, SRTs were measured for two types of target (monotonized or intonated) against two types of harmonic masker (monotonized or intonated).

2 General Methods

2.1 Stimuli

2.1.1 Target sentences

The speech material used for the target sentences was designed by Raake and Katz (2006) and consisted of 16 lists of 12 anechoic recordings of the same male voice digitized and down-sampled here at 44.1 kHz with 16-bit quantization. These recordings were semantically unpredictable sentences in French and contained four key words (nouns, adjectives, and verbs). For instance, one sentence was “la LOI BRILLE par la CHANCE CREUSE” (“the LAW SHINES by the HOLLOW CHANCE”). The sentences presented on average 65% of voiced parts. Among the 35% left, 20% were unvoiced parts and 15% were silent pauses between words or syllables. The mean F0 ($\overline{F0}$) of each target sentence was set to 117 Hz, corresponding to the averaged F0

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

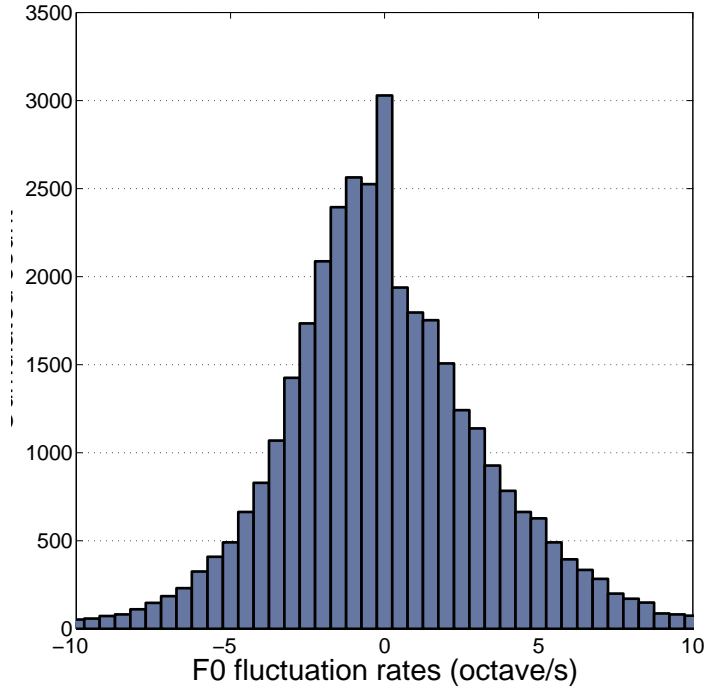


Figure IV.1 – F0 fluctuation rates of the sentences used in the present study.

across time and sentences of the corpus. On average over all sentences, the target F0 fluctuated across time between -9 and +9 semitones around $\overline{F0}$ (corresponding to an averaged spread of about 130 Hz) with an averaged standard deviation of 4.3 semitones (about 30 Hz). Figure IV.1 presents the F0 rates contained in the sentences used in the present study. They were calculated by passing every sentence into the PRAAT PSOLA speech analysis and resynthesis package (Boersma and Weenink, 2014) and extracting the F0 within each temporal window of 20 ms. The F0 rate was then derived from the derivative and expressed in octave per second. The averaged F0 rate was -0.37 oct/s with a 95% confidence interval (CI) of [-0.45 ; -0.30] with a standard deviation of 4.38 oct/s (CI = [4.33 ; 4.43]). To detail this distribution a little more closely, 74 % of the F0 rates were greater than 1 oct/s, 52 % were greater than 2 oct/s and 36% were greater than 3 oct/s.

Two F0 contours were tested: intonated and monotonized. Intonated targets had the same natural F0 contours as the original sentences, while monotonized target sentences had their F0 contours flattened over time at 117 Hz. Mean F0s and F0 contours of the sentences were manipulated using PRAAT PSOLA which calculated the F0 contour and resynthesized the

sentence with a specified F0 contour.

2.1.2 Maskers

The maskers were harmonic complexes with partials in random phase relationships, passed through a finite impulse response filter designed to match the averaged long-term excitation pattern of the sentence corpus (Fig. IV.2). For convenience, such a speech-shaped harmonic complex is hereafter referred to as “buzz”. Two buzz F0s were tested: 117 and 139 Hz, leading to a mean F0 difference with the target ($\overline{\Delta F0}$) of 0 and +3 semitones, respectively. In each experiment, buzzes were either kept unprocessed (monotonized buzzes) or natural speech F0 fluctuations were applied to their F0 contour (intonated buzzes). These fluctuations were extracted from two intonated sentences spaced by a 100 ms silence¹ and applied to the buzz with PRAAT PSOLA conserving its mean F0. In the case of intonated buzzes, the F0 contour was applied before the speech-shaped filter because the reverse order would have altered to speech-shaped spectral profile of the masker. Envelope-modulated buzzes were tested in experiment 2. They were obtained by multiplying a buzz with the envelope of two sentences spaced by a 100 ms silence. The envelope of the concatenated sentences was obtained by taking their modulus, following their peaks using a 10-ms smoothing window, and passing the signal through a second smoothing low-pass filter with a 40-Hz cutoff-frequency (Festen and Plomp, 1990). In the case of an envelope-modulated intonated buzz, both F0 contour and temporal envelope were extracted from the same pair of concatenated sentences producing a speech-like masker which was still unintelligible and clearly distinguishable from the speech target. The target sentence presented with an amplitude-modulated and/or intonated masker was always different from the concatenated sentences used to modulate or intonize the buzz. The buzzes were always longer than any of the target sentences to ensure that masking could occur throughout the entire target duration. All maskers had a duration of 3.8 s including a 10 ms cosine onset/offset whereas the targets all lasted less than 3.1 s, starting with 100 ms of silence.

2.2 Procedure

For each condition, one SRT was measured using a list of twelve target sentences and an adaptive method (Brand and Kollmeier, 2002). The twelve sentences were presented one after another against a different buzz. Listeners were instructed to type the words they heard on a computer keyboard after each presentation. The correct transcript was then displayed on a

1. This silence duration corresponded roughly to the average duration of the silences found between consecutive words in the speech material used.

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

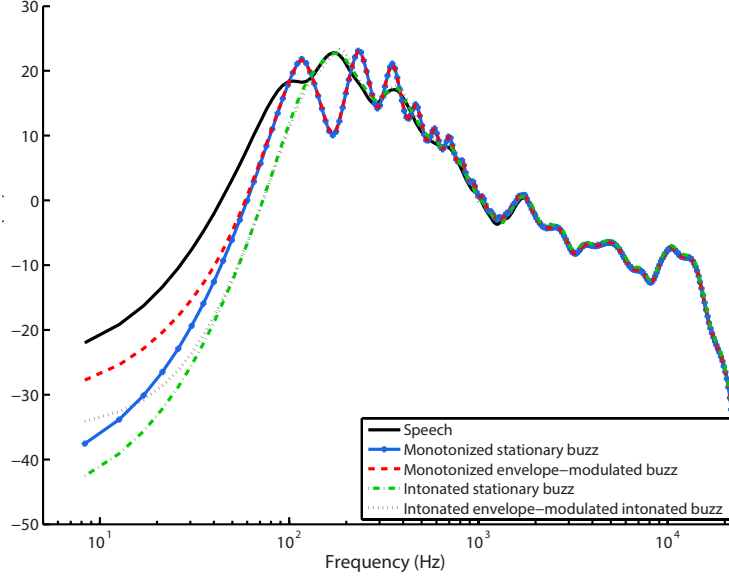


Figure IV.2 – (Color online) Long-term excitation patterns of the speech material (black line) and of each type of speech-shaped harmonic masker (buzz) with $F_0 = 117$ Hz.

monitor with the key words highlighted in capital letters. Listeners identified and self-reported the number of correct key words they perceived. For the first sentence of the list, listeners had the possibility to replay the stimuli, producing an increase in the target-to-masker ratio (TMR) of 3 dB, which was initially very low (-25 dB). Listeners were asked to attempt a transcript as soon as they believed that they could hear half of the words in the sentence. No replay was possible for the following sentences, for which the TMR was varied across sentences by modifying the target level while the masker level was kept constant at 70 dB SPL (unless the requested target level was too high so that it would have clipped the signal, in which very rare cases, the target level kept constant and the masker level was modified to achieve the same required TMR). For a given sentence, the TMR was increased if the score obtained at the previous sentence was greater than 2, it was decreased if the score was less than 2, and it remained unchanged if the score was 2. The sound level of the k^{th} ($2 \leq k \leq 12$) sentence of the list (L_k , expressed in dB SPL) was determined by Eq. IV.1 (Brand and Kollmeier, 2002):

$$L_k = L_{k-1} - 10 \times 1.41^{-i} \times (\text{SCORE}_{k-1} - 0.5) \quad (\text{IV.1})$$

where SCORE_{k-1} is the proportion of correct key words between 0 and 1 for the sentence $k - 1$ and i is the number of times $\text{SCORE}_{k-1} - 0.5$ changed sign since the beginning of the

sentence list. The SRT was taken as the mean TMR across the last eight sentences.

In each experiment, the SRT was measured for sixteen conditions presented in a pseudorandom order, which was rotated for successive listeners to counterbalance the effects of condition order and sentence lists, which were presented in a fixed sequence. Each target sentence was thus presented only once to every listener in the same order and, across a group of thirty-two listeners, two complete rotations of conditions were achieved. In each experiment, listeners began the session with two practice runs (diotic presentation of a naturally intonated speech target against a monotonized buzz), to get familiar with the task, followed by sixteen runs with breaks every four runs. Each experiment lasted between 1.5 and 2 hours depending on the duration of breaks and the rapidity of the listener.

2.3 Equipment

Signals were presented to listeners over Sennheiser HD 650 headphones in a doublewalled soundproof booth after having been digitally mixed, D/A converted, and amplified using a Lynx TWO sound card. A graphical interface was displayed on a computer screen outside the booth window. A keyboard and a computer mouse were inside the booth to interact with the interface and gather the transcripts.

2.4 Listeners

Thirty-two different listeners took part in each experiment. They were between 16 and 29 years old and self-reported normal hearing and French as their native language. None of them was familiar with the target sentences used during the test. Listeners were paid for their participation.

3 Experiment 1

3.1 Aim and design

Experiment 1 investigated whether listeners could independently benefit from the spatial separation between a target voice and a buzz and from their $\Delta F0$. The envelope of the buzzes was kept stationary in this experiment. The sources were virtually spatialized by convolving the signals with head-related impulse responses ([Gardner and Martin, 1994](#)). The target source was always simulated at 30° on the right of the listener while the buzz was simulated either

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

at the same location as the target (colocated condition) or at 30° on the left of the listener (separated condition).

3.2 Results

Figure IV.3 presents the mean SRTs across listeners measured in experiment 1 with a monotonized (left panel) or an intonated (right panel) buzz. The results of a repeated-measure analysis of variance (ANOVA) performed with four within-subject factors (target F0 contour \times buzz F0 contour \times $\overline{\Delta F0}$ \times spatial separation) are reported in Table IV.1. There was a main effect of spatial location, namely SRTs were lower for a buzz spatially separated from the target (triangles) compared to colocated conditions (circles). The size of this MR was substantial but depended on the buzz F0 contour, such that it was slightly larger with an intonated buzz (6.6 dB on average, right panel) than with a monotonized buzz (5.8 dB on average, left panel), resulting in a small interaction. There was a main effect of target F0 contour, SRTs were on average lower when the voice was naturally intonated (grey markers) rather than monotonized (black markers). There was a main effect of buzz F0 contour, SRTs were on average higher when the buzz was intonated (right panel) rather than monotonized (left panel), reflecting that variations in the buzz F0 were detrimental to target intelligibility. There was a main effect of $\overline{\Delta F0}$, reflecting that on average, target intelligibility was improved by separating the target and buzz $\overline{F0}$ s. There was also a rather complicated interaction between target F0 contour, buzz F0 contour and $\overline{\Delta F0}$ which was further interrogated by performing a simple effects analysis at each factorial combination of this interaction, averaged across spatial configurations. 1) As displayed in the left panel of Figure IV.4, $\overline{\Delta F0}$ benefit was significant only when sources were either both monotonized (Fig. IV.3, left panel, black markers) or both intonated (Fig. IV.3, right panel, grey markers), with an effect size of 3.5 dB and 1 dB, respectively [$F(1, 31) > 11$; $p < 0.01$]. This $\overline{\Delta F0}$ benefit was reduced as soon as one source was intonated. 2) Intonating the target F0 contour resulted in a significant benefit in every condition [$F(1, 31) > 9.07$; $p < 0.01$], except in presence of an intonated masker with $\overline{\Delta F0} = 0$. The largest benefit due to fluctuations in the target F0 was observed in the case of a monotonized buzz with $\overline{\Delta F0} = 0$ [$F(1, 31) = 146.4$; $p < 0.00001$]. 3) Natural fluctuations on the buzz F0 contour always induced an increase of SRTs ($F(1, 31) > 100$; $p < 0.0001$) except in the case of a monotonized target with $\overline{\Delta F0} = 0$.

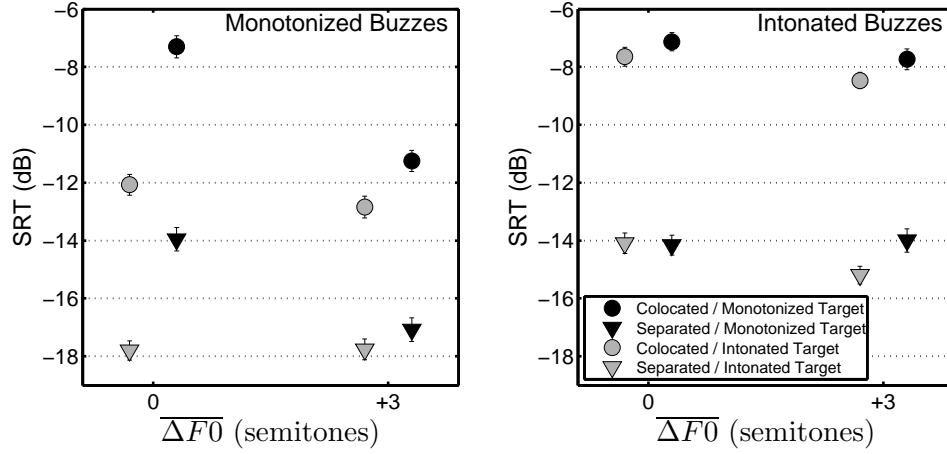


Figure IV.3 – Mean SRTs with standard errors across listeners measured in experiment 1 for monotonized and intonated buzzes as a function of the $\Delta F0$ between target and buzz. The target and buzz were either spatially colocated (circles) or separated. The target was either monotonized (black markers) or intonated (grey markers).

3.3 Discussion

3.3.1 Influence of the target F0 contour

The present study indicated that intonating the speech target resulted in lower SRTs, as it has been observed in previous studies. [Binns and Culling \(2007\)](#) and [Miller *et al.* \(2010\)](#) observed a detrimental effect on speech recognition of attenuating (or even reversing) the F0 fluctuations of the target in presence of a noise masker. This would reflect the specific role that the target F0 plays in the contribution of prosody to speech intelligibility ([Binns and Culling, 2007](#)). [Deroche *et al.* \(2014\)](#) also measured SRT for a voice masked by noise and found that the intonated voice led to SRTs at least 2 dB lower than the same voice monotonized. The size of this benefit seems in very good agreement with the present benefit of 1.7 dB (on average over all the other factors).

3.3.2 Benefit of spatial separation

The significant SRT difference between colocated and separated conditions observed in this experiment is in good agreement with previous studies conducted in anechoic environments ([Plomp, 1976](#); [Hawley *et al.*, 2004](#); [Beutelmann and Brand, 2006](#); [Jones and Litovsky, 2011](#)). In the presence of buzz maskers, listeners benefited from the binaural cues provided by the

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

Factors	F value	p value
Target F0 contour	138.409	0.000*
Buzz F0 contour	344.249	0.000*
$\overline{\Delta F0}$	62.275	0.000*
Location	1784.228	0.000*
Target F0 contour \times Buzz F0 contour	39.726	0.000*
Target F0 contour \times $\overline{\Delta F0}$	9.942	0.003*
Buzz F0 contour \times $\overline{\Delta F0}$	24.858	0.000*
Target F0 contour \times Location	3.086	0.089
Buzz F0 contour \times Location	4.977	0.033*
$\overline{\Delta F0} \times$ Location	3.349	0.077
Target F0 contour \times Buzz F0 contour \times $\overline{\Delta F0}$	39.936	0.000*
Target F0 contour \times Buzz F0 contour \times Location	1.583	0.217
Target F0 contour \times $\overline{\Delta F0} \times$ Location	0.609	0.441
Buzz F0 contour \times $\overline{\Delta F0} \times$ Location	0.604	0.443
Target F0 contour \times Buzz F0 contour \times $\overline{\Delta F0} \times$ Location	0.764	0.388

Table IV.1 – Results of the repeated-measures ANOVA with four within-subjects factors ($\overline{\Delta F0}$, target F0 contour, buzz F0 contour and spatial location) for experiment 1.

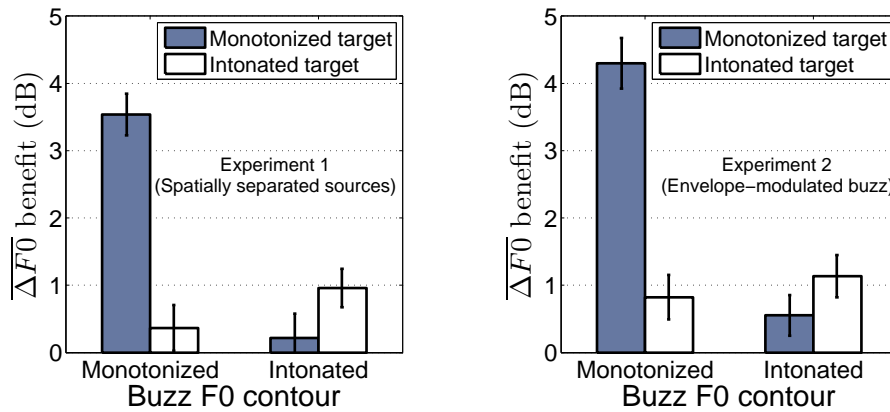


Figure IV.4 – Benefit due to $\overline{\Delta F0}$ (calculated as the difference between SRTs where $\overline{\Delta F0} = 3$ and 0 semitone) as a function of masker and target F0 contours for experiment 1 (left panel) and experiment 2 (right panel). The $\overline{\Delta F0}$ benefits have been averaged over spatial configurations in experiment 1, and over envelope-modulation conditions in experiment 2.

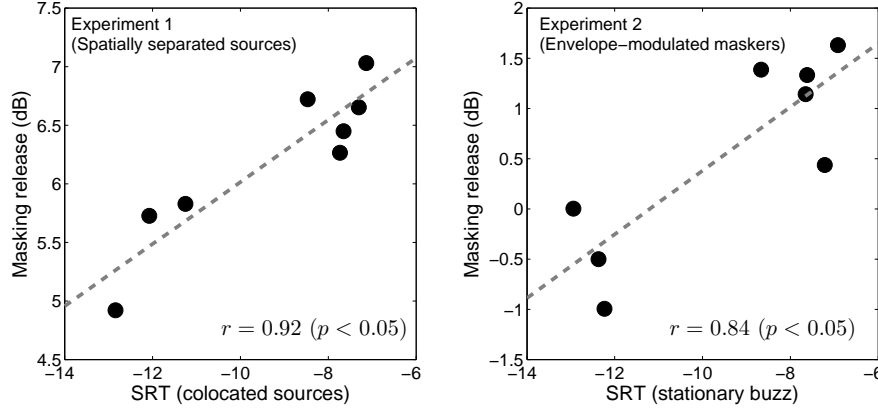


Figure IV.5 – Left panel: spatial masking release (SMR, calculated as the difference between SRTs in the colocated and separated conditions) measured in experiment 1 as a function of the SRT in the colocated condition. Right panel: benefit due to buzz envelope modulations (calculated as the difference between SRTs in the stationary and amplitude-modulated conditions) measured in experiment 2 as a function of the SRT in the stationary condition. The dashed line is the linear regression line.

spatial separation between target and masker, leading to a spatial masking release (SMR) of about 6 dB. Surprisingly, this effect was 0.8 dB larger for intonated than for monotonized buzzes. To further examine this interaction, one must note that SRTs were also lower with monotonized buzzes than with intonated buzzes, so the amount of SMR could have been limited by a ceiling effect in the case of monotonized buzzes. As an example, for both monotonized targets and buzzes (Fig. IV.3, left panel, black markers) the SMR was smaller when $\overline{\Delta F0} = 3$ semitones compared to when $\overline{\Delta F0} = 0$ (about 6 dB instead of about 7 dB). The same comparison can be made with the intonated targets and monotonized buzzes (Fig. IV.3, left panel, grey markers), for which the SMR was about 5 and 6 dB when target and masker $\overline{F0}$ were separated by 3 and 0 semitones, respectively. Left panel of figure IV.5 illustrates this ceiling effect by representing in each condition the SMR (the difference of SRTs between the separated and colocated conditions) as a function of the SRT in the colocated condition, which directly refers to the amount of masking there was to start with. The significant correlation between SMR and this SRT ($r = 0.91$; $p < 0.05$) supports the idea that the influence of buzz $F0$ contour could have simply been here a ceiling effect: when specific listening conditions allow a voice to be intelligible at a very adverse target-to-masker ratio, masking could have been already completely released, leaving no room for an additional cue to further improve intelligibility.

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

3.3.3 Benefit of F0 differences

The benefit due to $\overline{\Delta F0}$ depended on the F0 contour of the target and that of the buzz. In the case of a monotonized buzz and a monotonized target (Fig. IV.3, Left panel, black markers), speech intelligibility was improved when the $\overline{F0}$ of the buzz was shifted 3 semitones above that of the target (Fig. IV.4, left panel). This F0-segregation has been observed in earlier studies on double-vowels recognition (Culling and Darwin, 1993; de Cheveigné *et al.*, 1997) or in speech intelligibility studies using connected sentences (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Deroche and Culling, 2013). One account for this MR could be that listeners might be able to “glimpse” information from the target in between the resolved partials of the harmonic masker. This spectral glimpsing ability would not take place when target and masker share the same F0 since there is little target information in the masker spectral dips (Deroche *et al.*, 2014). This MR is also consistent with the harmonic cancellation theory: the auditory system would focus on the harmonic structure of the masker in order to cancel it (de Cheveigné *et al.*, 1995). When the buzz F0 is the same as that of the speech target, the cancellation process would distort the harmonic series belonging to the target. When F0s from the buzz and target differ, the masking partials could be cancelled with less alteration of the target partials, resulting in a larger benefit.

In the presence of an intonated buzz and a monotonized target (Fig. IV.3, right panel, black markers), shifting the buzz $\overline{F0}$ by 3 semitones did not produce any significant MR (Fig. IV.4, left panel). F0-segregation was strongly reduced compared to the monotonized buzz conditions. Because of the buzz F0 fluctuations, the spectral dips located in between resolved partials could have been blurred and listeners would not have been able to glimpse the target signal. The harmonicity could have also been disrupted and the cancellation of the masker partials could then have been then less effective. These two interpretations are presumably related to the temporal resolution of the auditory system. The reduction of F0-segregation observed in the presence of intonated buzzes might depend on the F0 rates contained in the F0 fluctuations. With very low F0 rates, F0-segregation would seem possible in the presence of an intonated masker and even if the rate limit for the harmonic process remains difficult to establish, it seems reasonable to admit that above 3.3 octave/s (which represents about 30% of the F0 rates contained in the speech material used in the present study, Fig. IV.1), the $\overline{\Delta F0}$ benefit is likely to be reduced (Deroche and Culling, 2011).

In the presence of an intonated target and a monotonized buzz (Fig. IV.3, left panel, grey markers, see also left panel of Fig. IV.4), variations in the target F0 created some instantaneous $\Delta F0$ s with the monotonized buzz. Those instantaneous differences could have been responsible

for the SRTs to be lowered by as much as a 3-semitone shift in F_0 between the corresponding monotonized sources (Fig. IV.3). Even though one should not forget that this comparison is confounded by the contribution of target prosody (section 3.3.1), this suggests that instantaneous ΔF_0 s could be sufficient to trigger F_0 -segregation even when no ΔF_0 exists on average. Furthermore, the fact that shifting the buzz $\overline{F_0}$ by 3 semitones had no significant effect for the intonated target might imply that the entire MR was already obtained with the instantaneous ΔF_0 s. Listeners were also able to glimpse target signal in between the resolved partials of the buzz. Since the excitation pattern of the intonated target did not present distinct spectral peaks and valleys (Fig. IV.2), the amount of target signal available in between the resolved partials of the buzz would only be marginally larger when the buzz F_0 is shifted 3 semitones above that of the target (Deroche *et al.*, 2014). Similarly, periodicity of the buzz was precisely defined when the buzz was monotonized, whether its F_0 was set at 117 Hz or 3 semitones above.

When both target and buzz F_0 s fluctuated (Fig. IV.3, right panel, grey markers), the benefit due to $\overline{F_0}$ separation was reduced but a significant $\overline{\Delta F_0}$ effect remained (Fig. IV.4, left panel). Although variations in the buzz F_0 are in general detrimental to the mechanisms underlying F_0 -segregation, it appears that natural variations would be apparently not large enough, fast enough, or both, to completely abolish F_0 -segregation.

Jackson and Moore (2013) conducted SRT measurements for $\overline{\Delta F_0}$ of 0, 2 and 4 semitones between a target speech and a masking background speech. Target and masker were either both intonated or both monotonized. They observed a $\overline{\Delta F_0}$ benefit when both target and masker were monotonized but not when they were intonated. In the present study, the same findings were observed. Separating the $\overline{F_0}$ of both monotonized target and masker by 3 semitones resulted in a significant masking release and this benefit was strongly reduced when both sources were intonated (even though this benefit was still significant in the present study). However, Jackson and Moore concluded that this lack of $\overline{\Delta F_0}$ benefit in the intonated case could be due to instantaneous ΔF_0 s which could have been sufficient to segregate the two intonated voices, even if they shared the same $\overline{F_0}$. The present study does not support this interpretation. By having tested all configurations of monotonized/intonated target/masker, it has been observed that the benefit due to instantaneous ΔF_0 s only occurred in the presence of an intonated target against a monotonized buzz. In the presence of a monotonized target against an intonated buzz, there were identical instantaneous ΔF_0 s, but they did not induce a significant benefit. In contrast with their conclusion, the present results indicate that instantaneous ΔF_0 s do not necessarily allow F_0 -segregation. The fact that SRTs do not decrease further by increasing $\overline{F_0}$ is no evidence that the full MR has been obtained, but it certainly implies a limiting factor. A comparison between SRTs for both intonated and both monotonized voices is not ideal to

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

evaluate whether the full MR was obtained because, as mentioned earlier, this comparison is confounded by the beneficial role of target F0 fluctuations for prosody (section 3.3.1). Their study used masking voices and it may well be that F0-segregation is then strongly sustained by streaming mechanisms being somehow more robust to fluctuations in the masker F0. Therefore, a safe conclusion at this point should be that for situations of speech-on-buzz masking, there is a very clear differential role for natural F0 fluctuations: those of the target voice are extremely beneficial whereas those of the buzz are extremely detrimental.

4 Experiment 2

4.1 Aim and design

Experiment 2 investigated whether listeners could independently benefit from temporal dips in the buzz envelope and the difference in F0 between target and buzz. The stimuli were presented diotically and the envelope of the buzz was either stationary or modulated (using a single voice).

4.2 Results

Figure IV.6 presents the mean SRTs across listeners measured in experiment 2, with a monotonized (left panel) or an intonated (right panel) buzz. The results of a repeated-measure ANOVA performed with four within-subject factors (target F0 contour \times buzz F0 contour \times $\overline{\Delta F0}$ \times envelope modulation of the buzz) are reported in Table IV.2. As already found in experiment 1, there was a main effect of target F0 contour, namely SRTs were on average lower when the target voice was naturally intonated rather than monotonized. There were a main effect of buzz F0 contour and a main effect of $\overline{\Delta F0}$: intelligibility was, on average, impaired when introducing variations in the buzz F0 contour, and improved when introducing a difference in mean F0 between target and buzz. These three factors interacted with each other, as in experiment 1. A simple effects analysis was performed on each combination factor of this interaction (target F0 contour \times masker F0 contour \times $\overline{\Delta F0}$), averaged over envelope modulation conditions (see Fig. IV.4, right panel). 1) The $\overline{\Delta F0}$ benefit was significant in all target and buzz contours conditions [$F(1, 31) > 6; p < 0.05$] except in the case of a monotonized target and intonated buzz. 2) Intonating the target significantly improved speech intelligibility in all configurations [$F(1, 31) > 9.5; p < 0.004$] except for a monotonized buzz with $\overline{\Delta F0} = 0$. 3) Intonating the buzz F0 had a significant effect on SRTs in all conditions [$F(1, 31) > 84; p <$

0.01], except for a monotonized target with $\overline{\Delta F0} = 0$.

There was a main effect of buzz envelope, namely SRTs were lower for a modulated buzz than a stationary buzz. This effect interacted with the masker F0 contour as illustrated in Figure IV.7. A simple effects analysis indicated that listeners significantly benefited from the envelope modulations of the buzz only for the intonated buzz [$F(1, 31) = 26.81$; $p < 0.01$] and not for the monotonized buzz. An interaction also occurred between the target F0 contour, the $\overline{\Delta F0}$ and the buzz envelope. A simple effects analysis was performed on each combination factor of this interaction, averaged over buzz F0 contours. A significant effect of the target F0 contour was observed in every factorial combinations, i.e. SRTs were significantly lowered by intonating the target [$F(1, 31) > 7.88$; $p < 0.008$]. The $\overline{\Delta F0}$ also led to a significant benefit in every factorial combinations [$F(1, 31) > 6.54$; $p < 0.01$]. Modulations in the temporal envelope of the buzz did not have a significant effect on SRTs in any condition, except in the presence of a monotonized target with $\overline{\Delta F0} = 0$ [$F(1, 31) = 10.25$; $p = 0.003$].

Factors	F value	p value
Target F0 contour	82.996	0.000*
Buzz F0 contour	291.687	0.000*
$\overline{\Delta F0}$	100.291	0.000*
Buzz envelope	7.681	0.009*
Target F0 contour \times Masker F0 contour	40.672	0.000*
Target F0 contour \times $\overline{\Delta F0}$	17.363	0.000*
Buzz F0 contour \times $\overline{\Delta F0}$	45.529	0.000*
Target F0 contour \times Buzz envelope	0.053	0.818
Buzz F0 contour \times Buzz envelope	19.632	0.000*
$\overline{\Delta F0} \times$ Buzz envelope	0.4	0.531
Target F0 contour \times Buzz F0 contour \times $\overline{\Delta F0}$	29.604	0.000*
Target F0 contour \times Buzz F0 contour \times Buzz envelope	0.181	0.673
Target F0 contour \times $\overline{\Delta F0} \times$ Masker envelope	4.790	0.036*
Buzz F0 contour \times $\overline{\Delta F0} \times$ Buzz envelope	0.429	0.517
Target F0 contour \times Buzz F0 contour \times $\overline{\Delta F0} \times$ Buzz envelope	1.076	0.307

Table IV.2 – Results of the repeated-measures ANOVA with four within-subjects factors ($\overline{\Delta F0}$, target F0 contour, buzz F0 contour and buzz envelope) for experiment 2.

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

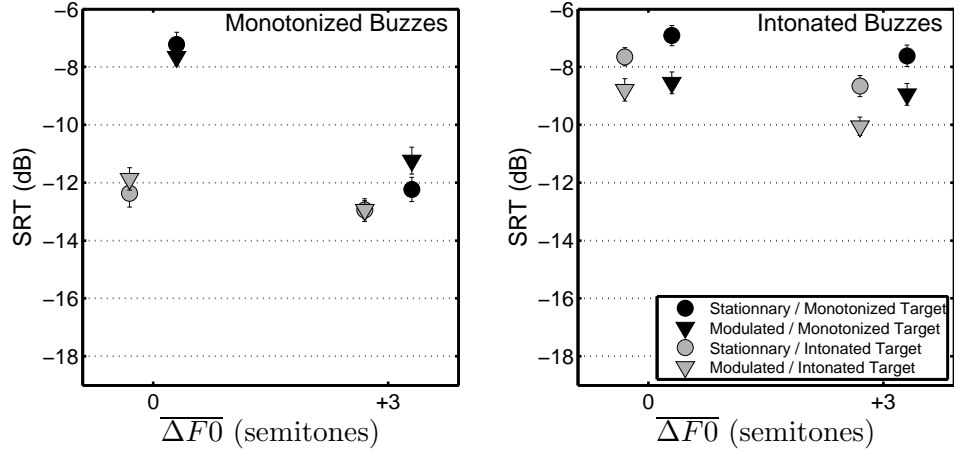


Figure IV.6 – Mean SRTs with standard errors across listeners measured in experiment 2 for monotonized and intonated buzzes as a function of the $\Delta F0$ between target and buzz. The amplitude of the envelope of the buzz was either stationary (circles) or one-voice modulated (triangles). The target was either monotonous (black markers) or intonated (grey markers).

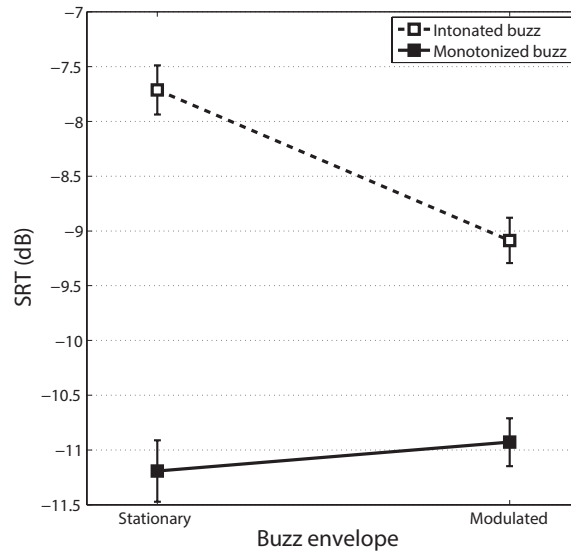


Figure IV.7 – Mean SRT measurements of experiment 2 as a function of the masker envelope and the masker contour (intonated maskers as white markers and monotonized maskers as black markers). Data of Fig. IV.6 have been averaged across target contour and $\Delta F0$.

4.3 Discussion

4.3.1 Comparison with experiment 1

Some results reported in experiment 1 were observed in experiment 2. The triple interaction between $\overline{\Delta F0}$, target F0 contour and buzz F0 contour occurred in both experiments (Fig. IV.4, right panel). The simple effect analysis performed on this interaction yielded the same findings, except that, in experiment 2, an additional significant $\overline{\Delta F0}$ benefit was revealed in the presence of an intonated target and a monotonized buzz. This may indicate that this $\overline{\Delta F0}$ benefit could have occurred in experiment 1, but that F0-segregation had no room to operate significantly because spatial unmasking already brought the SRT near -18 dB which could have been a floor of masking (ceiling of unmasking). Thus, the same $\overline{\Delta F0}$ benefit not observed in experiment 1 could have been limited by a potential ceiling effect of MR.

Despite this small difference between the two experiments, a key result was confirmed: the benefit due to $\overline{\Delta F0}$ separation was strongly reduced as soon as one of the two sources was intonated (Fig. IV.4). In the case of an intonated buzz, the spectral dips could have been too blurred by the buzz F0 fluctuations to listen through and/or the buzz partials fluctuated too rapidly in comparison to the temporal resolution of the auditory system to rely on harmonic cancellation. In the case of an intonated target, F0-segregation could have already operated thanks to instantaneous F0 differences with the monotonized buzz, providing the entire MR which could not be further increased by separating the $\overline{F0}$ s of the sources.

Like in experiment 1, speech intelligibility was significantly influenced by the target F0 contour. On average over envelope modulation conditions, SRTs were lower when the target voice was intonated rather than monotonized. This could illustrate again the benefit of prosody due to natural fluctuations of the target F0.

4.3.2 Benefit of envelope modulations of the buzz

On average, modulated-envelope buzzes led to lower SRTs compared to stationary buzzes. Listeners could have taken advantage of the momentary favorable TMRs in the temporal dips induced by the 1-voice amplitude modulation of the buzz envelope to glimpse the target signal. This MR was observed only with the intonated buzzes and did not exceed 2 dB unlike in previous studies which reported larger benefits: 4 dB in Hawley *et al.* (2004), 12 dB in Beutelmann *et al.* (2010) and 4 dB in Collin and Lavandier (2013). All these previous MRs were measured in the presence of a 1-voice envelope-modulated speech-shaped noise (SSN) colocated with the target in front of the listener in anechoic conditions (or with a very high direct-to-reverberant

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

ratio). Collin and Lavandier (2013) and the present study used the same speech material (Raake and Katz, 2006) and the same method to create the envelope-modulated maskers (Festen and Plomp, 1990). The difference in MR between these two studies should be related to the type of masker used: buzz in the present study and SSN in Collin and Lavandier (2013). Both maskers present a speech-shaped spectrum, but the resolved partials of the harmonic masker creates spectral dips (especially for monotonized buzzes) resulting in less masking than SSN (see Fig. IV.2). The modulated maskers used by Beutelmann *et al.* (2010) might have presented different properties regarding the envelope modulations in the masking noise which depend on the speech material used to modulate the masker envelope. Differences of language might induce differences in modulation parameters such as modulation depth or rate which can be responsible for intelligibility differences across the aforementioned studies. It should also be noted that Hawley *et al.* (2004) used frozen maskers. A non-negligible additional MR could have occurred due to anticipation effects on the occurrence of the masker temporal dips (Collin and Lavandier, 2013).

In the presence of temporal fluctuations in the masker envelope, both masking release (temporal dip listening) and modulation masking could occur (Kwon and Turner, 2001). Modulation masking refers to the fact that speech envelope modulations are less detected in the presence of similar masking envelope modulations. Modulation rates of both target and masker envelopes can overlap in the modulation domain, leading to a masking effect. The wide range of benefits induced by envelope-modulation in the masker reported in the literature (Festen and Plomp, 1990; Hawley *et al.*, 2004; Beutelmann *et al.*, 2010; Collin and Lavandier, 2013) might be due to different amounts of modulation masking in these different studies, which could have reduced the expected masking release associated with temporal dip listening. By introducing speech-like modulations into the buzz envelopes, the transmission of the target articulation could have been impaired by modulation masking, which would at least partly explained the limited differences between SRTs for stationary and envelope-modulated buzzes in this study. However, there is no obvious reason why modulation masking would be different amounts for the monotonized or intonated buzzes. Consequently, the role of modulation masking is not an obvious explanation for the observed interaction between the buzz envelope modulations and F0 contour.

The benefit due to the envelope modulations of the buzz depended on the buzz F0 contour. As illustrated in Figure IV.7, SRTs decreased when imposing modulations to the temporal envelope of the buzz only when this buzz was intonated, providing a benefit of 1.4 dB (white squares), but not when the buzz was monotonized (black squares). In experiment 1, the spatial masking release was also dependent on the buzz F0 contour, but this could have been caused by a ceiling effect since SRTs were already very low because of the F0-related benefit allowed

against monotonized buzzes but not with intonated buzzes. In experiment 2, the hypothesis of another ceiling effect could make sense considering that the temporal dip in a masker waveform is a perfect example where masking is at floor. To examine the hypothesis of a ceiling effect in experiment 2, the benefit due to the modulations of the buzz envelope was plotted in the right panel of Fig. IV.5 as a function of the SRT in the stationary condition. As in the case of spatial unmasking, a positive and significant correlation ($r = 0.84$; $p < 0.05$) indicates that the lower the SRTs, the lower the benefit of envelope modulations. This could explain why envelope modulations of the buzz resulted in a larger benefit when buzzes were intonated rather than monotonized. However, in experiment 1, spatial unmasking could bring SRTs down to -18 dB, so it is rather surprising that a ceiling effect could have occurred in experiment 2 for SRTs around -12 dB. Unless ceiling effects could occur at different levels of the auditory pathway and are then not directly comparable, these relatively high SRTs would suggest that the interpretation of a ceiling effect is not as obvious as in experiment 1.

Two other interpretations can be suggested to account for the F0-contour dependence of the benefit due to modulations in the buzz envelope. First, listeners might benefit from temporal fluctuations of the buzz envelope only in conditions where F0-segregation cues are not available, as if both benefits were mutually exclusive. Deroche *et al.* (2014) found a strong dependency of F0-segregation on the masker F0, using stationary monotonized buzzes and pseudo-stationary monotonized babbles (made of 400 simultaneous voices) and linked the $\overline{\Delta F0}$ benefit to the size of the spectral dips between the resolved partials of the maskers. Since this dependency had not been observed in conditions of a single monotonized masking voice (Bird and Darwin, 1998; Assmann, 1999), Deroche *et al.* (2014) concluded that the ability to glimpse in-between resolved masker partials may not hold for temporally-fluctuating maskers. The present data support this interpretation, suggesting that listeners could not glimpse spectrally and temporally at the same time. It should also be noted that other mechanisms, for instance based on the periodicity in the masker fine structure, would also suffer from temporal modulations in the masker envelope because of temporal interruptions in the F0 contour of the masker. However, this interpretation suffers from one counter-example. A benefit of modulating the buzz envelope in the case of a monotonized target against a monotonized buzz sharing the same $\overline{F0}$ ($\overline{\Delta F0} = 0$) would have been expected, since spectral glimpsing was then strongly reduced in this case.

Another interpretation could be that intonations in the buzz F0 contour could have been helpful cues for listeners to anticipate envelope modulations in this masker. Collin and Lavandier (2013) showed that the benefit due to modulations of the masker envelope is larger if listeners can predict the occurrences of the temporal gaps in the masker envelope compared to when these occurrences are unpredictable. This “predictability” benefit was about 1.5 dB

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

for 1-voice modulated SSNs, which correspond to the difference of modulated-envelope benefit observed between the intonated and monotonized buzzes here. Collin and Lavandier (2013) compared unfrozen and frozen modulated noises to highlight this benefit. Because here F0 contours and temporal envelopes applied to buzzes were extracted from the same sentences, listeners might have been able to rely on F0 patterns in order to anticipate the presence of temporal gaps in the masker envelope. It should be kept in mind that temporal “dip listening” has been previously observed without any F0 cue (Gustafsson and Arlinger, 1994; Hawley *et al.*, 2004; Beutelmann *et al.*, 2010; Collin and Lavandier, 2013), so this would not necessarily explain why there was no benefit at all for monotonized buzzes, simply that there could be more benefit with intonated buzzes. Here, if the MR due to anticipation had further improved an existing dip listening benefit associated with the fluctuations in the broadband temporal envelope of the masker, then some MR should have occurred in the presence of monotonized buzzes. One account for this discrepancy could be that more masking energy is attenuated in the temporal gaps of a SSN rather than in those of a monotonized buzz which presents peaks and valleys in the resolved region of the excitation pattern. A buzz being less masking than a SSN to start with, there is less benefit to get from the temporal “dip listening” mechanism. With this interpretation, the benefit observed with intonated buzzes could be an anticipation effect based on the F0 fluctuations which could have provided some prediction cues to the listeners regarding the time locations of the temporal gaps.

Finally, a triple interaction between the buzz envelope, the target F0 contour and the $\overline{\Delta F0}$ was observed. The benefits due to F0 fluctuations in the target and to the $\overline{\Delta F0}$ were significant in all conditions. However, the modulations in the buzz envelope resulted in a significant benefit only when the target was monotonized and with $\overline{\Delta F0} = 0$. This interaction would support the interpretation described above: listeners only relied on temporal gaps in the buzz envelope when spectral glimpsing was not effective. It is worth noting that this effect was observed only when averaging the data over the buzz F0 contour conditions, meaning that the benefit associated with envelope modulations was large enough in the presence of intonated buzzes to remain significant on average over the buzz F0 contour conditions.

5 Conclusions

SRT measurements were conducted to test whether masking releases resulting from spatial separation of sources, modulations of the masker envelope and F0 differences between competing sources added up or not when these cues were available at the same time. Spatial separation

and F0 differences were tested in experiment 1, while masker envelope modulations and F0 differences were tested in experiment 2. Two F0 contours (monotonized or intonated) were tested for both target and masker in each experiment. Several key results were highlighted:

1. The benefit due to a difference of mean F0 between target and masker was greatly reduced in the presence of a harmonic masker with a naturally fluctuating F0.
2. In addition to providing prosodic cues that facilitate intelligibility regardless of masking, naturally F0-fluctuating targets led to much lower SRTs compared with monotized targets, even though they shared the same mean F0 with the masker. This suggests that listeners could rely on instantaneous F0 differences to unmask speech from a harmonic masker. This masking release was not further improved by increasing the mean F0 of the masker by 3 semitones.
3. The masking release due to spatial separation of sources was of similar magnitude as in previous studies and was almost constant across tested conditions, suggesting that the benefits from spatial cues and F0-segregation cues could linearly add up. A small interaction with the masker F0 contour was observed and was attributed to a ceiling effect caused by low SRTs in the colocated conditions which could hardly have decreased further.
4. Listeners only benefited from the temporal dips in the masker envelope in the presence of an intonated masker, suggesting that either envelope modulations were detrimental to F0-segregation or the variations of the masker F0 contour constituted a helpful cue to anticipate the presence of temporal dips in the masker envelope. The hypothesis of another ceiling effect is not to be excluded, even if stationary SRTs were on a relatively higher range, leaving a priori “enough room” for temporal dip listening to operate.

Chapter IV. Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location

General conclusions and perspectives

Through the three studies presented in this thesis, experimental and modelling work has been conducted in order to extend the binaural model developed by [Lavandier and Culling \(2010\)](#) towards a speech intelligibility model accounting for speech maskers. The original model was handling only the cases of a near-field target masked by noises.

The influence of the room on target speech has first been implemented by combining the original model ([Lavandier and Culling, 2010](#)) and a U/D approach which attributes useful and detrimental roles to the early and late reflections, respectively, present in the impulse response of the target source. By separating these two types of reflection in the impulse response, the model kept its original framework but is now based on computations of U/D ratios rather than signal-to-noise ratios (SNRs). This early/late separation involves new parameters in the model which have been investigated to determine their influence on speech intelligibility predictions. This new version of the model was then validated on experimental data from previous studies of the literature which involved spatial unmasking and temporal smearing of speech ([Lavandier and Culling, 2008](#); [Rennies *et al.*, 2011](#)) as well as binaural de-reverberation ([Lavandier and Culling, 2008](#)). As a result, the proposed model yielded good prediction performances on each experimental dataset. But, this level of performance was obtained by adjusting, for each room, the temporal parameters allowing for the separation of early and late reflections, leading to a room-dependent model. A room-independent version of the model (i.e. with fixed parameters) was proposed and tested on another external dataset ([van Wijngaarden and Drullman, 2008](#)). Comparing the model predictions to the experimental measurements led to think that room-independence and prediction accuracy could not be achieved together with the current version of the model. Either the implementation needs to be revised in order to take into account the parameter adjustments, or the U/D approach presents a fundamental limit while applying

the model to different rooms. Despite this limitation, the U/D approach provides a unified interpretation of temporal smearing, spatial unmasking and binaural de-reverberation. The late reflections are regarded as an additional masking source which is induced by the room. As any other masker, it is processed by the binaural system which unmask the reverberant target: this is binaural de-reverberation. The room-dependence limitation highlighted in this study could be further investigated by considering either another approach to account for reverberant targets, or by implementing the early/late separation differently, i.e. find out on which (room-independent) parameters this separation process is based.

The case of target and masker with large differences in the spectrum was then investigated in a second study. The original model of [Lavandier and Culling \(2010\)](#) could predict infinitely high or low intelligibility if either the target or the masking source were completely filtered in a given frequency band (the SNR would be infinitely high or low). Speech reception thresholds were measured in four speech intelligibility experiments. Target or noise was either high-pass filtered or low-pass filtered, creating different SNRs in the rejected band by varying the attenuation level of the filter. The measured thresholds showed that the SNR range influencing speech intelligibility was larger than the range proposed by the SII standard [-15 dB; +15 dB] and was frequency-dependent. The results also suggested that filtering the masker had more impact on speech intelligibility than filtering the target, meaning that intelligibility would not increase linearly along with SNR in a given frequency band. A simple model based on a sum of weighted SNRs has been proposed to describe the experimental data. At the term of this PhD work, this model is limited to a monaural version only tested on the data which was used to define its parameters. When these parameters are implemented into another framework ([Lavandier et al., 2012](#)), it leads to poorer predictions than without. External datasets are needed to more precisely define how SNR influences speech intelligibility in each frequency band, and how these parameters could be implemented into binaural models.

The third study investigated three auditory mechanisms (spatial unmasking, temporal dip listening and F0 segregation) and the way they operate when they could be solicited simultaneously, as it can be the case in real-life situations. Harmonic maskers with F0 contours and envelopes extracted from speech sentences were used to get closer to speech maskers. This study first confirmed previous results on the individual benefits provided by each mechanism in isolation. Spatially separating the target from the masking source improved speech intelligibility compared to when they were colocated. Envelope modulations in the maskers helped the listeners to better understand the target speech compared to stationary maskers. Listeners benefited from the F0 separation between target and masker compared to when they shared

the same F0. The findings of this study also showed that the benefits from each mechanism linearly added up when considering spatial unmasking and F0 segregation, while an interaction was highlighted between dip listening and F0 segregation. Another main result revealed that F0 segregation was strongly disrupted in the presence of a harmonic masker with a fluctuating F0. In addition, the results showed that listeners could rely on instantaneous F0 differences induced between a naturally-intonated speech and a harmonic masker with a steady F0. In order to predict F0 segregation and its interaction with dip listening, additional signal analysis steps would be needed in the current models (spectro-temporal analysis, periodicity detection, harmonicity detection for instance).

All this work allowed to extend the scientific knowledge concerning the auditory mechanisms involved while listening speech masked by disturbing noise or harmonic sources in rooms. So far, each study (except chapter IV) proposed an extended version of the model developed by Lavandier and Culling (2010) with parameters allowing to take into account more complex situations. Each version yields good prediction performances in isolation and on datasets involved in the model definition. After having tested and validated each version of the model on more data, the next step would be to unify these different versions (Collin and Lavandier, 2013; Leclère *et al.*, 2015a, and the model proposed in chapter III) into a unique model able to predict a wide panel of situations related to speech in noise perception. In the longer term, a speech intelligibility model for cocktail-party situations in rooms should also include the perceptual mechanisms higher located in the auditory pathway and related to informational masking.

Résumé en français

1 Introduction

À l'heure actuelle, la communication occupe une place importante dans notre société. La compréhension de la parole est particulièrement essentielle pour les interactions sociales, la sécurité ou l'accessibilité aux bâtiments et transports publics. L'intelligibilité de la parole peut-être fortement réduite en présence de sources de bruits ou de parole dans des espaces clos, ce qui peut avoir des répercussions sur la santé dues à une augmentation de l'effort d'écoute, de la gêne ou encore de la fatigue.

Pour améliorer l'intelligibilité de la parole de ces situations bruyantes, il est nécessaire de comprendre quels sont les différents mécanismes auditifs et cognitifs qui participent à la compréhension de la parole masquée par du bruit ou d'autres sources de parole environnantes. Modéliser ces mécanismes, c'est à dire être capable de prédire l'intelligibilité de la parole perçue à partir de paramètres physiques pour une situation donnée, permettrait de concevoir des bâtiments destinés à accueillir de nombreuses personnes et ainsi fournir de bonnes conditions d'écoute aux usagers. Ces modèles peuvent également être utilisés pour développer des algorithmes à implémenter dans les prothèses auditives dans le but de restaurer les différentes déficiences des mécanismes auditifs dont souffrent les malentendants.

Ce travail de thèse se limite aux mécanismes auditifs chez les normo-entendants. Des études de la littérature dans les domaines de l'audition, de l'acoustique des salles et de la psychoacoustique ont permis d'identifier les facteurs qui influencent l'intelligibilité de la parole. Ces résultats ont débouché sur plusieurs modèles visant à prédire l'intelligibilité perçue à partir de paramètres physiques. Les premiers modèles proposés étaient limités à des situations basiques telles qu'une parole cible masquée par une seule source de bruit en écoute monaurale (écoute avec une seule oreille). Comme beaucoup d'études, ces travaux de thèse visent à étendre ces modèles vers des situations de communication plus complexes et réalistes (comme une situation dite de « cocktail-party » où d'autres conversations environnantes perturbent la source cible de

parole). Pour pouvoir prédire l'intelligibilité dans de telles situations, il est nécessaire d'étudier si les mécanismes auditifs identifiés dans le cas de la parole dans le bruit sont toujours valables en présence de voix concurrentes, qui présentent d'autres propriétés acoustiques que les bruits stationnaires (modulation d'enveloppe, fréquence fondamentale, intonation). Il est également nécessaire de comprendre si ces nouvelles propriétés acoustiques peuvent déclencher d'autres mécanismes auditifs de démasquage que ceux déjà identifiés en présence de bruits stationnaires.

Cette thèse expose dans un premier temps l'état de l'art concernant la connaissance scientifique sur l'intelligibilité de la parole dans le bruit en décrivant les mécanismes auditifs mis en jeu ainsi que les différents modèles prédictifs développés à l'heure actuelle. Le second chapitre traite de l'influence de la salle sur l'intelligibilité de la parole en étendant le modèle de [Lavandier and Culling \(2010\)](#) au cas d'une cible réverbérée. Des sources cible et masquante ayant des spectres différents font l'objet du troisième chapitre où une nouvelle version du modèle est présentée pour prendre en compte ces différences spectrales. Il est ensuite essentiel de déterminer comment ces mécanismes se comportent lorsqu'ils sont sollicités en même temps : opèrent-ils indépendamment ? ou bien interagissent-ils ? Cette question abordée dans le quatrième chapitre est cruciale pour permettre d'implémenter au mieux ces mécanismes dans un modèle unique. Enfin, les conclusions générales, les principaux résultats de ces différentes études ainsi que des perspectives de recherche potentielles sont résumés à la fin de ce manuscrit.

Cette thèse est composée de quatre chapitres : un état de l'art de la connaissance scientifique sur l'intelligibilité de la parole dans le bruit suivi de trois études de recherche réalisées dans le cadre de ce travail de thèse. Le présent résumé expose en français et de manière synthétique et vulgarisée le travail présenté dans le texte principal.

2 Intelligibilité de la parole dans le bruit

2.1 Définitions

Pour étudier l'intelligibilité de la parole, de nombreuses études (dont ce travail de thèse) considèrent une situation de communication où un locuteur parfait (la cible) s'adresse à un auditeur parfait à travers un canal de transmission (l'air et la salle). La présence de sources masquantes environnantes perturbe la transmission acoustique du locuteur cible vers l'auditeur qui ne comprendra alors qu'une fraction des mots émis par le locuteur (voir Fig. [1.1](#)). Cette fraction, souvent exprimée en pourcents, quantifie **l'intelligibilité de la parole cible**. Plus la parole est intelligible, plus le pourcentage de mots compris (par rapport aux nombres de mots

émis) sera élevé. En considérant un locuteur et un auditeur parfaits, l'intelligibilité est donc influencée par les sources masquantes et le canal de transmission.

Dans des situations de cocktail-party, la source cible est perturbée par d'autres voix concurrentes (Fig. 1.2). Sur un plan purement acoustique, ce type de source masquante présente d'autres propriétés que des sources de bruit. En effet, une voix possède une structure harmonique avec une fréquence fondamentale (F_0) et des formants (F_1 , F_2 ,...) qui varient au cours du temps (intonation) ainsi qu'une enveloppe modulée en amplitude (le niveau acoustique n'est pas constant pendant un discours, il y a des pauses entre les mots etc...). Il est donc important de comprendre en quoi ces propriétés acoustiques pourraient influencer la perception de la parole cible et ainsi agir sur l'intelligibilité.

Pour étudier les facteurs qui agissent sur l'intelligibilité de la parole, il faut faire appel à l'expérimentation. Un sujet humain écoute des phrases, des mots ou même des voyelles en présence de sources masquantes dont les propriétés acoustiques sont contrôlées. Il est ensuite demandé au sujet d'exprimer (oralement ou par écrit) la phrase, le mot ou la voyelle qu'il a entendu. En répétant l'opération plusieurs fois, il est donc possible de déterminer la proportion de mots compris par le locuteur par rapport au nombre de mots émis pour une condition de masquage donnée. En testant une autre source masquante (qui a été contrôlée différemment), il est ainsi possible de comparer les performances de l'auditeur et de déterminer comment a varié l'intelligibilité de la parole cible, c'est à dire dans quelle mesure la modification de la source masquante a influencé l'intelligibilité. Une amélioration de l'intelligibilité entre deux conditions s'appelle « *bénéfice* » ou « *démasquage* ».

Un des facteurs qui influence l'intelligibilité de la parole est le **rapport signal/bruit** (« *signal-to-noise ratio* » en anglais, SNR), qui représente la différence entre les niveaux acoustiques de la cible et du bruit (voir Eq. 1.1). Plus le niveau de la cible est fort et celui du bruit faible, plus le SNR sera grand (inversement pour une cible faible et un bruit fort). Un fort SNR conduit à une intelligibilité plus grande qu'un faible SNR : la parole est mieux comprise lorsque le locuteur parle fort et que les sources masquantes ont un faible niveau.

Pour mesurer l'intelligibilité, il est possible d'évaluer le pourcentage de mots compris par l'auditeur (par rapport au nombre de mots émis) ou bien de mesurer le SNR pour lequel l'auditeur reconnaît 50% des mots. Cette dernière mesure s'appelle le **seuil de reception de la parole** (« *speech reception threshold* » en anglais, SRT) et est généralement obtenu grâce à une méthode adaptative (Levitt, 1971; Brand and Kollmeier, 2002). Dans la procédure expérimentale, le SNR est varié d'une écoute à l'autre en fonction de la réponse du sujet à l'écoute précédente. Si le sujet a reconnu plus de 50% des mots à une écoute donnée, la tâche était plutôt simple. Elle va donc être rendue plus compliquée à l'écoute suivante en diminuant le

SNR. Si le sujet a reconnu moins de 50% des mots, la tâche était plutôt difficile et elle sera rendue plus simple en augmentant le SNR à l'écoute suivante. Ces variations de SNR en fonction des réponses de l'auditeur conduisent à déterminer pour quel SNR l'auditeur reconnaitra 50% des mots, c'est le SRT. Entre deux conditions, une baisse du SRT correspond à une augmentation de l'intelligibilité car la même proportion de mots a été comprise mais pour un SNR plus bas (c'est à dire dans des conditions d'écoute plus difficiles).

2.2 Mécanismes et modèles

Même en présence de sources masquantes, le système auditif peut faire appel à des mécanismes de démasquage pour améliorer la perception de la parole, conduisant ainsi à un bénéfice d'intelligibilité. Ces mécanismes de démasquage et leurs bénéfices associés dépendent des caractéristiques des signaux acoustiques émanant de chaque source reçus aux oreilles. Les mécanismes abordés dans cette thèse sont décrits dans les sections suivantes.

2.2.1 Démasquage spatial

Lorsque la source cible et la source masquante sont situées à différentes positions dans l'espace, l'intelligibilité est améliorée grâce au fait d'écouter avec deux oreilles plutôt qu'une (écoute binaurale en opposition à l'écoute monaurale). La parole est ainsi mieux démasquée lorsqu'elle est spatialement séparée de la source de bruit : ce mécanisme est appelé « **démasquage spatial** ». Les propriétés d'absorption de la tête impliquent qu'une source placée dans le plan horizontal (avec un azimuth non nul) produira un niveau acoustique différent sur chaque oreille : il y a une **différence interaurale de niveau** (« interaural level difference » en anglais, ILD). De plus, l'onde acoustique provenant de cette même source met un temps plus long pour atteindre l'oreille située la plus loin de la tête, créant ainsi une **différence interaurale de temps** (« interaural time difference » en anglais, ITD). Ces différences interaurales sont appelées « **indices binauraux** » et permettent la localisation dans le plan horizontal ([Moore, 2003](#)) mais également le démasquage spatial qui exploite les différences d'ILD et d'ITD générées par les sources cibles et masquantes pour réduire l'influence du masqueur sur la cible. Cela signifie que le démasquage spatial est efficace lorsque les sources sont séparées spatialement : si les deux sources occupent la même position, elles génèrent les mêmes ILDs et ITDs entre les oreilles et n'engendrent aucun démasquage.

2.2.2 Écoute dans les creux de modulation

L'intelligibilité est améliorée lorsque la source masquante présente une enveloppe qui fluctue dans le temps comparé à une enveloppe stationnaire. L'auditeur peut ainsi profiter des instants où le masqueur présente un faible niveau pour capter au mieux la source cible : comme si la cible était perçue « à travers » les creux de modulation du masqueur. Plus les creux de modulation sont larges, plus une grande quantité de signal cible est disponible et plus le gain d'intelligibilité sera élevé. Cette largeur des creux de modulation est déterminée par la fréquence et la profondeur de modulation. Dans le cas d'un bruit modulé par de la voix humaine, seuls les creux de modulation d'une seule voix sont bénéfiques pour l'auditeur. Lorsque plusieurs voix modulent simultanément le bruit, cela réduit les creux de modulation et donc le bénéfice d'intelligibilité ([Festen and Plomp, 1990](#); [Collin and Lavandier, 2013](#)).

2.2.3 Ségrégation par F0

Lorsque la source masquante présente une structure harmonique (voix ou masqueur harmonique), le système auditif peut mieux démasquer la parole cible lorsqu'elle présente une fréquence fondamentale (F0) différente de celle du masqueur ([Brokx and Nootboom, 1982](#)). Un masqueur harmonique présente des trous spectraux au travers desquels le système auditif peut capter le signal de la cible. Le principe est le même que pour l'écoute dans les creux de modulation mais dans le domaine fréquentiel. Le signal cible capté à travers ces trous spectraux est plus important lorsque les F0s de la cible et du masqueur diffèrent. À l'inverse, lorsque les sources présentent la même F0, leurs harmoniques se superposent et le signal cible disponible à travers les trous spectraux du masqueur est très résiduel. D'autres interprétations avancent également que le système auditif peut soit s'appuyer sur le caractère harmonique (fréquences situées à intervalles réguliers dans le spectre) du masqueur pour mieux l'identifier et ainsi l'ignorer, le supprimer ; soit s'appuyer sur le caractère harmonique de la cible, et ainsi l'augmenter, l'améliorer. Ces deux mécanismes sont également plus performants lorsque les F0s de la cible et du masqueur diffèrent puisque, les harmoniques de chacune des cibles ne se superposent pas et le processus d'annulation (pour le masqueur) et/ou augmentation (pour la cible) n'affecte que la source concernée. Dans le cas contraire, si les deux sources partagent la même F0, toute la série harmonique est confondue et le processus d'annulation/augmentation affecte les deux sources, ce qui n'engendre aucune amélioration de l'intelligibilité de la source cible.

2.3 Effet de la réverbération

Dans une salle, une seule onde acoustique parvient directement aux oreilles de l'auditeur depuis la source : **l'onde directe**. Les autres ondes acoustiques se réfléchissent sur les parois selon les propriétés d'absorption des matériaux et parviennent de manière détournée aux oreilles de l'auditeur qui reçoit donc une multiplicité de signaux acoustiques provenant des différentes réflexions de la salle. Ces réflexions sont des copies du son direct ayant subi des modifications temporelles et spectrales car 1) elles parcourent une distance plus longue avant d'arriver aux oreilles de l'auditeur et 2) les propriétés absorbantes des parois modifient le spectre de l'onde incidente. L'ensemble des réflexions forment ce que l'on appelle la "réverbération" et a un impact direct sur l'intelligibilité de la parole perçue dans une salle. Par la présence de réflexions, l'auditeur reçoit donc le même message de parole de manière décalée dans le temps. Selon ce décalage temporel, la même information peut parvenir de manière très tardive à l'auditeur alors que d'autres (nouveaux) messages ont été délivrés par l'onde directe, créant ainsi une superposition temporelle de signaux acoustiques qui se masquent entre eux à chaque instant et détériorent l'intelligibilité de la source cible ([Lochner and Burger, 1964](#); [Bradley and Bistafa, 2002](#); [Lavandier and Culling, 2008](#); [Rennies *et al.*, 2011](#)). Cet effet est appelé **étalement temporel** et agit même en absence de source masquante. Les propriétés d'absorption des parois modifient également le spectre de la source acoustique et agissent comme un filtre. Ce filtrage par la salle (aussi appelé coloration) va donc pondérer le contenu spectral de la source ce qui a un impact direct sur l'intelligibilité selon les fréquences concernées.

Puisque la réverbération modifie les propriétés temporelles et spectrales de l'onde acoustique émise par la source, les indices acoustiques sur lesquels se basent les mécanismes de démasquage peuvent également être modifiés et donc conduire à un démasquage différent qu'en conditions anéchoïques. Plus de précisions sur la façon dont les mécanismes de démasquage sont influencés par la réverbération sont reportées dans le texte principal de ce manuscrit.

2.4 Modèles d'intelligibilité

L'intelligibilité de la parole dans le bruit a fait l'objet de nombreuses investigations pour la quantifier, la mesurer ou la prédire. Des indices d'intelligibilité ont été développés très tôt (AI, SII, STI) et visent à quantifier si les conditions d'écoute sont optimales ou non sur une échelle de 0 à 1. Dans une approche plus récente, beaucoup d'études visent à modéliser les mécanismes de démasquage et leurs interactions afin de décrire et prédire des résultats expérimentaux à partir d'un nombre limité de paramètres. Devant la multiplicité des modèles existants et les différentes approches sur lesquelles ils reposent, le lecteur est convié à se reporter au chapitre

d'état de l'art de ce manuscrit pour une description détaillée de nombreux modèles. Seul le modèle de Lavandier et Culling sera brièvement décrit ici, puisqu'il est au coeur de ces travaux de thèse.

Lavandier and Culling (2010) ont développé un premier modèle d'intelligibilité qui a par la suite été amélioré par Jelfs *et al.* (2011); Lavandier *et al.* (2012)². Ce modèle permet de prédire le gain d'intelligibilité entre deux conditions (une différence de SRT). Le mécanisme de démasquage spatial est implémenté dans ce modèle binaural, ce qui permet de tenir compte de la position des sources cibles et masquantes dans les prédictions (une meilleure intelligibilité sera prédite dans le cas de sources séparées par rapport à des sources co-localisées). À partir des réponses impulsionnelles binaurales (« binaural room impulse response » en anglais, BRIR) de la source cible et du masqueur, le modèle va pouvoir extraire les indices binauraux (ILDs, ITDs) de chaque source et ainsi calculer le démasquage spatial. Les prédictions de ce modèle sont limitées au cas des sources cibles en champ proche perturbées par des bruits stationnaires dans les salles.

3 Modèle de prédiction pour les sources distantes

Lorsque la distance entre la cible et l'auditeur augmente dans une salle, l'influence de la réverbération devient de moins en moins négligeable et crée de l'étalement temporel, même en absence de sources masquantes.

Cet étalement temporel a déjà été étudié dans la littérature et implémenté dans divers indicateurs d'acoustique des salles (STI, U/D, voir chapitre I). Mais ces indicateurs sont monauraux, c'est à dire qu'il ne prennent en compte les signaux reçus que sur une seule oreille. Ils négligent donc le bénéfice apporté par l'écoute binaurale qui permet à l'auditeur de démasquer une source cible de parole parmi des sources de bruit spatialement séparées de la cible. À l'inverse, le modèle de Lavandier and Culling (2010) permet de prédire le démasquage spatial mais est limité au cas d'une source cible située en champ proche car l'effet de l'étalement temporel n'est pas pris en compte. Ce premier travail cherche à concilier ces précédents travaux en étendant le modèle de Lavandier and Culling (2010) au cas des sources cibles réverbérées avec une approche utile/nuisible (useful/detrimental en anglais, U/D) pour prédire l'effet de l'étalement temporel sur l'intelligibilité.

L'approche U/D a été introduite par Lochner and Burger (1964) puis reprise par Bradley

2. Ces améliorations concernent surtout une différence d'implémentation plutôt qu'une révision majeur du concept d'origine proposé par Lavandier and Culling (2010). Ces versions sont fondamentalement équivalentes et renvoient à la même structure de modélisation.

(1986) et Bradley *et al.* (1999) pour développer un indicateur d'intelligibilité dans les salles. Cette approche considère les premières réflexions comme étant utiles car elles sont perceptivement intégrées au son direct et renforcent ainsi le message transmis. Au contraire, les réflexions tardives sont responsables de l'effet nuisible sur l'intelligibilité. Pour rendre compte de l'influence de la salle sur l'intelligibilité de la parole, l'approche U/D compare donc l'énergie des réflexions précoces à l'énergie des réflexions tardives pour une réponse impulsionnelle donnée. Plus la source cible est éloignée de l'auditeur, plus la réponse impulsionnelle sera riche en réflexions tardives et plus le rapport U/D sera faible. À l'inverse, une réponse impulsionnelle associée à une source proche présentera un grand rapport U/D et donc un faible effet de l'étalement temporel.

Cette approche a donc été implémentée dans la dernière version du modèle de Lavandier and Culling (2010) qui prend les réponses impulsionnelles de la cible et des différents masqueurs en entrée de modèle (Lavandier *et al.*, 2012). La réponse impulsionnelle de la cible est tout d'abord séparée en une réponse « précoce » et une réponse « tardive ». Puisque les réflexions tardives sont considérées comme nuisibles à l'intelligibilité, la réponse « tardive » est regroupée avec les réponses des masqueurs (comme s'il s'agissait d'un masqueur additionnel) pour ainsi former la composante « nuisible ». La réponse « précoce » constitue la composante « utile ». Le modèle original (Lavandier *et al.*, 2012) est ensuite appliqué aux composantes « utile » et « nuisible » de la même manière qu'il était appliqué aux composantes « cible » et « masqueur » dans la version précédente.

Cette séparation entre les réflexions utiles et nuisibles implique l'introduction de nouveaux paramètres dans le modèle : la limite temporelle à partir de laquelle les réflexions deviennent nuisibles (appelée « early/late limit » en anglais, ELL) ainsi que la transition entre utile et nuisible (une réflexion devient-elle subitement nuisible ? Existe-t-il une durée de « transition » entre les réflexions utiles et nuisibles ?). Ces paramètres ont été étudiés en testant plusieurs ELLs et durées de transition de manière systématique pour déterminer l'impact de chacun de ces paramètres sur les prédictions du modèle.

Cette nouvelle version du modèle a donc été testée sur des données expérimentales publiées dans la littérature qui impliquent à la fois du démasquage spatial et de l'étalement temporel (Lavandier and Culling, 2008; Rennies *et al.*, 2011). Les prédictions étaient répétées en faisant varier de manière indépendante les paramètres temporels liés à la séparation utile/nuisible.

Les résultats de ce test systématique indiquent que l'approche U/D combinée au modèle de démasquage spatial de Lavandier *et al.* (2012) permettait de prédire avec précision l'effet de

l'étalement temporel et du démasquage spatial sur l'intelligibilité de la parole cible. Les deux paramètres temporels induits par la séparation des réflexions utiles/nuisibles influencent les prédictions. Lorsque l'ELL est choisie trop longue, des réflexions sensées nuire à l'intelligibilité sont considérées comme utiles et l'étalement temporel est ainsi sous-estimé par rapport à celui mesuré chez les sujets.

Cependant, les bonnes performances du modèle à prédire à la fois les effets d'étalement temporel et de démasquage spatial étaient obtenues en ajustant les paramètres du modèle pour chaque étude/salle. Cela signifie que les paramètres donnant les meilleures prédictions pour les données d'une étude étaient différents sur les données d'une autre étude : le modèle était dépendant de la salle et l'approche proposée dans cette étude ne peut, à l'heure actuelle, donc pas être généralisée à n'importe quelle salle.

Ce résultat soulève deux questions : la séparation utile/nuisible ne serait-elle pas basée sur un facteur acoustique non pris en compte dans cette étude et qui serait, lui, indépendant de la salle ? Cette dépendance de la salle est-elle fondamentalement liée à l'approche U/D ?

Malgré cette limitation, le modèle proposé permet une interprétation unifiée du démasquage spatial et de l'étalement temporel. En considérant les réflexions tardives comme nuisibles, elle sont interprétées par le modèle comme une nouvelle source de bruit qui perturbe la compréhension de la cible. Plus cette nouvelle source est énergétique, c'est à dire plus la cible est réverbérée, plus l'étalement temporel aura un effet nuisible sur l'intelligibilité. Comme en présence de n'importe quel masqueur, le système auditif tente de démasquer spatialement la cible du masqueur grâce à l'écoute binaurale, créant ainsi un léger bénéfice comparé à l'écoute monaurale dans des environnements réverbérants ([Lavandier and Culling, 2008](#), « binaural squelch »).

4 Intelligibilité pour des sources de spectres différents

Il a été précédemment vu que l'intelligibilité d'une source cible était fortement liée au SNR. Lorsque le SNR augmente, l'intelligibilité est améliorée et lorsque le SNR diminue, l'intelligibilité est détériorée. Bien que le SNR puisse prendre des valeurs infinies, cela n'est pas le cas de l'intelligibilité qui est forcément bornée. Lorsque l'auditeur a compris tous les mots émis par son locuteur, augmenter le SNR n'améliorera pas l'intelligibilité qui est déjà maximale. Même chose, lorsque l'auditeur n'a compris aucun mot, baisser le SNR ne conduira pas à une intelligibilité plus faible. Il doit donc exister des SNRs limites, au-delà desquels l'intelligibilité

cesse d'être influencée. De plus, le modèle concerné par ces travaux de thèse ([Lavandier and Culling, 2010](#)) ne prévoit pas d'étape pour limiter le SNR, ce qui implique que, en l'état, de très forts (ou faibles) SNRs peuvent être calculés par le modèle et ainsi conduire à une intelligibilité infiniment grande (ou petite), ce qui n'a pas de sens d'un point de vue perceptif.

Le « speech intelligibility index » (SII, voir chapitre [I](#)) inclut de telles limites en admettant que les SNRs qui influencent l'intelligibilité sont compris dans l'intervalle $[-15 \text{ dB}; +15 \text{ dB}]$. Cela signifie que, dans une bande fréquentielle donnée, l'intelligibilité ne serait plus modifiée par des variations de SNR au dessus de $+15 \text{ dB}$ ou en dessous de -15 dB .

Ces hypothèses ont été testées dans quatre expériences en mesurant des SRTs avec une source cible et de bruit en écoute diotique. Chaque source (cible/masqueur) était filtré en passe-bas ou passe-haut avec une fréquence de coupure à 1400 Hz . Différentes atténuations étaient testées en bande coupée, permettant ainsi de faire varier le SNR dans cette bande. D'après notre hypothèse d'expérience, l'intelligibilité devrait dans un premier temps (pour les faibles niveaux d'atténuation) être influencée par les variations de SNR, puis rester constante à partir d'une certaine valeur d'atténuation.

Les résultats des quatre expériences confirment l'hypothèse avancée : le SRT varie linéairement en fonction de l'atténuation du filtre pour de faibles atténuations et une asymptote est atteinte pour les atténuations plus élevées. Cette asymptote a été atteinte pour différents niveaux d'atténuations selon l'expérience. Dans le cas de la cible filtrée en passe-haut, la limite de -15 dB proposée par le SII a été confirmée tandis que l'intelligibilité continuait d'être détériorée pour de plus fortes atténuations dans le cas de la cible filtrée en passe-bas, jusqu'à stagner à partir d'une atténuation de 37 dB . Dans le cas du masqueur filtré, l'intelligibilité était améliorée jusqu'à des niveaux d'atténuation de 43 dB (passe-haut) et 36 dB (passe-bas). Il a également été observé que filtrer la cible ou le masqueur n'impactait pas l'intelligibilité de la même manière. En effet, lorsque le SRT évolue linéairement avec l'atténuation (avant l'atténuation limite donc), la pente diffère selon la source filtrée, indiquant ainsi que, dans une bande de fréquence, un incrément de SNR résultait en un plus grand accroissement du SRT dans le cas du masqueur filtré (quand le SNR est positif) comparé au cas de la cible filtrée (quand le SNR est négatif).

Un modèle simple a été proposé pour décrire les données collectées dans ces quatre expériences en implémentant les paramètres nécessaires pour prédire 1) les asymptotes obtenues pour de forts niveaux d'atténuations, 2) la différence de pentes obtenue à faibles atténuations entre les cas où la cible est filtrée et où le masqueur est filtré. Le modèle donne une bonne

description des SRTs mesurés et fût ainsi testé sur d'autres données externes à l'étude (Lavandier *et al.*, 2012). Cette fois, les prédictions du modèle étaient plus erronées, ce qui soulève l'importance de tester sa validité et son pouvoir prédictif sur d'autres données qui n'ont pas servi à définir ses paramètres.

Les résultats de cette étude mettent également en question les paramètres utilisés dans le SII, qui est une norme largement utilisée dans beaucoup de modèles d'intelligibilité. Les limitations de cette étude peuvent être repoussées par de futures études qui auraient pour but d'augmenter la résolution fréquentielle ou encore d'étudier les mêmes problématiques en écoute binaurale.

5 Intelligibilité de la parole en présence de masqueurs harmoniques

Lorsque les sources masquante sont des voix, elles possèdent des propriétés acoustiques différentes de celles des bruits stationnaires et sur lesquelles des mécanismes de démasquage se basent. En effet, en plus du démasquage spatial, les modulations d'enveloppe permettent l'écoute dans les creux de modulation et les différences de fréquence fondamentale (F_0) entre cible et masqueur permettent la ségrégation par F_0 .

Ces mécanismes ont été brièvement décrits précédemment et ont fait l'objet de nombreuses études de la littérature. Toutefois, ils étaient bien souvent étudiés indépendamment les uns des autres pour acquérir le maximum de connaissances sur un mécanisme isolé. Or, dans une situation réaliste (cocktail-party par exemple), il est fort probable que ces mécanismes soient sollicités en même temps puisque l'on peut très bien être en présence d'une voix concurrente qui, est spatialement séparée de la cible, possède des modulations d'enveloppe et possède une F_0 différente de la cible. Pour pouvoir prédire l'intelligibilité de ce genre de situation, il est nécessaire de comprendre si les mécanismes fournissent des bénéfices qui peuvent s'additionner ou bien s'il existe des interactions potentielles lorsque ces mécanismes sont sollicités au même moment.

Cette étude présente deux expériences conçues dans le but de déterminer si la ségrégation par F_0 interagit avec le démasquage spatial (expérience 1) ou avec l'écoute dans les creux de modulation (expérience 2) en mesurant des SRTs en présence d'une voix naturelle (avec des intonations) ou monotone (pas de variation de F_0) perturbée par huit masqueurs harmoniques différents. Dans l'expérience 1, le masqueur était intonisé (sa F_0 variait au cours du temps)

ou monotone (F0 fixe au cours du temps); sa F0 était supérieure à celle de la cible de 0 ou 3 demi-tons; sa position spatiale était soit la même que la cible (co-localisée) soit différente. Dans l'expérience 2, seule la variation de la position était remplacée par une variation d'enveloppe (stationnaire ou modulée en amplitude). Les modulations d'enveloppe et de F0s étaient extraites de phrases, ce qui rendait ces types de masqueurs très proches des voix en matière de propriétés acoustiques sans être intelligible.

Les résultats de l'expérience 1 ont dans un premier temps mis en évidence le bénéfice de la ségrégation par F0 et du démasquage spatial seuls. Séparer spatialement les sources cibles et masquantes et augmenter la F0 du masqueur 3 demi-tons au-dessus de celle de la cible a conduit à un gain d'intelligibilité. De plus, il a été observé que la ségrégation par F0 était très réduite lorsque le masqueur présentait des variations de F0 au cours du temps. Les trous spectraux au travers desquels l'auditeur pouvait capter le signal cible « bougent » au cours du temps et le système auditif semble incapable de suivre ce mouvement qui conduit à une forte réduction du bénéfice apporté par la ségrégation par F0. L'intonation de la cible a amélioré l'intelligibilité de manière générale et de façon plus prononcée en présence d'un masqueur monotone à la même F0 qui, dans ce cas, présentait des différences instantanées de F0 avec la cible. Lorsque la ségrégation par F0 et le démasquage spatial étaient sollicités en même temps, leurs bénéfices étaient cumulés linéairement, indiquant ainsi que l'auditeur pouvait bénéficier de ces deux mécanismes de manière indépendante.

L'expérience 2 a confirmé des résultats déjà observés dans l'expérience 1 concernant la réduction de ségrégation par F0 en présence d'un masqueur intonisé et concernant le bénéfice dû aux F0s instantanées créées entre une cible intonisée et un masqueur monotone partageant la même F0. De plus, cette deuxième étude a permis de mettre en évidence une interaction entre la ségrégation par F0 et l'écoute dans les creux de modulation. En effet, les modulations d'enveloppe du masqueur n'ont amélioré l'intelligibilité que dans le cas où le masqueur était intonisé. Plusieurs interprétations ont été proposées pour ce résultat, dont le fait que l'auditeur ne puisse pas à la fois écouter dans les trous spectraux et dans les trous temporels. L'écoute dans les creux de modulation serait bénéfique seulement lorsque la ségrégation par F0 n'agit pas (en présence de masqueurs intonisés).

6 Conclusions et perspectives

Au cours des trois études présentées dans cette thèse, des travaux expérimentaux et de modélisation ont été réalisés dans le but de développer le modèle de [Lavandier and Culling \(2010\)](#) vers un modèle d'intelligibilité pouvant prédire l'influence de voix masquantes. Le modèle original était limité aux cas d'une cible en champ proche masquée par des bruits.

L'influence de la salle a été étudié et modélisé en combinant le modèle original ([Lavandier and Culling, 2010](#)) avec une approche U/D. Cette nouvelle implémentation a permis de prédire avec précision différents résultats expérimentaux faisant intervenir plusieurs effets perceptifs (démasquage spatial, étalement temporel). Toutefois, ce modèle nécessitait d'être ajusté en fonction de la salle pour obtenir de bonnes performances de prédiction. Soit l'approche U/D présente une limite fondamentale empêchant d'appliquer le modèle à plusieurs salles avec des paramètres fixes, soit l'implémentation de cette approche doit être étudiée et révisée plus en détail. Malgré cette limitation, ce modèle permet d'interpréter l'étalement temporel en termes de démasquage spatial.

Le cas d'une cible et d'un masqueur avec de larges différences dans le spectre ont été étudiés dans une seconde étude. Le modèle original de [Lavandier and Culling \(2010\)](#) pouvait prédire une intelligibilité infiniment grande ou petite si la cible ou le masqueur était complètement filtré dans une certaine bande de fréquence. Un travail expérimental a permis de déterminer les SNRs limites à partir desquels l'intelligibilité de la cible cessait d'être influencée. Les résultats indiquent que les intervalles de SNR ayant un impact sur l'intelligibilité sont plus larges que ceux déjà proposés dans la littérature (SII, [ANSI S3.5, 1997](#)) et dépendent de la bande de fréquence considérée. Un modèle monaural simple a été proposé pour décrire ces données. Lorsque ses paramètres sont appliqués à un autre modèle plus complexe ([Lavandier et al., 2012](#)) pour prédire d'autres données extérieures à l'étude, les prédictions ne sont plus aussi précises. Cela montre l'importance d'autres jeux de données pour mieux définir comment le SNR influence l'intelligibilité dans chaque bande de fréquence, et comment ces paramètres peuvent être implémentés dans des modèles binauraux.

La troisième étude s'intéressait à trois mécanismes de démasquage (démasquage spatial, écoute dans les creux de modulation et ségrégation par F0) et à la façon dont ils opèrent lorsqu'ils sont sollicités simultanément, comme cela pourrait être le cas dans la réalité. Des masqueurs harmoniques avec des intonations et des enveloppes extraites de signaux de parole

ont été utilisés dans le but de se rapprocher des voix concurrentes. La première expérience a confirmé des résultats connus dans la littérature concernant ces mécanismes de démasquage lorsqu'ils opèrent seuls. De plus, les résultats ont également montré que le démasquage spatial et la ségrégation par F0 pouvaient agir indépendamment : les bénéfices fournis par chaque mécanisme s'additionnent linéairement. En revanche, la ségrégation par F0 et l'écoute dans les creux de modulations interagissaient montrant ainsi que chaque mécanisme ne pouvait être bénéfique que dans des conditions particulières. Pour pouvoir prédire ces différents effets, de nouvelles analyses de signaux devraient être implémentées dans le modèle (analyse spectro-temporelle, détection de F0, détection d'harmonicité par exemple).

Tous ces travaux ont permis d'étendre la connaissance scientifique concernant les mécanismes auditifs impliqués lors de l'écoute de parole masquée par du bruit ou des sources harmoniques dans les salles. Jusqu'à présent, chaque étude (mis à part la dernière), a débouché sur une version étendue du modèle original développé par [Lavandier and Culling \(2010\)](#) avec de nouveaux paramètres permettant de prendre en compte des situations de communication plus complexes. Chaque version donnait de bonnes performances de prédiction de manière isolée et sur des données qui ont servi au développement du modèle. Après avoir testé et validé chaque version sur plus de données externes, la prochaine étape consisterait à unifier ces différentes versions ([Collin and Lavandier, 2013](#); [Leclère et al., 2015a](#), et le modèle proposé dans la deuxième étude) en un modèle unique, capable de prédire un large panel de situations impliquant la perception de la parole dans le bruit. À plus long terme, un modèle d'intelligibilité pour des situations de cocktail-party devrait également inclure les mécanismes perceptifs situés plus haut sur un plan cognitif et liés au masquage informationnel.

Bibliography

- Abed, A.-E. H. M. and Cain, G. D. (1984). “The host windowing technique for FIR digital filter design”, IEEE Transactions on Acoustics Speech and Signal Processing **ASSP.32**, 683–694.
- Allen, J. B., Berkley, D. A., and Blauert, J. (1977). “Multimicrophone signal-processing technique to remove room reverberation from speech signals”, J. Acoust. Soc. Am. **62**, 912–915.
- ANSI S3.5 (1969). “Methods for the calculation of the articulation index”, American National Standards Institute, New York .
- ANSI S3.5 (1997). “Methods for calculation of the speech intelligibility index”, American National Standards Institute, New York .
- Arweiler, I. and Buchholz, J. M. (2011). “The influence of spectral characteristics of early reflections on speech intelligibility”, J. Acoust. Soc. Am. **130**, 996–1005.
- Arweiler, I., Buchholz, J. M., and Dau, T. (2013). “The influence of masker type on early reflection processing and speech intelligibility (L)”, J. Acoust. Soc. Am. **133**, 13–16.
- Bacon, S. P. and Grantham, D. W. (1989). “Modulation masking: Effects of modulation frequency, depth, and phase”, J. Acoust. Soc. Am. **85**, 2575–2580.
- Beranek, L. L. (1947). “The design of speech communication systems”, Proceedings of the Institute of Radio Engineers **35**, 880–890.
- Beutelmann, R. and Brand, T. (2006). “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners”, J. Acoust. Soc. Am. **120**, 331–342.

- Beutelmann, R., Brand, T., and Kollmeier, B. (2010). “Revision, extension, and evaluation of a binaural speech intelligibility model”, J. Acoust. Soc. Am. **127**, 2479–2497.
- Binns, C. and Culling, J. F. (2007). “The role of fundamental frequency contours in the perception of speech against interfering speech”, J. Acoust. Soc. Am. **122**, 1765–1776.
- Bird, J. and Darwin, C. J. (1998). *Effects of a difference in fundamental frequency in separating two sentences* (Whurr, London).
- Boersma, P. and Weenink, D. (2014). “Praat: doing phonetics by computer [computer program]. version 5.3.85, retrieved 19 september 2014 from <http://www.praat.org/>”, .
- Bolt, R. H. and MacDonald, A. D. (1949). “Theory of speech masking by reverberation”, J. Acoust. Soc. Am. **21**, 577–580.
- Bradley, J. S. (1986). “Predictors of speech intelligibility in rooms”, J. Acoust. Soc. Am. **80**, 837–845.
- Bradley, J. S. and Bistafa, S. R. (2002). “Relating speech intelligibility to useful-to-detrimental sound ratios (L)”, J. Acoust. Soc. Am. **112**, 27–29.
- Bradley, J. S., Reich, R. D., and Norcross, S. G. (1999). “On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility”, J. Acoust. Soc. Am. **106**, 1820–1828.
- Bradley, J. S., Sato, H., and Picard, M. (2003). “On the importance of early reflections for speech in rooms”, J. Acoust. Soc. Am. **113**, 3233–3244.
- Brand, T. and Kollmeier, B. (2002). “Efficient adaptive procedures for thresholds and concurrent slope estimates for psychophysics and speech intelligibility tests”, J. Acoust. Soc. Am. **111**, 2801–2810.
- Brandewie, E. and Zahorik, P. (2010). “Prior listening in rooms improves speech intelligibility”, J. Acoust. Soc. Am. **128**, 291–299.
- Brokx, J. P. L. and Nooteboom, S. G. (1982). “Intonation in the perceptual separation of simultaneous voices”, Journal of Phonetics **10**, 23–26.
- Bronkhorst, A. W. (2000). “The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions”, Acta Acust. united Ac. **86**, 117–128.

- Bronkhorst, A. W. and Plomp, R. (1988). “The effect of head-induced interaural time and level differences on speech intelligibility in noise”, *J. Acoust. Soc. Am.* **83**, 1508–1516.
- Bronkhorst, A. W. and Plomp, R. (1992). “Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing”, *J. Acoust. Soc. Am.* **92**, 3132–3139.
- Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). “Informational and energetic masking effects in the perception of multiple simultaneous talkers”, *J. Acoust. Soc. Am.* **110**, 2527–2538.
- Carhart, R., Tillman, T. W., and Johnson, K. R. (1966). “Binaural masking of speech by periodically modulated noise”, *J. Acoust. Soc. Am.* **39**, 1037–1050.
- Cherry, E. C. (1953). “Some experiments on the recognition of speech with one and with two ears”, *J. Acoust. Soc. Am.* **25**, 975.
- Collin, B. and Lavandier, M. (2013). “Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers”, *J. Acoust. Soc. Am.* **134**, 1146–1159.
- Cooke, M. (2003). “Glimpsing speech”, *J. Phonetics* **31**, 579.
- Culling, J. F. and Darwin, C. J. (1993). “Perceptual separation of simultaneous vowels: Within and across formant grouping by f_0 ”, *J. Acoust. Soc. Am.* **93**, 3454–3467.
- Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (2004). “The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources”, *J. Acoust. Soc. Am.* **116**, 1057–1065.
- Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (2005). “Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources”, *J. Acoust. Soc. Am.* **118**, 552.
- Culling, J. F., Hodder, K. I., and Toh, C. Y. (2003). “Effects of reverberation on perceptual segregation of competing voices”, *J. Acoust. Soc. Am.* **114**, 2871–2876.
- Culling, J. F., Summerfield, Q., and Marshall, D. H. (1994). “Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels”, *Speech Communication* **14**, 71–95.

- Dau, T., Verhey, J., and Kohlrausch, A. (1999). “Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers”, *J. Acoust. Soc. Am.* **106**, 2752–2760.
- de Cheveigné, A. (1993). “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing”, *J. Acoust. Soc. Am.* **93**, 3271–3290.
- de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997). “Concurrent vowel identification. i. effects of relative amplitude and f0 difference”, *J. Acoust. Soc. Am.* **101**, 2839–2847.
- de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). “Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement”, *J. Acoust. Soc. Am.* **97**, 3736–3748.
- Deroche, M. L. D. and Culling, J. F. (2011). “Voice segregation by difference in fundamental frequency : Evidence for harmonic cancellation”, *J. Acoust. Soc. Am.* **130**, 2855–2865.
- Deroche, M. L. D. and Culling, J. F. (2013). “Voice segregation by difference in fundamental frequency: Effect of masker type”, *J. Acoust. Soc. Am.* **134**, EL465.
- Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (2014). “Roles of the target and masker fundamental frequencies in voice segregation”, *J. Acoust. Soc. Am.* **136**, 1225–1236.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2002). “Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing”, *J. Acoust. Soc. Am.* **111**, 2897–2907.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2005). “Recognition of filtered words in noise at higher-than-normal levels: Decreases in scores with and without increases in masking”, *J. Acoust. Soc. Am.* **118**, 923–933.
- Dunn, H. K. and White, S. D. (1940). “Statistical measurements on conversational speech”, *J. Acoust. Soc. Am.* **11**, 278–288.
- Durlach, N. I. (1963). ““Equalization and Cancellation Theory of Binaural Masking-Level Differences””, *J. Acoust. Soc. Am.* **35**, 1206–1218.

- Durlach, N. I. (1972). “Binaural signal detection: Equalization and cancellation theory”, in *Foundations of Modern Auditory Theory*, edited by J. Tobias, volume II, 371–462 (Academic, New York).
- Durlach, N. I., Mason, C. R., Jr., G. K., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). “Note on informational masking (I)”, *J. Acoust. Soc. Am.* **113**, 2984.
- Egan, J. P. and Hake, H. W. (1950). “On the masking pattern of a simple auditory stimulus”, *J. Acoust. Soc. Am.* **22**, 622–630.
- Ewert, S. D. and Dau, T. (2000). “Characterizing frequency selectivity for envelope fluctuations”, *J. Acoust. Soc. Am.* **108**, 1181–1196.
- Festen, J. M. (1993). “Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice”, *J. Acoust. Soc. Am.* **94**, 1295–1300.
- Festen, J. M. and Plomp, R. (1990). “Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing”, *J. Acoust. Soc. Am.* **88**, 1725–1736.
- French, N. R. and Steinberg, J. C. (1947). “Factors governing the intelligibility of speech sounds”, *J. Acoust. Soc. Am.* **19**, 90–119.
- Gardner, B. and Martin, K. (1994). “HRTF measurements of a KEMAR dummy-head microphone”, Technical Report, MIT Media Lab Perceptual Computing.
- Glasberg, B. R. and Moore, B. C. J. (1990). “Derivation of auditory filter shapes from notched-noise data”, *Hearing Res.* **47**, 103–138.
- Gustafsson, H. Å. and Arlinger, S. D. (1994). “Masking of speech by amplitude-modulated noise”, *J. Acoust. Soc. Am.* **95**, 518–529.
- Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer”, *J. Acoust. Soc. Am.* **115**, 833–843.
- Houtgast, T. and Steeneken, H. J. M. (1973). “The modulation transfer function in room acoustics as a predictor of speech intelligibility”, *Acustica* **28**, 66–73.

- Houtgast, T. and Steeneken, H. J. M. (1985). “A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria”, *J. Acoust. Soc. Am.* **77**, 1069–1077.
- Houtgast, T., Steeneken, H. J. M., and Plomp, R. (1980). “Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics”, *Acustica* **46**, 60–72.
- Howard-Jones, P. A. and Rosen, S. (1993). “The perception of speech in fluctuating noise”, *Acustica* **78**, 258–272.
- ISO 3382 (1997). “Acoustics - measurement of the reverberation time of rooms with reference to other acoustical parameters”, International Organization for Standardization, Geneva .
- Jackson, H. M. and Moore, B. C. J. (2013). “Contribution of temporal fine structure information and fundamental frequency separation to intelligibility in a competing-speaker paradigm”, *J. Acoust. Soc. Am.* **133**, 2421–2430.
- Jelfs, S., Culling, J. F., and Lavandier, M. (2011). “Revision and validation of a binaural model for speech intelligibility in noise”, *Hearing Res.* **275**, 96–104.
- Jones, G. L. and Litovsky, R. Y. (2011). “A cocktail party model of spatial release from masking by both noise and speech interferers”, *J. Acoust. Soc. Am.* **130**, 1463–1474.
- Jørgensen, S. and Dau, T. (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing”, *J. Acoust. Soc. Am.* **130**, 1475–1487.
- Jørgensen, S., Ewert, S. D., and Dau, T. (2013). “A multi-resolution envelope-power based model for speech intelligibility”, *J. Acoust. Soc. Am.* **134**, 436–446.
- Koenig, W. (1950). “Subjective effects in binaural hearing”, *J. Acoust. Soc. Am.* **22**, 61–62.
- Kryter, K. D. (1962). “Methods for the calculation and use of the articulation index”, *J. Acoust. Soc. Am.* **34**, 1689–1697.
- Kwon, B. J. and Turner, C. W. (2001). “Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?”, *J. Acoust. Soc. Am.* **110**, 1130–1140.

- 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. and Culling, J. F. (2010). “Prediction of binaural speech intelligibility against noise in rooms”, *J. Acoust. Soc. Am.* **127**, 387–399.
- 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.
- Leclère, T., Lavandier, M., and Culling, J. F. (2015a). “Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation”, *J. Acoust. Soc. Am.* **137**, 3335–3345.
- Leclère, T., Lavandier, M., and Deroche, M. L. D. (2016). “Speech intelligibility against a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location (under review)”, *J. Acoust. Soc. Am.* .
- Leclère, T., Théry, D., Lavandier, M., and Culling, J. F. (2015b). “Speech intelligibility for target and masker with different spectra”, in *International Symposium on Hearing (in press)*.
- Levitt, H. (1971). “Transformed up-down methods in psychoacoustics”, *J. Acoust. Soc. Am.* **49**, 467–477.
- Levitt, H. and Rabiner, L. R. (1967). “Predicting binaural gain in intelligibility and release from masking for speech”, *J. Acoust. Soc. Am.* **42**, 820–829.
- Libbey, B. and Rogers, P. H. (2004). “The effect of overlap-masking on binaural reverberant word intelligibility”, *J. Acoust. Soc. Am.* **116**, 3141–3151.
- Licklider, J. C. R. (1948). “The influence of interaural phase relations upon masking of speech by white noise”, *J. Acoust. Soc. Am.* **20**, 150–159.
- Lochner, J. and Burger, J. (1964). “The influence of reflections on auditorium acoustics”, *J. Sound and Vib.* **1**, 426–454.
- Marrone, N., Mason, C. R., and Jr., G. K. (2008). “Tuning in the spatial dimension: Evidence from a masked speech identification task”, *J. Acoust. Soc. Am.* **124**, 1146–1158.

- Miller, S. E., Schlauch, R. S., and Watson, P. J. (2010). “The effects of fundamental frequency contour manipulations on speech intelligibility in background noise”, *J. Acoust. Soc. Am.* **128**, 435–443.
- Moncur, J. P. and Dirks, D. (1967). “Binaural and monaural speech intelligibility in reverberation”, *J. Speech Hear. Res.* **10**, 186–195.
- Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing* (Elsevier Academic Press).
- 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.
- Mosko, J. D. and House, A. S. (1971). “Binaural unmasking of vocalic signals”, *J. Acoust. Soc. Am.* **49**, 1203–1212.
- Nábělek, A. K. and Robinson, P. K. (1982). “Monaural and binaural speech perception in reverberation for listeners of various ages”, *J. Acoust. Soc. Am.* **71**, 1242–1248.
- Peters, R. W., Moore, B. C. J., and Baer, T. (1998). “Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people”, *J. Acoust. Soc. Am.* **103**, 577–587.
- Plomp, R. (1976). “Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)”, *Acustica* **34**, 200–211.
- Pollack, I. (1948). “Effects of high pass and low pass filtering on the intelligibility of speech in noise”, *J. Acoust. Soc. Am.* **20**, 259–266.
- Raake, A. and Katz, B. F. G. (2006). “SUS-based method for speech reception threshold measurement in French”, in *Proceedings of Language Resources and Evaluation Conference*, 2028–2033.
- Rennies, J., Brand, T., and Kollmeier, B. (2011). “Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet”, *J. Acoust. Soc. Am.* **130**, 2999–3012.
- Rennies, J., Warzybok, A., Brand, T., and Kollmeier, B. (2014). “Modeling the effects of a single reflection on binaural speech intelligibility”, *J. Acoust. Soc. Am.* **135**, 1556–1567.

- Rhebergen, K. S. and Versfeld, N. J. (2005). “A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners”, *J. Acoust. Soc. Am.* **117**, 2181–2192.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2005). “Release from informational masking by time reversal of native and non-native interfering speech”, *J. Acoust. Soc. Am.* **118**, 1274.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise”, *J. Acoust. Soc. Am.* **120**, 3988–3997.
- Roman, N. and Woodruff, J. (2013). “Speech intelligibility in reverberation with ideal binary masking : Effects of early reflections and signal-to-noise ratio threshold”, *J. Acoust. Soc. Am.* **133**, 1707–1717.
- Shinn-Cunningham, B. G., Schickler, J., Kopčo, N., , and Litovsky, R. (2001). “Spatial unmasking of nearby speech sources in a simulated anechoic environment”, *J. Acoust. Soc. Am.* **110**, 1118–1129.
- Soulodre, G. A., Popplewell, N., and Bradley, J. S. (1989). “Combined effects of early reflections and background noise on speech intelligibility”, *J. Sound and Vib.* **135**, 123–133.
- Studebaker, G. A. and Sherbecoe, R. L. (2002). “Intensity-importance functions for bandlimited monosyllabic words”, *J. Acoust. Soc. Am.* **111**, 1422–1436.
- Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., and Gwaltney, C. A. (1999). “Monosyllabic word recognition at higher-than-normal speech and noise levels”, *J. Acoust. Soc. Am.* **105**, 2431.
- Summerfield, Q. and Assmann, P. F. (1991). “Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony”, *J. Acoust. Soc. Am.* **89**, 1364–1377.
- van Wijngaarden, S. J. and Drullman, R. (2008). “Binaural intelligibility prediction based on the speech transmission index”, *J. Acoust. Soc. Am.* **123**, 4514–4523.
- Wallach, H. (1939). “On sound localization”, *J. Acoust. Soc. Am.* **10**, 270.
- Wan, R., Durlach, N. I., and Colburn, H. S. (2010). “Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers”, *J. Acoust. Soc. Am.* **128**, 3678–3690.

Bibliography

- Wan, R., Durlach, N. I., and Colburn, H. S. (2014). “Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers”, *J. Acoust. Soc. Am.* **136**, 768–776.
- Warzybok, A., Rennie, J., Brand, T., Doclo, S., and Kollmeier, B. (2013). “Effects of spatial and temporal integration of a single early reflection on speech intelligibility”, *J. Acoust. Soc. Am.* **133**, 269–282.
- Watkins, A. J. (2005). “Perceptual compensation for effects of reverberation in speech identification”, *J. Acoust. Soc. Am.* **118**, 249–262.
- Zurek, P. M. (1993). “Binaural advantages and directional effects in speech intelligibility”, in *Acoustical factors affecting hearing aid performance*, edited by G. Studebaker and I. Hochberg, 255–276 (Allyn and Bacon, Needham Heights, MA).

Abstract - Résumé

This PhD work aims to propose a model predicting the perceived intelligibility of a target speech masked by competing sources in rooms. An existing model developed by Lavandier and Culling (2010) is already able to predict speech intelligibility of a near-field target in the presence of multiple noise sources. The present work deals with new implementations and experimental work needed to extend the model to the case of a distant target and to the case of masking voices, which present different acoustical properties than noises (envelope fluctuations, fundamental frequency, modulations of fundamental frequency). The detrimental effect of reverberation on the target speech has been successfully implemented. This new version of the model provides a unified interpretation of several perceptual effects previously observed in the literature but it presents a room dependency which limits its predictive power. Experimental work has been conducted to determine how the model could account for sources presenting different spectra, and to account for several auditory mechanisms operating simultaneously (F0 segregation, spatial unmasking and temporal dip listening).

Ce travail de thèse vise à proposer un modèle pouvant prédire l'intelligibilité d'une voix cible masquée par des sources concurrentes dans les salles. Un modèle a déjà été développé par Lavandier et Culling (2010) et est capable de prédire l'intelligibilité d'une cible en champ proche perturbée par plusieurs sources de bruit. Le travail présenté ici traite des nouvelles implémentations et expérimentations nécessaires pour étendre le modèle au cas de cibles distantes et au cas de voix concurrentes, qui présentent des propriétés acoustiques différentes des bruits stationnaires (fluctuation d'enveloppe, fréquence fondamentale, modulations de fréquence fondamentale). L'effet nuisible de la réverbération sur la parole cible a été implémenté avec succès. Cette nouvelle version du modèle permet une interprétation unifiée de plusieurs effets perceptifs observés dans la littérature mais il présente une dépendance de la salle, ce qui limite son aspect prédictif. Des travaux expérimentaux ont été menés pour déterminer comment le modèle pourrait prendre en compte le cas de sources cibles et masquantes avec des spectres différents ainsi que le cas où plusieurs mécanismes auditifs opèrent simultanément (ségrégation par F0, démasquage spatial et écoute dans les creux de modulation).