UNIVERSITÉ DE PROVENCE — AIX-MARSEILLE I

ÉCOLE DOCTORALE SCIENCES POUR L'INGÉNIEUR :

MÉCANIQUE, PHYSIQUE, MICRO ET NANOÉLECTRONIQUE (ED353)

THÈSE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ DE
PROVENCE AIX-MARSEILLE I

Discipline
ACOUSTIQUE

Présentée et soutenue publiquement par
THIBAUD NECCIARI
Le 25 octobre 2010

MASQUAGE AUDITIF TEMPS-FRÉQUENCE :

Mesures psychoacoustiques et application à l'analyse-synthèse des sons

AUDITORY TIME-FREQUENCY MASKING:

Psychoacoustical measures and application to the analysis-synthesis of sound signals

JURY :	
Pr. WEBER Reinhard (Carl von Ossietzky University Oldenburg)	Rapporteur
Dr. PRESSNITZER Daniel (LPP - CNRS UMR 8158)	Rapporteur
Pr. TORRÉSANI Bruno (Université de Provence)	Examinateur
Dr. LABACK Bernhard (Acoustics Research Institute)	Examinateur
Dr. BALAZS Peter (Acoustics Research Institute)	Examinateur
Dr. KRONLAND-MARTINET Richard (LMA - CNRS UPR7051)	Directeur de thèse
Dr. SAVEL Sophie (LMA - CNRS UPR7051)	Co-directrice de thèse

Résumé

De nombreuses applications audio, telles que les outils d'analyse-synthèse ou les codeurs audio, nécessitent des représentations des signaux linéaires et adaptées aux signaux non stationnaires. Typiquement, ces représentations sont de types « Gabor » ou « ondelettes ». Elles permettent de décomposer n'importe quel signal en une somme de fonctions élémentaires (ou « atomes ») bien localisées dans le plan temps-fréquence (TF). Dans le but d'adapter ces représentations à la perception auditive humaine, ce travail porte sur l'étude du masquage auditif dans le plan TF.

Dans la littérature, le masquage a été considérablement étudié dans les plans fréquentiel et temporel. Peu d'études se sont intéressées au masquage dans le plan TF. D'autre part, toutes ces études ont employé des stimuli de longue durée et/ou large bande, donc pour lesquels la concentration d'énergie dans le plan TF n'est pas maximale. En conséquence, les résultats ne permettent pas de prédire les effets de masquage entre des atomes TF. Au cours de cette thèse, le masquage a donc été mesuré dans le plan TF avec des stimuli — masque et cible — dotés d'une localisation TF maximale : des sinusoïdes modulées par une fenêtre Gaussienne de courte durée (ERD = 1,7 ms) et à support fréquentiel compact (ERB = 600 Hz). La fréquence du masque était fixée à 4 kHz et son niveau à 60 dB SL. Masque et cible étaient séparés en fréquence, en temps, ou en TF. Les résultats pour les conditions TF fournissent une estimation de l'étalement du masquage TF pour un atome. Les résultats pour les conditions fréquence et temps ont permis de montrer qu'une combinaison linéaire des fonctions de masquage fréquentiel et temporel ne fournit pas une représentation exacte du masquage TF pour un atome. Deux expériences supplémentaires ont été menées afin d'étudier les effets du niveau et de la fréquence du masque Gaussien sur le pattern de masquage fréquentiel. Une diminution du niveau du masque de 60 à 30 dB SL a provoqué un renversement de l'asymétrie des patterns de masquage et un rétrécissement de l'étalement spectral du masquage, conformément à la littérature. La comparaison sur une échelle ERB des patterns mesurés à 0,75 et 4 kHz a révélé un étalement spectral du masquage similaire pour les deux fréquences. Ce résultat est cohérent avec l'analyse fréquentielle à facteur de qualité constant du système auditif.

La thèse s'achève sur une tentative d'implémentation des données psychoacoustiques dans un outil de traitement du signal visant à éliminer les atomes inaudibles dans les représentations TF des signaux sonores. Les applications potentielles d'une telle approche concernent les outils d'analyse-synthèse ou les codeurs audio.

Mots clefs : psychoacoustique, masquage auditif, analyse-syntèse, temps-fréquence, Gaussienne, Gabor, ondelettes

Abstract

Many audio applications, such as sound analysis-synthesis tools or audio codecs, call for specific signal representations enabling the analysis, processing, and synthesis of non stationary signals. Most of them are concerned with time-frequency (TF) representations such as the Gabor and wavelet transforms that allow decomposing any real-world sound into a set of elementary functions (or “atoms”) well localized in the TF domain. On the purpose of adapting these representations to the human auditory perception, the present study investigated auditory masking in the TF domain.

Masking has been extensively investigated with simultaneous (frequency masking) and non-simultaneous (temporal masking) presentation of masker and target. A few studies examined TF relations of masking between masker and target. Because those studies involved stimuli that are not maximally compact in the TF plane (*i.e.*, they were temporally and/or spectrally broad), their results are not suitable for predicting masking effects between TF atoms. In this study, we investigated auditory TF masking with masker and target signals having minimum spread in the TF plane, namely Gaussian-shaped sinusoids (referred to as Gaussians). The masker had a carrier frequency of 4 kHz and a level of 60 dB SL. Masker and target were separated either in frequency, in time, or both. The results of the TF conditions provide the TF spread of masking for stimuli that are maximally concentrated in the TF domain. The results of the simultaneous and non-simultaneous conditions allowed to show that a simple superposition of frequency and temporal masking functions does not provide an accurate representation of the measured TF masking function for Gaussian maskers. Two additional experiments were carried out that examined the effects of masker level and masker frequency in simultaneous conditions. Decreasing the masker level from 60 to 30 dB SL resulted in a reversal of the masking patterns’ asymmetry and a narrowing of the frequency spread of masking. The frequency spread of masking at 0.75 kHz was similar to that obtained at 4 kHz when compared on an ERB scale. This is compatible with the constant-Q frequency analysis by the human auditory system.

Finally, a first attempt was made to implement the gathered masking data in a sound signal processing algorithm allowing to remove the perceptually irrelevant atoms in the TF representations of audio signals. Potential applications of such an approach are, for instance, audio codecs and sound analysis-synthesis tools.

Keywords: psychoacoustics, auditory masking, analysis-synthesis, time-frequency, Gaussian, Gabor, wavelets

Remerciements / Acknowledgements

Je tiens avant tout à remercier mes proches qui m'ont soutenu sans cesse tout au long de ces années et sans qui je n'en serais certainement pas à ce point (et quel point !) aujourd'hui : mes parents, grand-parents, Jean-Pierre & Vannina, Régis & Marlène, mes cousines et surtout toi, ô ma douce et tendre « Reine du Matin » : Chen. Je reconnais t'avoir fait passer des moments difficiles où j'avais plutôt la tête dans mes recherches que dans nos conversations mais tu as toujours été compréhensive et je t'en remercie. Merci pour tous les moments difficiles que tu m'as aidé à surmonter par ton réconfort, ton amour (et ta cuisine) ! Je suis persuadé que l'achèvement de ce travail sera d'un aussi grand soulagement pour toi que pour moi… Je regrette seulement que nous n'ayons pas eu l'occasion de nous affronter au ping-pong sur la « official S2M table ». Mais ce n'est que partie remise !

Je tiens ensuite à remercier tout particulièrement mes directeurs de thèse, Sophie Savel (les dames d'abord !) et Richard Kronland-Martinet, pour leur soutien, leurs encouragements et la confiance qu'il m'ont apportés tout au long de ces quatre années. Il y a eu des moments difficiles, certes, mais ils ont toujours été là pour m'aider à les surmonter, un grand merci ! Et puis de toute façon, ces dit « moments difficiles » ne représentent rien par rapport à tous les très bons souvenirs accumulés pendant la thèse, que ce soit à Marseille, Vienne, Paris, Lyon ou ailleurs. Votre humour et votre joie de vivre ont toujours su agrémenter les réunions de quelques rigolades ou apaiser les tensions quand elles se faisaient sentir (quoique… ?!). Tout cela a contribué au bon déroulement de la thèse et je vous en remercie. Je ne peux énumérer ici tous les moments forts et mémorables — scientifiques ou pas — qui ont jalonné cette thèse tellement ils sont nombreux. J'évoquerai seulement les barbecues chez Doumè, le « maître du ribs », les soirées au 1516, la réunion « temps-fréquence » aux Goudes (plus productive que jamais), les restos autrichiens, et bien d'autres ! J'espère bien pouvoir encore partager des moments comme ceux-ci en votre compagnie.

Sophie, maître Yoda de la psychoacoustique, le jeune padawan te remercie pour les enseignements et les valeurs que tu lui as inculqués. Même s'il n'est pas encore apte à élever des serpents, il peut les déjà les approcher. Même s'il ne saurait pas mettre en place une expérience sur la localisation auditive, il sait au moins mesurer et analyser des données de masquage. Même si ces enseignements ont pu se révéler difficiles par moments, ils ont sûrement porté leurs fruits et c'est maintenant au padawan de savoir les aiguiser et les utiliser, à bon escient…

Richard, je me suis souvent posé la question de savoir si je devais te remercier ou te reprocher de m'avoir lancé dans ce projet « temps-fréquence » mêlant psychoacoustique et traitement du signal. Etant issu d'une formation en traitement du signal et n'ayant aucune notion en psychoacoustique, je ne savais pas dans quoi je m'engagais en signant ce « contrat » de trois ans. Maintenant, avec le recul, je pense que je peux te remercier ! J'ai beaucoup appris par cette thèse pluridisciplinaire et j'ai maintenant une nouvelle carte à mon jeu de connaissances. Merci également pour m'avoir ré-appris à écrire correctement un produit de convolution. La preuve que la pluridisciplinarité n'a pas que du bon. . .

Je me dois ensuite de remercier toutes les bonnes âmes du LMA qui m'ont accompagné toutes ces années et avec qui j'ai passé de très bons moment, au sein ou à l'extérieur du laboratoire avec certains. Je commencerai par mes chers collègues de chez PA : Michel Jevaud (tiens le coup mon cher ! Maintenant tu préfères quoi, l'environnement cosmique ou les circuits TDT ?), Guy Rabau, Georges Canévet, Michèle Laurent, Jacques Chatron (le roi du Zeph', pas comme tout le monde !) et bien sûr Sabine Meunier (cachaça$e^{\text{bière}} \propto$ architectes^2, c'est bien ça la relation ?) pour sa disponibilité et l'aide précieuse qu'elle m'a apporté chaque fois que j'en ai eu besoin. Je n'oublierai certainement pas les auvergnats Françoise et Pyo pour leur soutien à tous moments. Un soutien qui s'est manifesté par des rigolades, des apéros, des longueurs de brasse, et du St Nectaire ! Un grand merci !

Quant à la troupe des analyseurs/synthétiseurs de sons et des acousticiens/physiciens des instruments de musique — j'ai nommé l'équipe S2M — je tiens à remercier Sølvi Ystad pour son sourire et sa bonne humeur permanents, Christophe Vergez pour ses rumeurs (« I have four words for you : I love this company ! »), Philippe Guillemain pour ses parties de ping-pong enragées, Thierry Voinier pour ses conseils (tu crois qu'on va l'avoir cette antilope ?) et Didier Ferrand pour sa franchise. Sans oublier les anciens et présents doctorants pour leur soutien, leurs encouragements et la bonne humeur qu'ils sèment au sein de l'équipe : Dr. Mathieu Barthet, Dr. Fabrice Silva (merci d'avoir accepté de subir mes tests et désolé pour leur longueur), Anaïk, les Adrien M. & S., Samy, J-F « el moustacho » S., sans oublier le montagnard Charles et le dernier docteur en date, Charly. Merci également à Bastien pour sa participation aux tests.

Enfin, merci à Erick Ogam pour son soutien (et surtout pour l'antilope sauce mangue que je n'ai pas — encore ? — goûtée). Merci à Mitsuko pour ses encouragements, à Clément François pour ses imitations et son pinard, à Damián Marelli pour ses conseils et à Philippe Depalle pour ses jeux de mots.

Je terminerai cette liste de remerciements (en français) en saluant mes collègues bûcherons « Ninou-boyz » de Charnier, J'Hell'M & Charognard (on va finir par le sortir cet album !), Aude & Olivier de METAL C.O.M.M.A.N.D. ainsi que les anciens combattants de l'ISEN : Seb « La Boulasse », Sacha « Loser » et Alban « Denis Boon ».

Enfin, salutations aux amis du territoire Corse : SSX et le Colonel ! *A populu fattu, bisognu à marchjà !*

To the non-French speakers. . .

First, I would like to acknowledge the Austrian "dream team" at the Acoustics Research Institute (ARI) — namely Bernhard Laback and Peter Balazs — for having entrusted me with the time-frequency masking project. Wasn't it kind of a challenge?! I hope I succeeded, at least I did my best to do it. . . I would also thank the whole staff at ARI for the cordial welcome in Vienna: Bernard Laback and Peter Balazs (of course!), Werner Deutsch, Wofgang Kreuzer, Piotr Majdak (Metallica rules!), Georg Rieckh, Monika Dörfler, Timo Becker, Matthew Goupell, and all those who I may have forgotten. All the time I spent in the Austrian capital was only good time. I must also credit the staff at the 1516 brewery in Vienna (when extra beers meet good food. . . what else?!), the "beer man" at the Graffiti (don't ask, just drink!) and the staff at the BOOTS store (don't walk without it. . .).

I wish to deeply acknowledge Peter "XXL" Balazs and his wife Claudia for the warm welcome in Tulln. I shall never thank you enough for all the time you have devoted to me when I was in Vienna. I wish all the best for you and your growing family: Barbara, Michael, *who's next?*

To Bernhard and XXL: many thanks for the great support you have communicated throughout the PhD. There have been some hard times but you were always here with your ideas (and jokes) and "will to fight". It helped me a lot, thank you.

Finally, I would like to dedicate the end of this thank list to all those Rock and Metal bands that help me taking the everyday stress, anxiety and anger away. Unfortunately, I cannot cite all bands here (it would be too long!) but I must al least credit the four riders of Manowar (we were born to win number one, not number two!), the metal and leather made Judas Priest (you keep the country clean!), Billy Gibbons & ZZ Top, Tom & Sodom (the saw is the law!), Metallica for their historic gig in Nîmes, Slayer, Lemmy & Motörhead (I got Rock'n'roll to save me from the gold, and it satisfies my soul!), the Brazilian guys from Krisiun and Violator (the chemical assault really happened at La Penne sur Huveaune, June'10!) and Immolation for their intense gig in Luynes (May'07).

How do you fill a schnitzel?. . .
. . . well, you cut it in two pieces!

Contents

Introduction

The present study is part of a large project exploring the time-frequency (TF) representations of sound signals. This project, conducted in collaboration between the *Laboratoire de Mécanique et d'Acoustique* in Marseille (France) and the *Acoustics Research Institute* in Vienna (Austria), aims to develop new audio signal representations based on both the properties of TF representations and the results from psychoacoustical experiments on auditory masking. A particularity of this study is its multidisciplinarity because it requires knowledges in both the psychoacoustics and signal processing fields.

In the field of sound signal processing, many applications such as sound analysis-synthesis tools, sound characterization techniques or perceptual audio codecs, seek to extract relevant information (*e.g.*, synthesis parameters, sound descriptors, or perceptual cues) from the signal based on its representation. Because the accuracy of this extraction mostly depends on the choice of the signal representation, audio applications call for specific tools enabling the analysis, processing and resynthesis of non stationary signals, *i.e.*, whose properties evolve with respect to time.

Although the Fourier transform is a very useful tool in signal processing, it is not appropriate for the analysis of non stationary signals such as real-world sounds. The Fourier transform allows to decompose any arbitrary signal into a basis of complex exponential functions (*i.e.*, sine waves) which are maximally concentrated in frequency but are of *infinite* duration. Thus, the Fourier analysis does not allow to track the temporal evolution of the spectral components in the signal. To do so, the sine-wave basis functions used in the Fourier transform must be replaced with functions which are more concentrated in time (and hence, less concentrated in frequency). This is the field of TF and time-scale representations. TF representations such as the Gabor transform allow to decompose any arbitrary signal into a set of elementary functions or "atoms" well concentrated in the TF domain. The time and frequency resolution in the TF representation is fixed by the spectro-temporal properties of the elementary functions. Alternatively, TF transforms can be thought of as a bank of bandpass filters with fixed bandwidth. Time-scale representations such as the wavelet transform decompose signals into a family of functions which are scaled (*i.e.*, dilated or compressed) versions of a prototype function called "mother wavelet". Alternatively, wavelet transforms can be thought of as a bank of bandpass filters with constant relative bandwidth. This constitutes a multi-resolution analysis in that the time and frequency resolutions depend on the scale factor. Large scales correspond to dilated impulse responses (*i.e.*, low frequencies), while small scales correspond to compressed impulse responses (*i.e.*, high frequencies). Both TF and time-scale representations allow perfect reconstruction of the input signal.

For those reasons, TF and time-scale representations are standard tools in sound signal processing. For example, in Pielemeier and Wakefield (1996), various TF and time-scale representations are assessed for the characterization of musical signals. Specifically, the cited study provides estimators of frequencies and magnitudes of the spectral components in musical instrument signals. In Aramaki and Kronland-Martinet (2006), a synthesis model of impact sounds is proposed. Impact sounds are modeled by a sum of exponentially damped sinusoids and bands of noise. The damping laws and the spectral modes, which provide the synthesis parameters, are extracted through a time-scale analysis of natural impact sounds. Many others audio applications of TF representations could be listed here. . .

An analogy can be established between the wavelet analysis and the frequency analysis performed by the human auditory system. Both analyses can be modeled as a bank of bandpass filters with constant relative bandwidth. Consequently, wavelet transforms are better adapted to auditory perception than TF transforms. To date, however, no time-scale or TF representations is able to take human auditory perception into account in signals representation, *i.e.*, by representing only the perceptually relevant components of the signal. Obtaining sparser signal representations adapted to auditory perception constitutes a crucial interest to improve the performance of many audio applications such as those cited above. Processing only the "useful" information contained in the signal may indeed result in faster algorithms.

Therefore, in this study, we aim to develop new audio signal representations which are (1) based on the properties of TF (or time-scale) representations and (2) adapted to human auditory perception. More specifically, we focus on auditory masking. Auditory masking occurs when the detection of a sound (referred to as the "target" in psychoacoustics) is degraded by the presence of another sound (the "masker"). This effect is quantified by measuring the degree to which the detection threshold of the target (in dB) increases in the presence of the masker, referred to as the "amount of masking". Given that any real-world sound can be decomposed into a set of elementary atoms, acquiring knowledge on the "basic" spread of TF masking produced by an atom may allow predicting the masking interactions between atoms, and thus identifying which components of the signal are perceptually relevant.

Masking has been extensively investigated with simultaneous (frequency masking) and non-simultaneous (temporal masking) presentation of masker and target. However, little is known about masking in the TF domain. To study frequency masking, the frequency separation (ΔF) between target and masker is varied. In the common method of masking patterns, the masker frequency is fixed, and the amount of masking is measured for various target frequencies. Studies on frequency masking generally involve relatively long-duration maskers to keep the spectrum narrow (typically, long-lasting sinusoids). To investigate temporal masking, masker and target frequencies are identical, and the temporal separation (ΔT) between masker and target is varied. Studies on temporal masking generally involve relatively broadband maskers (typically, broadband noise or clicks), allowing the precise control of the temporal properties of the stimuli.

The results from those studies were used to develop models of frequency, temporal and TF masking that are currently implemented in perceptual audio codecs (such as MPEG-1 Layer III, best known as simply “MP3”). To reduce the digital size of audio files, audio codecs decompose sounds into TF segments and exploit the properties of auditory masking to reduce the bit rates in segments which are subject to masking. To date, the prediction of masking in TF segments is based on masking data for stimuli that are not maximally concentrated in the TF plane (*i.e.*, they are temporally and/or spectrally broad). This raises the following question: to which degree the spread of TF masking produced by an elementary sound (*i.e.*, short *and* narrowband) can be predicted by these models? In other words, can a representation of TF masking be achieved by combining data measured separately in the time and frequency domains?

Given that the spectro-temporal characteristics of the stimuli used in most studies of masking are not compatible with the atomic decomposition offered by TF analysis, we can presume that the results from these studies are not suitable for prediction of masking between TF atoms. In the same idea, Boullet (2005) showed that loudness models (ISO 532B, 1975; ANSI S3.4, 2007) established from data for stationary sounds are not able to predict the loudness of non stationary sounds.

A few studies measured masking in various ΔF *and* ΔT configurations (Fastl, 1979; Kidd Jr. and Feth, 1981; Soderquist et al., 1981; Lopez-Poveda et al., 2003; Yasin and Plack, 2005). Those studies involved long-duration sinusoidal maskers *versus* short sinusoidal targets. Overall, it is important to collect data on the spread of TF masking for signals with maximal concentration in the TF domain. The results of such measurements may first allow quantifying the deviation between a linear combination of temporal and frequency masking and actual TF masking data. Second, such measurements describing the basic TF spread of masking may allow predicting the masking interactions between atoms in TF representations, and thus identifying which components of the representation are perceptually relevant.

The signal that has the best localization in the TF plane is the Gaussian. It has Gaussian shapes in both domains, and minimizes the TF uncertainty principle. Additionally, narrowband and very short Gaussians are assumed to excite a limited number of spectro-temporal observation windows of the auditory system compared to broadband and/or long signals. For those reasons, we used Gaussian-shaped sinusoids as masker and target signals to investigate TF masking.

This dissertation is divided into three parts. The first part (Chap. 1–4) provides the theoretical background of the study. The second part (Chap. 5–11) provides the experimental contribution. In the third part (Chap. 12–13), a modeling attempt is made in the context of the TF representations of sound signals.

Part I

THEORETICAL BACKGROUND

Contents of the First Part

Chapter 1

Time-frequency analysis of sound signals

Contents

In the field of sound signal processing, many applications such as sound analysis-synthesis tools, sound characterization techniques or perceptual audio codecs, seek to extract relevant information from the signal based on its representation. Therefore, the choice of the signal representation remains a fundamental aspect. Although the Fourier transform is a very useful tool in signal processing, it is not appropriate for the analysis of non-stationary signals such as real-world sounds. Indeed, the Fourier transform allows to decompose any arbitrary signal into a basis of complex exponential functions (*i.e.*, *infinite* sinusoids), but it does not allow to track the temporal evolution of the spectral components of the signal. For example, consider $s(t)$ as a real-valued signal and its time-reversed version $s(-t)$. The long-term power spectra of $s(t)$ and $s(-t)$ computed by Fourier analysis are identical, although both signals are actually different. To be able to differentiate $s(t)$ from $s(-t)$, specific tools enabling the analysis of non-stationary signals are required. Time-frequency (TF) representations offer such an option. Therefore, TF representations are widely used in sound signal processing applications like sound analysis-synthesis tools, sound characterization techniques, or audio codecs.

Because TF representations and their applications to sound signal processing constitute the framework of the present study, in this first chapter, the background theory of TF representations is established. In particular, the work presented below focuses on two methods of signals representation that are of interest in the remaining of this PhD work, namely the short-time Fourier (or Gabor) transform, and the wavelet transform.

1.1 Preliminaries

In this section, the mathematical notations and operators used in this and subsequent chapters are defined. For a more exhaustive mathematical formalism and proofs, see, *e.g.*, Vetterli and Kovačević (1995); Gröchening (2001); or Flandrin (1993, in French).

1.1.1 Notations and definitions

Let $\mathbb{C}$, $\mathbb{R}$ and $\mathbb{Z}$ denote the spaces of complex, real and integer numbers, respectively. In this study, we deal with continuous signals and sampled versions of it. These signals are real- or complex-valued and of finite duration. Thus, they can be considered as vectors of $L^2(\mathbb{R}, \mathbb{C})$, where L^2 is a Hilbert space (*i.e.*, a space of vectors). To be conform with the signal processing literature, we denote by t the continuous time variable, and k the discrete time variable such that $t = k\,T_S = k/F_S$, with T_S and F_S being the sampling period and sampling frequency, respectively ($k \in \mathbb{Z}$, $T_S, F_S \in \mathbb{R}$).

Given a complex number $u = x + jy$ where $j = \sqrt{-1}$ and a, $b \in \mathbb{R}$, $\overline{u}$ denotes the complex conjugation of u such that $(\overline{x + jy}) = (x - jy)$. Also, $\Re(u)$ and $\Im(u)$ denote the real and imaginary parts of u, respectively.

The inner product of two functions f, h in $L^2(\mathbb{R}, \mathbb{C})$ is

$$\langle f, h \rangle = \int_{-\infty}^{+\infty} \overline{f(t)}\, h(t)\, \mathrm{d}t \tag{1.1}$$

and the L^2 norm of a function is defined from the inner product as

$$||f|| = \sqrt{\langle f, f \rangle} = \sqrt{\int_{-\infty}^{+\infty} |f(t)|^2\, \mathrm{d}t}$$

1.1.2 Short theory on the Fourier analysis

Given an integrable function $f(t) \in L^2(\mathbb{R})$ and the function $h_\omega(t) = e^{j\omega t}$ ($\forall t \in \mathbb{R}$), the Fourier transform $\mathcal{F} : f(t) \mapsto \hat{f}(\omega)$ is

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t)\, e^{-j\omega t}\, \mathrm{d}t = \langle h_\omega, f \rangle \quad \text{(analysis)} \tag{1.2}$$

and the inverse Fourier transform is

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega)\, e^{j\omega t}\, \mathrm{d}\omega \quad \text{(synthesis)} \tag{1.3}$$

The Fourier transform is linear and verifies several properties, some of which are listed below for later use.

Linearity Since the Fourier transform is an inner product (see Eq. (1.1)), it follows

$$\alpha f(t) + \beta h(t) \quad \longleftrightarrow \quad \alpha \hat{f}(\omega) + \beta \hat{h}(\omega) \quad (\alpha, \beta \in \mathbb{R})$$

Shifting A shift in time by τ results in multiplication by a phase factor in the Fourier domain,

$$f(t-\tau) \quad \longleftrightarrow \quad e^{-j\omega\tau} \hat{f}(\omega) \tag{1.4}$$

Conversely, a shift in frequency results in a phase factor in the time domain

$$e^{j\omega_0 t} f(t) \quad \longleftrightarrow \quad \hat{f}(\omega - \omega_0)$$

Scaling Scaling in time results in inverse scaling in the frequency domain

$$f(at) \quad \longleftrightarrow \quad \frac{1}{|a|} \hat{f}\left(\frac{\omega}{a}\right) \quad (a \in \mathbb{R}^*) \tag{1.5}$$

Convolution The convolution of two functions $f(t)$ and $h(t) \in L^2(\mathbb{R})$ is

$$(f * h)(t) = (h * f)(t) = \int_{-\infty}^{+\infty} f(\tau)\, h(t-\tau)\, \mathrm{d}\tau \tag{1.6}$$

The convolution theorem states that

$$(f * h)(t) \quad \longleftrightarrow \quad \hat{f}(\omega)\, \hat{h}(\omega) \tag{1.7}$$

and by symmetry we get

$$f(t)\, h(t) \quad \longleftrightarrow \quad \frac{1}{2\pi} \left(\hat{f} * \hat{h}\right)(\omega)$$

Parseval's formula

$$\begin{aligned} \langle f, h \rangle &= \langle \hat{f}, \hat{h} \rangle \\ \int_{-\infty}^{+\infty} \overline{f(t)}\, h(t)\, \mathrm{d}t &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \overline{\hat{f}(\omega)}\, \hat{h}(\omega)\, \mathrm{d}\omega \end{aligned} \tag{1.8}$$

which reduces, when $h(t) = f(t)$, to

$$\int_{-\infty}^{+\infty} |f(t)|^2\, \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2\, \mathrm{d}\omega \tag{1.9}$$

Relation (1.9) verifies the energy conservation for signal $f(t)$ by Fourier transform.

1.2 Time-frequency representations

To perform the spectral analysis of a signal at a certain time location τ (and thus to "observe" transient phenomena contained in the signal), the basic idea is to replace the complex exponential basis functions used in the Fourier analysis (see Eq. (1.2)) by functions which are more concentrated in time (and thus, more extended in frequency). Given an arbitrary signal $s(t) \in L^2(\mathbb{R})$, and an analysis window $g(t) \in L^2(\mathbb{R})$ limited in time and centered at τ, the Fourier transforms of the signal windowed at several adequate times yields the short-time Fourier transform ($STFT$) of $s(t)$

$$STFT(\tau, \omega) = \int_{-\infty}^{+\infty} s(t)\, \overline{g(t-\tau)}\, e^{-j\omega(t-\tau)}\, \mathrm{d}t = \langle\, g_{\tau,\omega}, s\,\rangle \tag{1.10}$$

with

$$g_{\tau,\omega}(t) \;=\; g(t-\tau)\, e^{j\omega(t-\tau)}$$

or, in the frequency domain

$$STFT(\tau, \omega) \;=\; \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{s}(\xi)\, \overline{\hat{g}(\xi-\omega)}\, e^{j\xi\tau}\, \mathrm{d}\xi \;=\; \langle\, \hat{g}_{\tau,\omega}, \hat{s}\,\rangle \tag{1.11}$$

with

$$\hat{g}_{\tau,\omega}(\xi) \;=\; \hat{g}(\xi-\omega)e^{-j\xi\tau}$$

Equations (1.10) and (1.11) provide the TF analysis formulas. The resulting STFT provides a two-dimensional TF representation of $s(t)$. The square modulus of the $STFT$ is called a *spectrogram* and provides a distribution of the signal energy in the TF plane. The inner products at the rightmost parts of Equations (1.10) and (1.11) reflect the fact that computing the $STFT$ is equivalent to decomposing $s(t)$ on a set of elementary functions or “atoms” which correspond to temporal shifts and frequency modulates of the window $g(t)$. From Equations (1.11) and (1.3), $STFT(\tau, \omega)$ can be interpreted as the inverse Fourier transform of $\hat{s}(\xi)\, \overline{\hat{g}(\xi-\omega)}$. This leads to another formulation for the $STFT$

$$\text{for } \xi = \omega_i,\; STFT(\tau, \omega_i) \;=\; (s * \overline{g_{-\omega_i}})\,(\tau)$$

which can be interpreted as the filtering of $s(t)$ through a bandpass filter centered at ω_i and whose impulse response is $\overline{g(-t)}$.

The time and frequency resolution in the TF representation depends on the choice of the elementary functions, *i.e.*, on the duration and bandwidth of $g(t)$. This is illustrated in Figure 1.1. More precisely, the time and frequency resolution in the TF domain cannot be arbitrarily small because the time-bandwidth product is lower bounded

$$\Delta t\, \Delta f \geq \frac{1}{4\pi} \tag{1.12}$$

which is referred to as the TF “uncertainty principle”, or Gabor-Heisenberg inequality. It means that one can only trade time resolution for frequency resolution, or *vice versa*. When a Gaussian window is used as $g(t)$, Equation (1.12) becomes an equality. This means that a Gaussian window minimizes the TF uncertainty and has minimum spread of energy in the TF plane (*i.e.*, maximal concentration in the TF domain). Therefore, Gaussian windows are most often used as analysis windows for the computations of $STFT$s, which are then called Gabor transforms (Gröchening, 2001). An important property of TF representations is that once an analysis window $g(t)$ with Δt and Δf has been chosen, then the TF resolution is

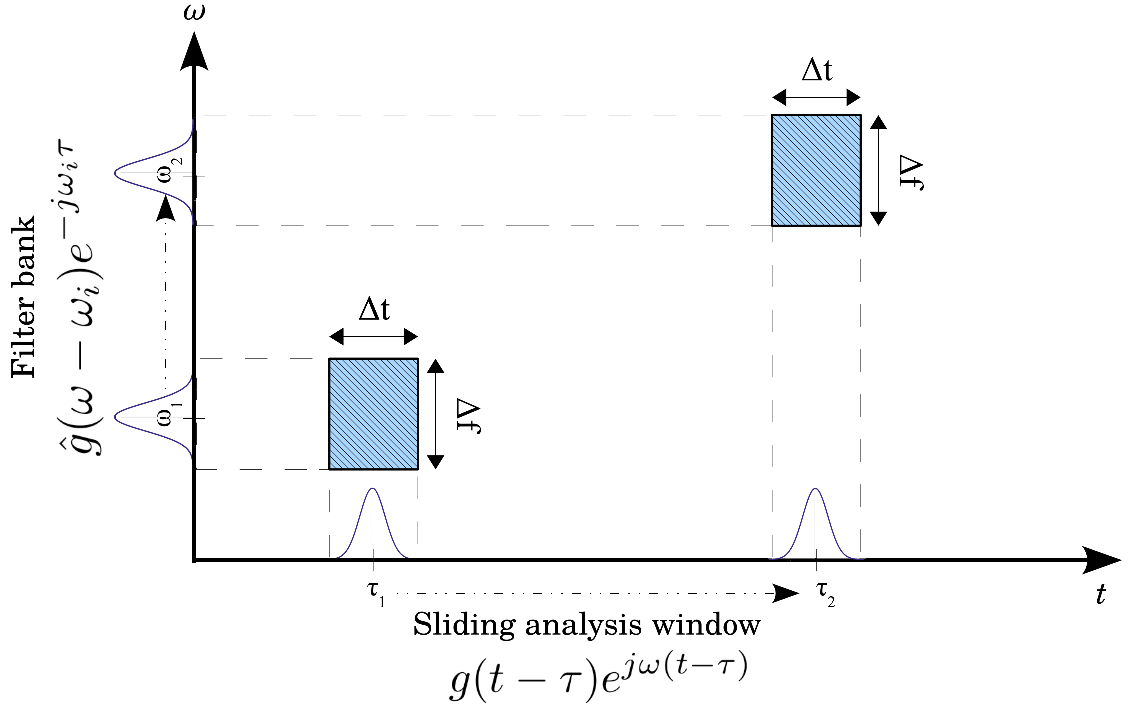


Figure 1.1: The short-time Fourier transform (STFT) of a signal is obtained by computing the Fourier transforms of local portions of the signal at several time locations. This is achieved using a time-limited analysis window $g(t)$ that is shifted in time. An alternative view is to consider that the signal is filtered through a bandpass filter centered at $\omega = \omega_i$ and whose impulse response is $\overline{g(-t)}$. The time and frequency resolution in the STFT depends on the duration (Δt) and bandwidth (Δf) of $g(t)$.

fixed over the entire TF domain (see Fig. 1.1). Another property of the $STFT$ is its invertibility. The synthesis formula is

$$s(t) = \frac{1}{||h||^2} \iint_{-\infty}^{+\infty} STFT(\tau, \omega)\, h(t-\tau)\, e^{j\omega(t-\tau)}\, \mathrm{d}\tau\, \mathrm{d}\omega \tag{1.13}$$

where $h(t)$ is the synthesis window. In most cases, the analysis and synthesis windows are the same (*i.e.*, $g(t) = h(t)$). Note that the synthesis imposes a constraint on the choice of $g(t)$, namely to be of finite energy (*i.e.*, $0 < ||g||^2 < +\infty$). Finally, Equation (1.13) can be interpreted as follows: any signal $s(t)$ can be decomposed into a sum of TF atoms which are functions with minimum spread in the TF plane. In the case of the Gabor transform, the signal is decomposed into a sum of Gaussian components with appropriate amplitudes and phases (Gröchening, 2001). An example of Gabor transform is provided in Figure 1.2a below.

1.3 Time-scale representations

1.3.1 Analysis and synthesis

While the resolution Δt and Δf is fixed in TF representations, time-scale representations allow to vary the time and frequency resolution in the TF plane so as to obtain a multi-resolution analysis. This is achieved by using a family of functions which are *scaled* (*i.e.*, dilated or compressed) versions of a prototype function $g(t) \in L^2(\mathbb{R})$ such that

$$g_a(t) \;=\; \frac{1}{\sqrt{a}}\, g\left(\frac{t}{a}\right) \qquad (a \in \mathbb{R}^{*+}) \tag{1.14}$$

where a is a *scale factor* allowing to compress ($a < 1$) or dilate ($a > 1$) the prototype function $g(t)$ ($a = 1$). The factor $1/\sqrt{a}$ allows to normalize the functions so that all scaled versions of the prototype have the same energy (*i.e.*, $||g_a(t)||^2 = ||g(t)||^2$). Thus, computing the inner product between an arbitrary signal $s(t) \in L^2(\mathbb{R})$ and scaled and time-shifted versions of $g(t)$ yields the continuous wavelet transform (CWT) of $s(t)$

$$\begin{aligned} CWT_s(b,a) &= \langle\, g_{a,b},\, s\,\rangle \\ &= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(t)\, \overline{g\left(\frac{t-b}{a}\right)}\, \mathrm{d}t && (1.15) \\ &= \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{+\infty} \hat{s}(\omega)\, \overline{\hat{g}(a\omega)}\, e^{jb\omega}\, \mathrm{d}\omega \quad \text{(Parseval)} && (1.16) \end{aligned}$$

where $b \in \mathbb{R}$ is the time variable, and $CWT_s(b,a)$ provides a two-dimensional representation of the signal in the time-scale plane. In the same way as the spectrogram is defined for the $STFT$, the square modulus of the CWT provides a distribution of the signal energy in the time-scale plane, and is called a *scalogram*. Overall, the wavelet decomposition consists in measuring the “similarity” between the signal and the functions $g_{a,b}(t)$, called *wavelets*. The prototype function $g(t)$ for $a = 1$ is called the *mother wavelet*. Any real-valued or complex analytic function can be elected to $g(t)$ on condition that it oscillates in time, *i.e.*, $g(t)$ is a bandpass function. To achieve the reconstruction of a signal from its wavelet transform (see the synthesis formula in Eq. (1.19)), the wavelets must be of finite energy. An admissibility condition can be formulated as

$$C_g \;=\; \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|\hat{g}(\omega)|^2}{\omega}\, \mathrm{d}\omega \;<\infty \tag{1.17}$$

This is usually verified in practice since $\hat{g}(\omega)$ is the response of a bandpass filter with a high roll-off. The admissibility condition for $g(t)$ is then reduced to

$$\int_0^{+\infty} g(t)\, \mathrm{d}t \;=\; g(0) \;=\; 0 \tag{1.18}$$

namely that $g(t)$ be of zero mean. As for the STFT, the wavelet decomposition can be viewed as a bank of bandpass filters with impulse responses $\overline{g_a(-t)}$.

Indeed, Equation (1.16) can be interpreted as the inverse Fourier transform of $\sqrt{a}\,\hat{s}(\omega)\,\overline{\hat{g}(a\omega)} = (s * \overline{g_a})\,(t)$, which reflects a filtering process. Considering a mother wavelet $\hat{g}(\omega)$ with a pulsation ω_0 and a passband in the interval $[\omega_{min}\,;\,\omega_{max}]$, then for any scale factor $a \neq 1$, a "new" bandpass filter is defined with center pulsation $a\omega_0$ and a passband in $[a\omega_{min}\,;\,a\omega_{max}]$. While the center pulsation and passband of the filter are modified with a, the quality factor "Q" of the filer, defined as the ratio between the passband and the center pulsation (see, e.g., Tooley, 2006, pp. 77–78), remains constant and does not depend on a

$$Q \;=\; \frac{a\omega_{max} - a\omega_{min}}{a\omega_0} \;=\; \frac{\omega_{max} - \omega_{min}}{\omega_0} \;=\; constant$$

In contrast to the $STFT$ that can be considered as a bank of bandpass filters with a fixed bandwidth (*i.e.*, a fixed resolution in the TF plane), the CWT can be thought of as a bank of bandpass filters with constant relative bandwidth, also referred to as a "constant-Q" analysis. Such a constant-Q analysis provides the CWT with a resolution that depends on the scale. Since the filter bank impulse responses in (1.15) are dilated as scale increases, large scales correspond to stretched signals (*i.e.*, low frequencies), while small scales correspond to compressed signals (*i.e.*, high frequencies). Note that a is linked to frequency according to $\omega = \omega_0/a$ where ω_0 is the pulsation of the mother wavelet. Overall, time-scale representations can be interpreted as TF representations but whose frequency axis is logarithmically scaled (see Fig. 1.2b).

Finally, the wavelet synthesis formula is

$$s(t) \;=\; \frac{1}{C_g} \iint_{a>0,\,\mathbb{R}} CWT(b,a)\, g_{a,b}(t)\, \frac{\mathrm{d}a\mathrm{d}b}{a^2} \tag{1.19}$$

where C_g is the normalization constant defined in (1.17). Equation 1.19 can be interpreted as the decomposition of $s(t)$ into wavelets.

1.3.2 Properties

Because the CWT is an inner product between the signal and the wavelets, it is linear. Moreover, the CWT ensures energy conservation. Given $s(t) \in L^2(\mathbb{R})$ and its wavelet transform $CWT_s(a,b)$, the following property is verified

$$\int_{-\infty}^{+\infty} |s(t)|^2\,\mathrm{d}t \;=\; \frac{1}{C_g} \iint_{-\infty}^{+\infty} |CWT_s(a,b)|^2\, \frac{\mathrm{d}a\mathrm{d}b}{a^2}$$

A property specific to the CWT is the "reproducing kernel", which states that

$$CWT(a',b') \;=\; \frac{1}{C_g} \iint_{\mathbb{R}} K_g(a',b',a,b)\, CWT(a,b)\, \frac{\mathrm{d}a\mathrm{d}b}{a^2} \tag{1.20}$$

where

$$K_g(a',b',a,b) \;=\; \langle\, g_{a,b}, g_{a',b'}\,\rangle$$

is the reproducing kernel. Equation (1.20) means that the reproducing kernel ensures a strong correlation between all components of the wavelet transform. In other words, any component at location a', b' depends upon remote component at location

a, b through the reproducing kernel. Note that $K_g(a', b', a, b)$ is the wavelet transform of $g_{a,b}$ at location a', b'.

The CWT has other interesting properties such as the time and frequency localizations, which can be very useful for the estimation of signals' properties from the CWT. For a review see, *e.g.*, Vetterli and Kovačević (1995, Sec. 5.1.2, p. 316).

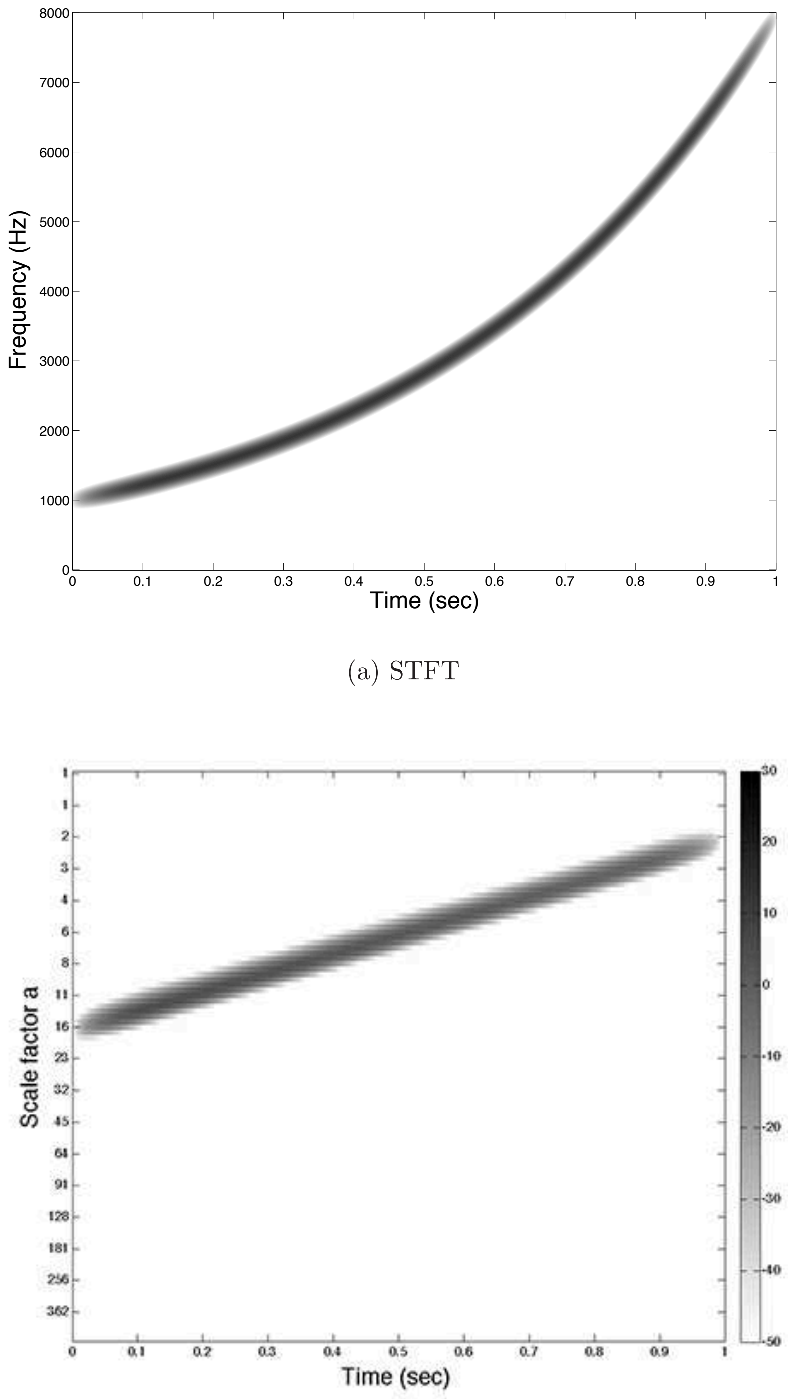
1.4 Examples

Figure 1.2 shows the TF and time-scale representations of an exponential chirp (see figure caption for details on signal generation). Both representations were computed[1] with a Gaussian window as $g(t)$. Because the Gabor transform (Fig. 1.2a) has a frequency axis which is linearly scaled, it clearly depicts the exponential evolution of the spectral content of $s(t)$. On the other hand, the wavelet transform (Fig. 1.2b) has a frequency axis which is logarithmically scaled. Therefore, in the time-scale domain, the exponential chirp is represented as a linear function of time.

To show up the multi-resolution property of the CWT, Figure 1.3 shows the spectrogram (Fig. 1.3a) and scalogram (Fig. 1.3b) of two Dirac pulses. Particularly, it can be seen that the two pulses are well time-localized at high frequencies (*i.e.*, small scales) in the scalogram. Increasing the scale factor a results in "global views" of the signal. Conversely, the spectrogram provides the same resolution over the entire TF domain.

Finally, Figure 1.4 shows the spectrogram and scalogram of a clarinet sound playing a A3 (fundamental frequency = 220 Hz). Both the STFT and the CWT allow distinguishing the fundamental (bottom horizontal stripe) from the harmonic components.

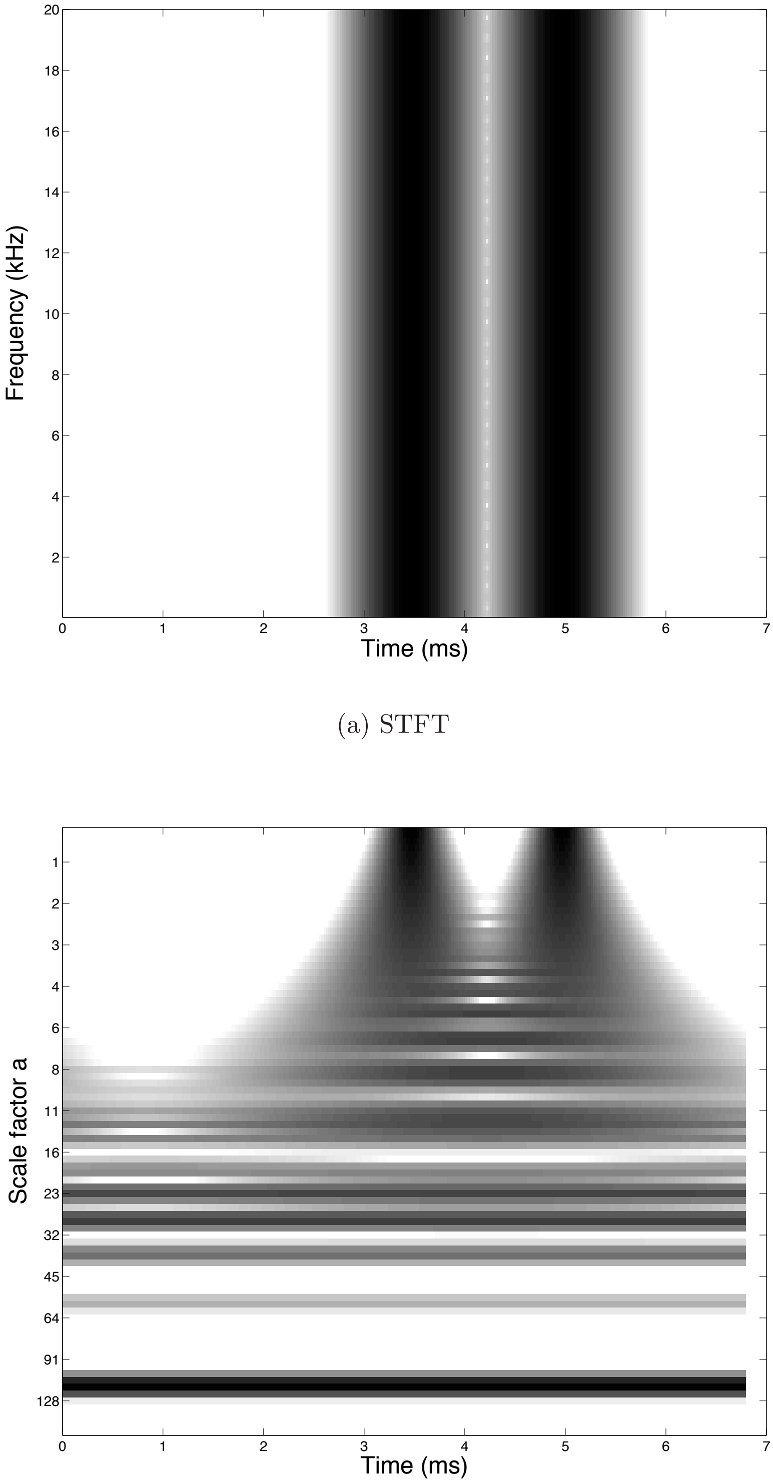
[1] The Gabor transform was computed with the help of the Linear Time-Frequency Analysis Toolbox (LTFAT) by Soendergaard (2010). The wavelet transform was computed with the discrete wavelet analysis-synthesis scheme described in Chapter 13.



(a) STFT

(b) CWT

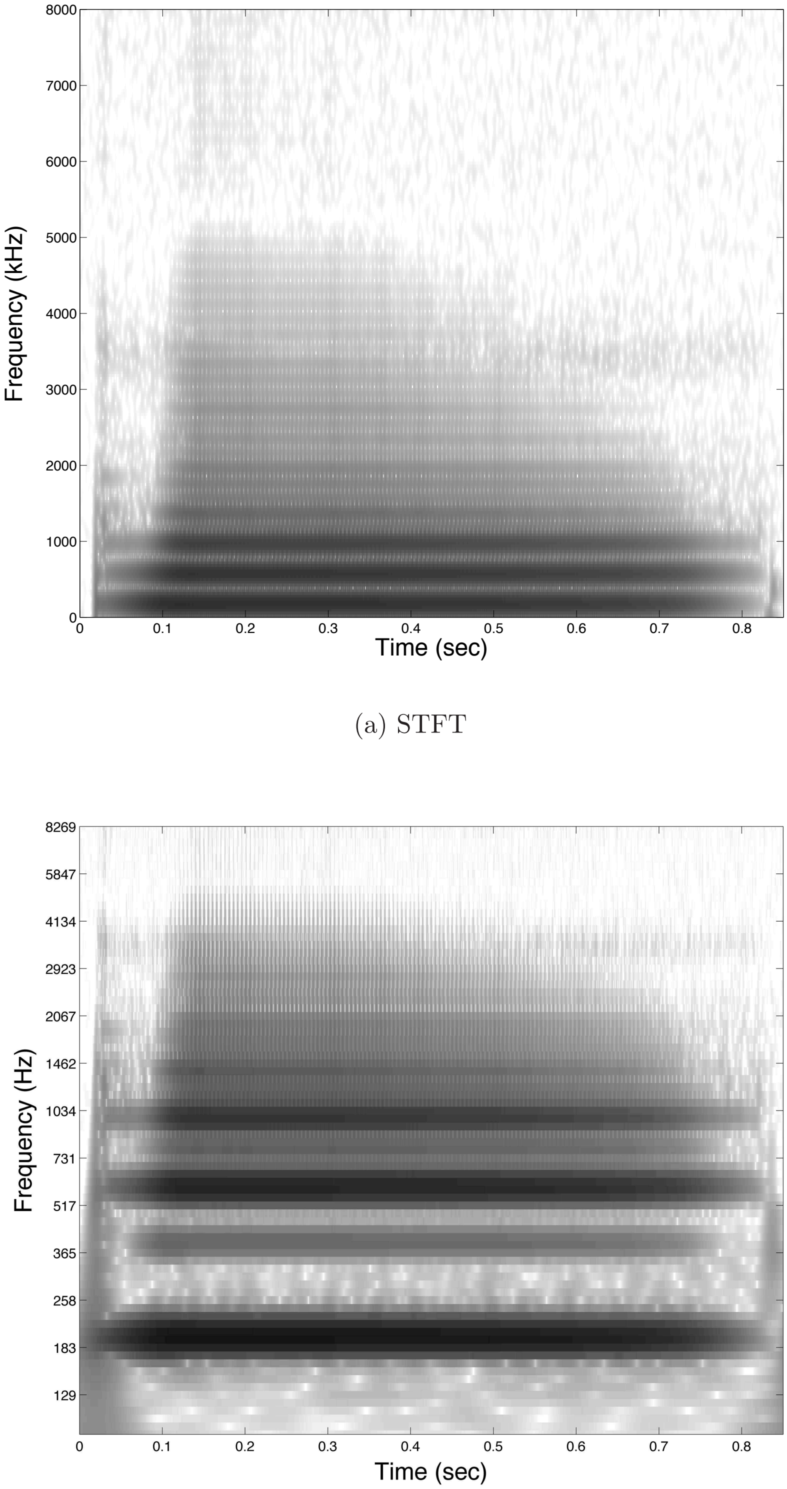
Figure 1.2: Spectrogram (a) and scalogram (b) of an exponential chirp of the form $s(t) = \sin\left(2\pi f_0 \left(\frac{\beta^t - 1}{\ln \beta}\right)\right)$ where f_0 is the starting frequency (fixed to 1000 Hz), and β (fixed to 8) controls the temporal evolution of the spectral components.



(a) STFT

(b) CWT

Figure 1.3: Spectrogram and scalogram of two Dirac pulses.



(a) STFT

(b) CWT

Figure 1.4: Spectrogram and scalogram of a clarinet sound playing a A3 (fundamental frequency = 220 Hz).

Summary

While the Fourier analysis is a very useful tool in signal processing, it is not appropriate for the analysis of non-stationary signals such as real-world sounds whose spectral properties evolve with respect to time. Most of sound signal processing applications call for specific tools enabling the analysis, processing and synthesis of non-stationary sounds. More specifically, such applications search for a signals representation leading to a first-sight interpretation of the signal properties. Most of audio applications are concerned with time-frequency (TF) and time-scale representations, which constitute the framework of the present study.

TF representations (such as the short-time Fourier or Gabor transform) allow to decompose complex sounds into a sum of elementary functions or "atoms" well localized in the TF plane. TF analysis schemes can be considered as a bank of bandpass filters with fixed bandwidth. Therefore, in such representations, the TF resolution is fixed over the entire domain and depends on the analysis window (basis function). The resolution cannot be arbitrarily small because it is constrained by the TF uncertainty principle stating that one can only trade time resolution for frequency resolution, or *vice versa*. A Gaussian window minimizes the TF uncertainty and has maximal concentration in the TF plane.

Time-scale representations (such as the continuous wavelet transform, CWT decompose complex sounds into a family of functions (*wavelets*) which are scaled versions of a prototype function (*mother wavelet*). The prototype must oscillate in time, *i.e.*, be a bandpass function. The scale factor is related to frequency and determines the time and frequency resolution. Large scales correspond to dilated signals (low frequencies), while small scales correspond to compressed signals (high frequencies). The CWT can be considered as a bank of bandpass filters with constant relative bandwidth, *i.e.*, *constant-Q analysis*.

Chapter 2

Time-frequency processing by the human auditory system

Contents

The mathematical properties of time-frequency and time-scale representations and their applications to sound signals analysis were described in Chapter 1. The present chapter focuses on how the human auditory system processes sounds. Such processes involve, among several physiological and neurological mechanisms, both temporal and frequency analysis. The comprehension of these indissociable temporal and frequency processing is essential for understanding and interpreting auditory masking, which is the main concern of the present PhD work. After a brief anatomic and functional description of the peripheral auditory system, the time-frequency processing of sounds is presented in this chapter. Temporal resolution, temporal integration and frequency selectivity are described.

2.1 Anatomy and functions of the auditory system

The auditory system is composed of two parts: (1) the peripheral system, illustrated in Figure 2.1, which is composed of the outer, middle and inner ears, and (2) the central nervous system. The current chapter mainly focuses on the former. For a detailed review on the central system see, for example, Romand et al. (1992, in French); Palmer (1995); Gelfand (1998, Chap. 2).

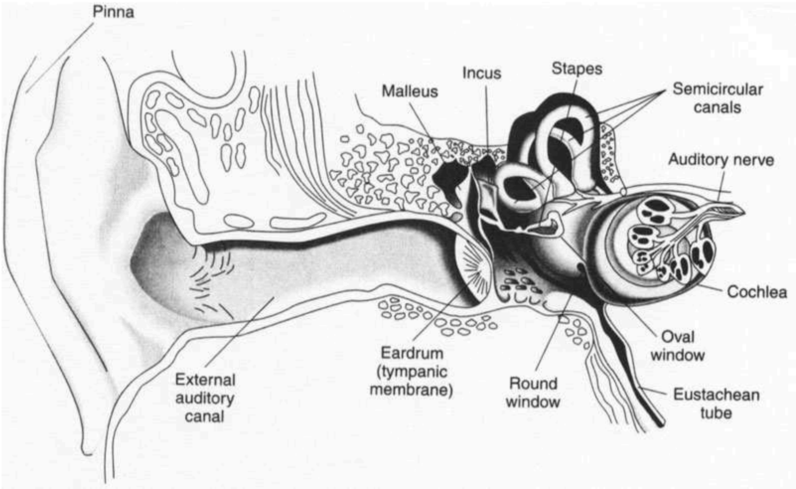


Figure 2.1: Structure of the human peripheral auditory system.

2.1.1 The outer and middle ears

The outer ear is composed of the pinna and the external auditory canal. Both can be thought of as resonant systems with proper frequencies situated in the high frequency region of the audible spectrum (3 kHz and above). The pinna has an irregular shape composed of multiple cavities. Besides providing several resonant frequencies, these irregularities induce multiple reflections so as to delay some components of the incident sounds. Theses properties provide the pinna an important role in sound localization.

When a sound is presented to the ear, it travels the auditory canal, which acts like a pressure amplifier around its resonant frequency, and causes the eardrum to vibrate. These vibrations are transmitted through the middle ear by three ossicles: the Malleus, Incus and Stapes. The Stapes is in contact with the oval window, the frontier between the middle and inner ears.

The main function of the middle ear is to ensure the efficient transfer of sound from the air to the fluids present in the cochlea, the spiral-shaped structure in the inner ear. Thus, the middle ear acts like an impedance matcher. The transmission of sound through the middle ear is the most efficient at frequencies from 500 to 4000 Hz (Aibara et al., 2001). The middle ear transfer function also depends on the level of incoming sounds. The ossicular chain is maintained by small muscles that contract upon exposure to intense sounds. This contraction, known as the “middle-ear reflex", reduces the transmission of sound through the middle ear and may help to prevent damage in the delicate structures of the cochlea. Nevertheless, the activation of this reflex is slow, about 150 ms after the onset of a high-intensity stimulation (Liberman and Guinan, 1998; Pang and Guinan, 1997), and is therefore unable to protect the inner ear from impulsive sounds. Note that the activation of the middle-ear reflex can also result in a reduction of the masking of high-frequency sounds by lower ones. This is discussed in Section 3.2.

2.1.2 The inner ear

The inner ear comprises two cavities: (1) the semicircular canals, which are part of the balance organ and do not seem to play a role in auditory processing, and (2) the cochlea, which contains the specialized receptors of hearing. Figure 2.2(a) shows a partial section of the cochlea. Figure 2.2(b) presents a cross section of the cochlear duct. The latter is filled with fluids and is divided into three juxtaposed canals: (1) Scala vestibuli, (2) Scala media and (3) Scala tympani. The first two canals are separated by the Reissner's membrane while the other two are separated by the basilar membrane (BM). It is the response of the BM to sound vibrations arriving to the oval window that is of primary importance.

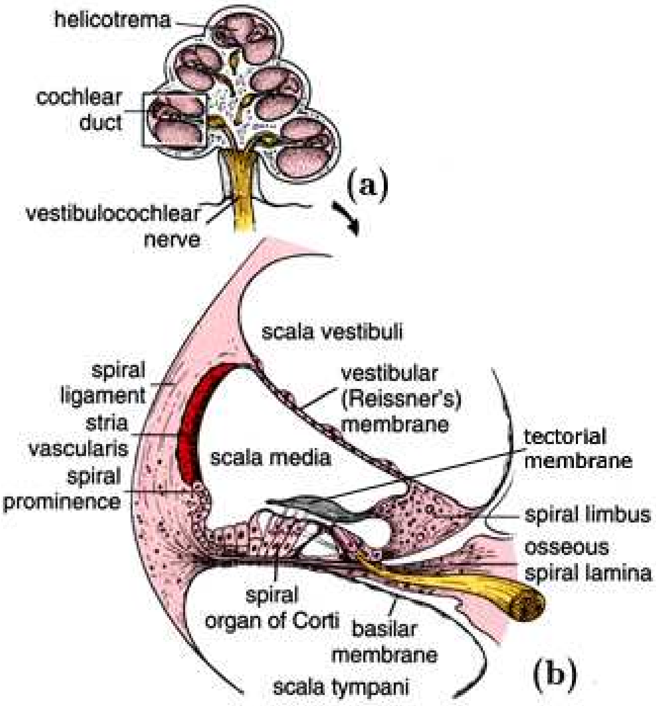


Figure 2.2: (a) Partial section of the cochlea. (b) Cross section of the cochlear duct showing the three canals (Scala vestibuli, Scala media and Scala tympani) separated by the Reissner's membrane and the basilar membrane. Along the basilar membrane runs the organ of Corti.

2.1.2.1 The basilar membrane

One end of the cochlea, at the connection with the oval window, is the “base", while the other end is the “apex" (see Fig. 2.3). When the movements of the Stapes set the Oval window in motion, a compression is produced in the Scala vestibuli. Because the Scala vestibuli and Scale tympani are connected at the apical end of the cochlea (the helicotrema), the compression propagates up to the round window where it is damped. The response of the BM to sounds of different frequencies is strongly affected by its mechanical properties. Indeed, there is a widening of the BM along the cochlea from its base to its apex, resulting in a gradation of stiffness. Because of this stiffness gradient, a sinusoidal stimulation of the BM results in the formation of a pressure wave traveling from the base to the apex. The position of the peak in the pattern of vibration depends on the frequency of stimulation. High-frequency sounds rather produce maximum displacement of the BM near the base with little movement on the rest of the membrane. Low-frequency sounds produce a pattern of vibration which extends all the way along the BM but that reaches a maximum before the apex. Figure 2.4 shows the envelopes of the vibration patterns of the BM obtained by von Békésy (1960) with seven low frequencies on the cochlea of a human cadaver. Although it is now known that BM responses to sinusoidal stimulations are more sharply tuned than those shown in Figure 2.4, it illustrates the important point that sounds of different frequencies produce maximum displacement at different places along the BM. The frequency that gives the maximum response at a particular point on the BM is called the “characteristic frequency" (CF) of that point.

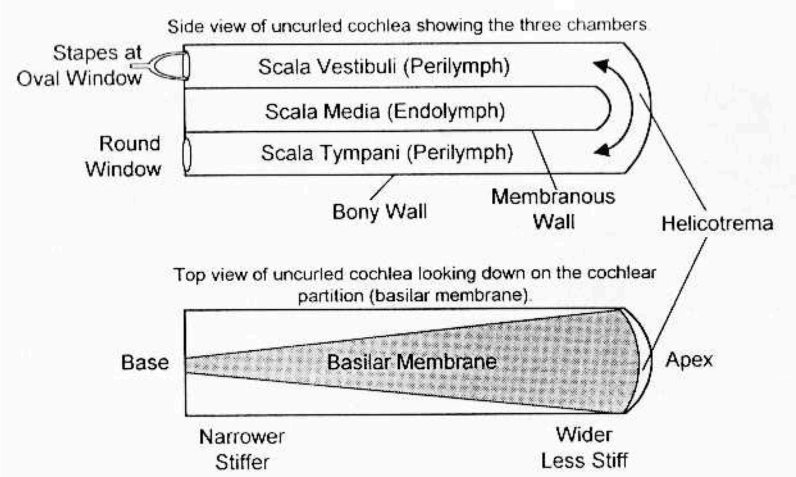

Side view of uncurled cochlea showing the three chambers
Stapes at Oval Window
Scala Vestibuli (Perilymph)
Scala Media (Endolymph)
Round Window
Scala Tympani (Perilymph)
Bony Wall
Membranous Wall
Helicotrema
Top view of uncurled cochlea looking down on the cochlear partition (basilar membrane)
Base
Basilar Membrane
Apex
Narrower Stiffer
Wider Less Stiff

Figure 2.3: Schematic representation of the uncoiled cochlea showing the different canals and the connections with the Oval window and the Round window. From Gelfand (1998).

What about the response of the cochlea to an impulsive sound like a click? Because such a sound contains energy at all frequencies, it will cause all points on the BM to vibrate in a sinusoidal manner at their own CFs. Thus, the envelope of the resulting traveling wave is not as prominent as it is with sinusoidal stimuli. This illustrates the fact that the cochlea behaves like a Fourier analyzer (see Sec. 2.3).

2.1.2.2 The organ of Corti

Along the BM runs the organ of Corti that contains the auditory sensory receptors: the hair cells. The hair cells transduce the mechanical vibrations of the BM into nervous impulsions to be processed by the central auditory system. The hair cells are divided into two groups: (1) outer hair cells (OHCs), and (2) inner hair cells (IHCs). The tip of each cell is covered with hairs, called stereocilia, which are in contact with the tectorial membrane (see Figs. 2.2b and 2.5).

When the BM starts to vibrate as a result of an acoustical perturbation, a shearing motion is created between the BM and the tectorial membrane. This motion causes a displacement of the stereocilia. An illustration of this displacement

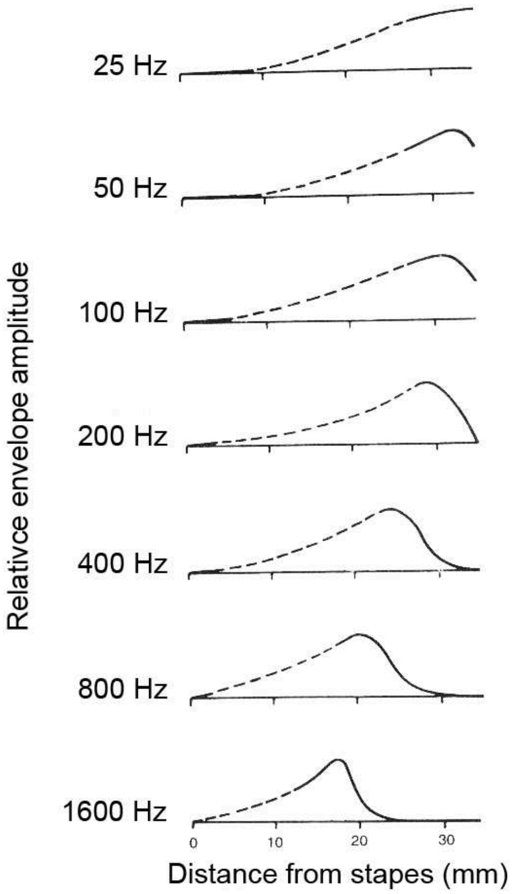


Figure 2.4: Envelopes of vibration patterns of the BM measured by von Békésy (1960) on the cochlea of a human cadaver with seven low-frequency sinusoids.

is given in Figure 2.5. This displacement results in a voltage difference between the inside and the outside of the cells and, finally, in the initiation of action potentials in auditory nerve fibers. A great majority of afferent nerve fibers (*i.e.*, which convey information from the cochlea to the central system) connects to IHCs. Conversely, most of the efferent nerve fibers (*i.e.*, which convey information from the central system to the cochlea) connect to OHCs.

For the OHCs, the displacement of the stereocilia results in cell contractions of two types: (1) fast contractions (up to 10 kHz) in phase with the stimulating signal, and (2) slower contractions (below 1 kHz) that modulate the fast ones. These two types of contractions provide the OHCs a double role. The fast contractions, called "active mechanisms" of the cochlea (Davis, 1983), result in an amplification of the BM vibrations. The slow contractions result in a filtering of the vibration pattern. Overall, the OHCs actively influence the mechanics of the cochlea so as to produce high sensitivity and sharp tuning. These active mechanisms are schematized in Figure 2.6 (adapted from Pujol, 1990). Note that Ruggero and Rich (1991) showed that drugs or other agents (*e.g.*, aspirin) selectively affecting the activity of the OHCs result in a loss of sharp tuning and in a reduction of the sensitivity of the BM.

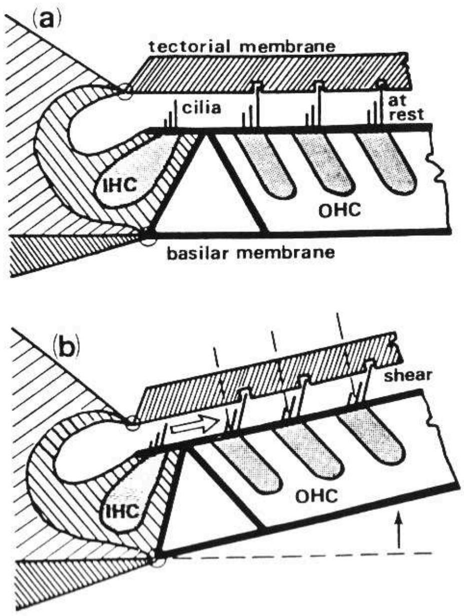


Figure 2.5: Relative positions of the BM and tectorial membrane (a) at rest, and (b) during elevation towards the Scala vestibuli. This deflection causes the stereocilia to bend. From Gelfand (1998).

2.1.3 Sound coding in the auditory nerve

The auditory nerve is composed of 30 000 afferent neurons that convey information from the cochlea to the central system, and of about 500 efferent neurons that perform a feedback control on the cochlea. The electrical activity in a single nerve fiber is characterized by a spontaneous firing rate ranging from 0 to 250 spikes per second in the absence of sound stimulation. Auditory nerve fibers are classified into three groups on the basis of their spontaneous rates (Liberman, 1978): 61% of fibers have high spontaneous rates (18–250 spikes per second), 23% have medium rates (0.5–18 sps), and 16% have low rates (< 0.5 sps). An acoustical stimulation results in an increase of the spontaneous firing rate. The threshold of a neuron is the lowest sound level for which a change in the response of this neuron can be measured (see Fig. 2.6). High spontaneous rates tend to be associated with low thresholds and vice versa. The most sensitive neurons have thresholds close to 0 dB SPL whereas the least sensitive neurons may have thresholds of 80 dB SPL or more. As the BM vibrates on a broader area when the stimulus level increases, a greater number of neurons is solicited, each of these neurons responding to a given range of levels. Intensity coding in the auditory nerve is thus provided by this parallel working neural network that provides the ear a huge dynamics of 120 dB SPL.

Furthermore, the fibers show frequency selectivity. For a given level, the response of a single fiber will be maximal for a stimulus frequency corresponding to the CF of that fiber. The CF corresponds to the frequency at which the threshold of the fiber is the lowest. The frequency selectivity of a fiber is often illustrated by tuning curves that show the threshold of a fiber as a function of frequency. An example of tuning curves is given in Figure 2.7 (adapted from Aran, 1988). It is important to

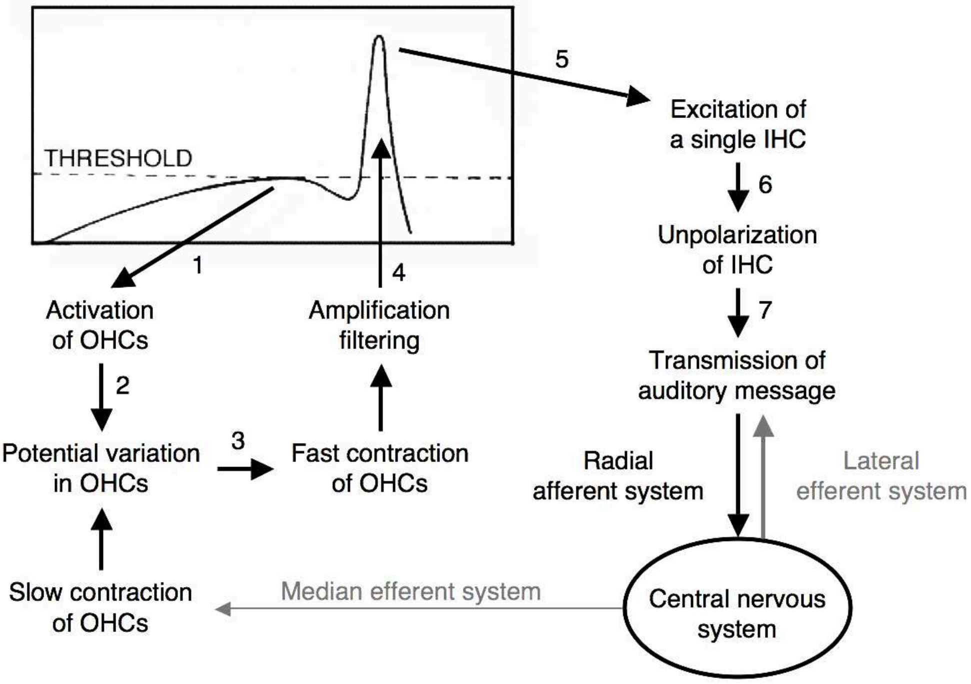


Figure 2.6: Diagram summarizing the active mechanisms of the cochlea. Adapted from Pujol (1990). When an acoustical perturbation causes the BM to vibrate, OHCs are activated (1, 2 and 3). The fast contractions result in an amplification of the vibration (4), the excitation and unpolarization of a single IHC (5, 6), and transmit the auditory message to the central system (7). The latter exerts a feedback on IHCs via the lateral efferent system and modulates the OHCs contractions via the median efferent system.

mention that there is a coincidence between the CF of a nerve fiber and the peak of the BM vibrating pattern where the fiber is connected. There is a frequency-to-place organization along the cochlea which is called the “tonotopic organization" of the auditory system. This property has consequences on the frequency analysis of the ear (see Sec. 2.3).

Overall, the sound signal transmitted from the cochlea to the auditory nerve is the combination of all the neural impulses released by the auditory nerve fibers. Afterwards, the coded signal is transmitted to the auditory cortex and submitted to higher-level processes.

2.1.4 Auditory nonlinearities

The functioning of the auditory system described above, by its complexity, underlies several non linear phenomena. In an attempt to understand the origins of these nonlinearities, a number of authors have developed computational cochlear models (for a review see, *e.g.*, Zwicker, 1985; Lopez-Poveda and Meddis, 2001; Bian and Chen, 2008). These models allowed to establish that the BM is the main source of nonlinearity, although some nonlinearities are suggested to occur at higher levels, in the central system.

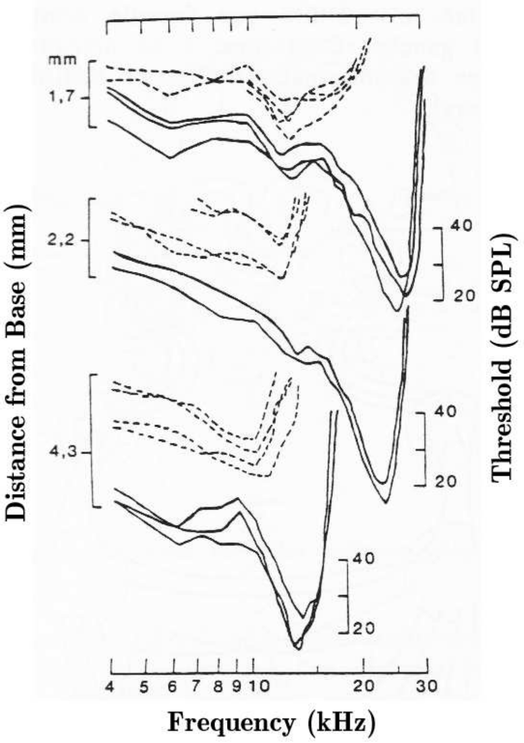


Figure 2.7: Examples of tuning curves measured in three different areas in the cochlea of guinea pigs. The left ordinate represents the distance from the base of the cochlea. The right ordinate is the threshold, in dB SPL. Solid curves are results for healthy guinea pigs. Dashed curves correspond to subjects for whom inner hair cells have been removed in the measuring area. Dashed curves show less frequency selectivity. Adapted from Aran (1988).

The following paragraphs briefly describe some of these non linear phenomena that play an important role in auditory masking results: BM compression, combination tones, and two-tone suppression.

2.1.4.1 Basilar membrane compression

As described in Section 2.1.2, there is a frequency-to-place affectation along the BM, so that each point of the BM produces its maximum response when excited at its CF. In fact, the response of the BM is not linear but rather compressive: the magnitude of the vibration does not grow in direct proportion to the level of the input stimulus. Typical responses at one point along the BM to tones of different frequencies and intensities, measured by Ruggero et al. (1997), are shown in Figure 2.8. The different curves represent tones of different levels, ranging from 10 to 90 dB SPL. The CF of the measurement point is defined as the frequency that produces the greatest response at low levels, in this case 10 kHz. Two important features can be deduced from this graph. First, the BM tuning is very sharp at low levels and broadens at high levels. Second, the change in response as a function of input level varies with frequency. For frequencies well below the CF, the response growth is linear (*i.e.*, a 10-dB change in sound pressure level produces a 10-dB change in BM response). For frequencies around the CF, the growth is highly compressive. This is illustrated in Figure 2.9, where some results of Figure 2.8 are re-plotted. The solid line of Figure 2.9 indicates how the response of the BM to a tone at the CF (10 kHz) is linear at very low levels (below 20 dB SPL) and then becomes highly compressive at higher levels. On the other hand, the dashed line shows that the response to a tone whose frequency is well below the CF (5 kHz in this case) is linear. Whereas physiological data showed that the input-output function of the BM is less compressive at low CFs (*i.e.*, at the apical region of the cochlea, see for example Cooper and Yates, 1994), more recent psychophysical data revealed that the low-frequency region is as much compressive as the high-frequency region (Lopez-Poveda et al., 2003).

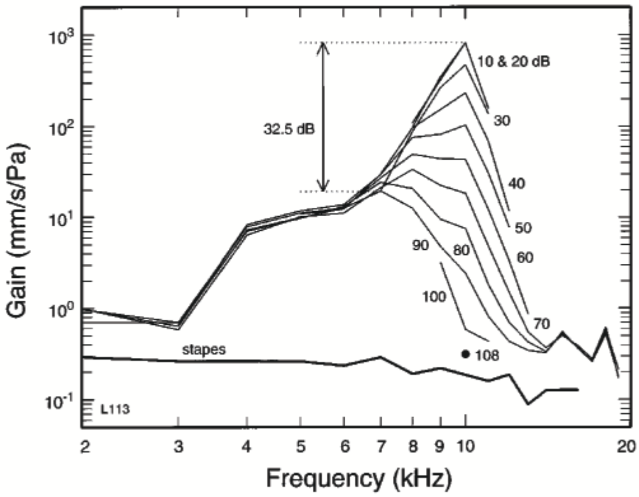


Figure 2.8: Responses of the chinchilla BM at a CF of 10 kHz to fixed-level tones with a variable frequency. The level of the tones ranged from 10 to 90 dB SPL. From Ruggero et al. (1997).

Moreover, some works showed that the nonlinearities of the BM decrease as the cochlear active mechanisms are inhibited by drugs or other agents (*e.g.*, aspirin, see Ruggero and Rich, 1991), and that the responses are linear after death (Ruggero et al., 1997). This suggests that the cochlear active mechanisms are implicated in the non linear input-output response of the cochlea.

2.1.4.2 Combination tones

When two or more sound components are presented simultaneously, a listener can hear other components that are not present in the source signal. These other components, called “combination tones" (CTs), are thought to be induced by the non linear activity of the inner ear. Although this perceptive phenomenon is hard to demonstrate, it has been observed in numerous studies dealing with auditory perception (see, *e.g.*, Plomp, 1965). Moreover, CTs could account for the irregularities observed in the masking curves obtained with sinusoidal stimuli (see Sec. 3.2). Thus, several methods were developed to evaluate the conditions in which CTs are present and need to be taken into account (Plomp, 1965; Goldstein, 1967; Ashihara, 2006; Bian and Chen, 2008). One of the best studies is that by Goldstein

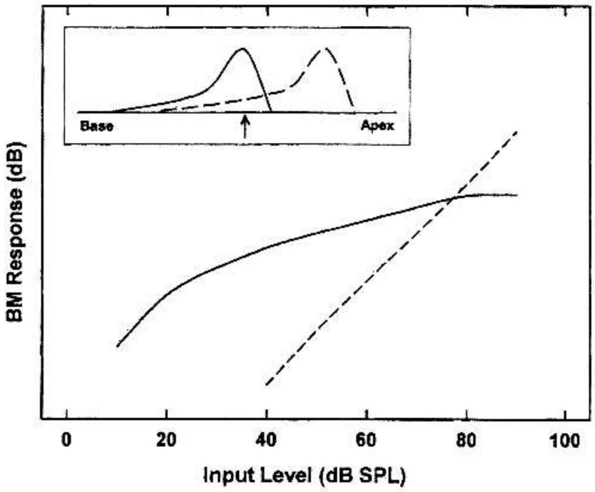


Figure 2.9: Re-plot of the data in Figure 2.8 showing the BM input-output function for a tone at the CF (10 kHz, solid line) and a tone one octave below (5 kHz; dashed line). The insert is an illustration of the traveling wave envelopes for the two tones; the arrow indicates the measurement place. From Oxenham and Bacon (2003).

(1967), who provided results for a huge variety of stimulus conditions. Below is a summary of his results dealing with the most significant properties of CTs. We focus on the properties that are of primary importance in the experimental part of the present PhD work.

Given two simultaneous primary tones of respective frequencies f_1 and f_2 (with $f_1 < f_2$) and levels L_1 and L_2, the most prominent CTs that can be perceived are the difference tone (DT) $f_1 - f_2$, and the cubic difference tone (CDT) $2f_1 - f_2$. The DT is audible only for primary levels above 50 dB SL[1]. Its level is independent of the frequency ratio f_2/f_1, but is proportional to $L_1 + L_2$. The CDT can be perceived for primary levels as low as 15–20 dB SL. Its level, "L_{CDT}", strongly depends on the frequency ratio f_2/f_1. For $f_2/f_1 = 1.1$ and $L_1 \geqslant L_2$, L_{CDT} can be as high as L_1 - 25 dB. Then, L_{CDT} roughly decreases as the ratio f_2/f_1 increases. For example, L_{CDT} can decrease down to 40–60 dB below L_1 for $f_2/f_1 = 1.5$. For ratios $f_2/f_1 < 1.1$, the CDT frequency is too close to the primary frequencies to be resolved by the ear. When $L_1 < L_2$, L_{CDT} is very low and can be neglected in psychoacoustical experiments. Higher order CTs of the form $(f(n) = (n+1)f_1 - nf_2,\ n \in \mathbb{Z})$ have been measured. They appeared to behave in the same manner as the CDT ($n = 1$), but have such low levels that they are of negligible influence.

Although the mechanisms responsible for CTs generation are not fully understood, it is widely accepted that CTs result from cochlear nonlinearities. The fact that the perception of CTs strongly depends on the frequency ratio f_2/f_1 suggests that CTs are generated where frequency analysis takes place (Plomp, 1965; Goldstein, 1967; Smoorenburg, 1972; Fahey and Allen, 1985; Bian and Chen, 2008).

[1] The sensation level (SL), expressed in dB, refers to the level above the absolute threshold of a sound for a listener (see Sec. 3.1).

2.1.4.3 Two-tone suppression

Sachs and Kiang (1968) measured in cats the discharge rates of single auditory nerve fibers when either one or two tones were presented simultaneously. The following procedure was employed: a continuous tone was presented at the CF of a single nerve fiber, and the level of the tone was adjusted so that the fiber was responding at a rate slightly faster than its spontaneous rate. Then, another tone with a sweeping frequency (below and above the CF) was presented simultaneously, and the discharge rate of the fiber was measured as functions of the frequency and level of that "sweep tone". Frequencies of the sweep tone near the CF caused little change in the discharge rate, while sweep tone frequencies away from the CF produced a response rate below that for the continuous tone alone. Then, the authors defined a "response area" and an "inhibitory area" for each fiber. An example is shown in Figure 2.10. Each fiber tested (more than 300 in total) revealed such inhibitory response areas on both sides of the CF. For all fibers tested, the inhibitory areas for sweep frequencies above the CF always covered lower stimulus levels (by down to 10 dB) than the inhibitory areas for sweep frequencies below the CF.

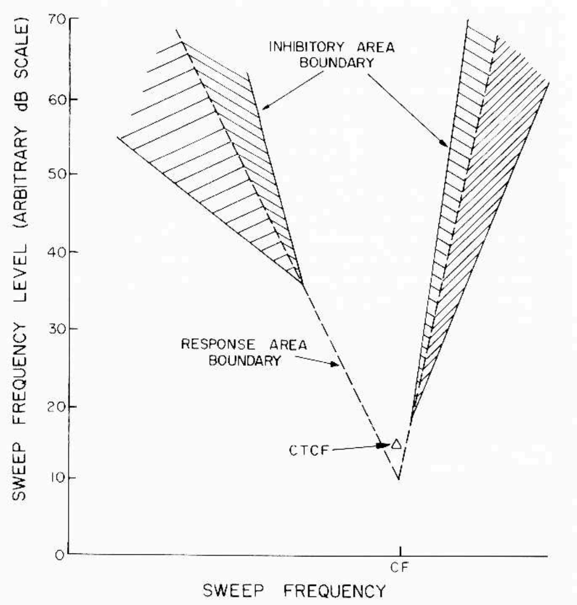


Figure 2.10: Schematic representation of the inhibitory and excitatory response areas of an auditory nerve fiber. The dashed line shows the response area boundary of the fiber in response to a sinusoidal signal. The diamond shows the response to a Continuous Tone set in level near the CF of the fiber (denoted CTCF). The shaded portions represent those frequency-level combinations of the sweep tone that cause a decrease in the response of the fiber. From Sachs and Kiang (1968).

This phenomenon was originally called "two-tone inhibition" by Sachs and Kiang because the authors suggested some neural interactions to account for their results. However, the term "two-tone suppression" is now generally preferred, because the effect does not appear to involve neural inhibition but would rather occur on the BM (Zwicker, 1985; Rhode and Cooper, 1993).

Summary

The peripheral auditory system is composed of the outer, middle, and inner ears. The outer and middle ears act like an amplifier and an impedance matcher, respectively, of the incoming pressure variation. The inner ear contains the specialized receptors of hearing, located in the cochlea. Along the cochlea runs the basilar membrane (BM). Incoming sounds produce traveling waves along the BM. The resulting vibration patterns contain peaks localized at specific places on the BM, related to the frequency content of the input signal. High frequencies produce maximum vibration close to the base, whereas low frequencies produce maximum vibration close to the apex. The mechanical vibrations of the BM are then coded into a series of nervous impulses via hair cells, the sensory receptors located in the organ of Corti on the BM. The nervous impulses are conveyed to the central auditory system via auditory nerve fibers.

This complex functioning of the peripheral auditory system underlies several nonlinearities such as BM compression, combination tones, and suppression.

2.2 Temporal processing

The present section focuses on the temporal processing of sound signals by the auditory system. By "temporal processing" one must distinguish *temporal resolution*, which corresponds to the ability of the auditory system to detect rapid changes in sounds over time, from *temporal integration*, which refers to the improvement of detection threshold as the signal duration increases.

2.2.1 Temporal resolution

Temporal resolution (or temporal acuity) refers to the ability of the auditory system to detect and follow rapid changes in sounds over time. The system must be fast enough to retain the temporal structure of complex sounds such as speech or music, for which a loss of information could affect their perception. Several authors attempted to measure the temporal resolution of the ear. Most of them used gap-detection tasks, and defined temporal resolution as the smallest time interval necessary for a change to be detected. Nevertheless, because a change in the temporal structure of a sound affects its spectrum, there is a difficulty in trying to accurately measure the detection of temporal changes. Indeed, in some cases, listeners are able to use spectral information in order to detect temporal changes (see, *e.g.*, Leshowitz, 1971). Thus, experiments on temporal resolution must be designed to minimize the spectral cues that are available to the listener so that discrimination can only be achieved by the detection of envelope changes. Three different methods can be considered. A first one (phase-detection task) consists in changing the time envelope of the stimuli using phase changes that do not affect their magnitude spectra. A second method (gap-detection task) requires the use of an additional background stimulus to mask the spectral changes. The last method (system analysis), quite different, is based on the modeling of temporal resolution.

The main psychoacoustical results obtained with each of these methods are presented below. Then, the role of peripheral filtering as a limiting factor of temporal resolution is evoked.

2.2.1.1 Psychoacoustical results

In the phase-detection method, listeners must discriminate stimuli pairs with identical durations and magnitude spectra but with different phases. This is achieved by generating the waveform either forward or backward in time. Ronken (1970) was the first to use this method. He asked listeners to discriminate between two pairs of clicks, each click having the same duration (250 μs). In one pair, the first click was higher in amplitude than the second click, and conversely for the second pair. By varying the time interval between clicks in each pair, Ronken found that the pairs were accurately discriminated with a time interval as low as 2 ms. Green (1973) replicated the same procedure except that he used brief sinusoidal pulses at 1, 2, or 4 kHz instead of clicks. He reported a temporal resolution of 1–2 ms, consistent with Ronken's results, with no effect of signal frequency.

The most common method used for measuring temporal resolution is the gap-detection task. It consists in asking the listeners to detect a brief temporal gap in a broad or narrow band of noise (or a sinusoid). The sound portions before and after the gap are often referred to as “markers". In the case of broadband noise, introducing a small temporal gap within the stimulus does not affect its magnitude spectrum, so that spectral cues cannot be used as an indication of the presence of the gap. Plomp (1964) and Penner (1977) obtained gap-detection thresholds of 2–3 ms for a broadband noise level of 30 dB SL, and greater thresholds (5–20 ms) for lower SLs.

When narrowband noise or sinusoidal markers are used, the stimulus spectrum is affected by the presence of the brief temporal gap. To avoid spectral cues, an additional background stimulus is introduced to mask the spectral changes. Usually, a broadband masker with a spectral notch located at the markers' center frequency is used (Shailer and Moore, 1983, 1985, 1987; Eddins et al., 1992). Furthermore, the use of such markers allows to test the dependence of temporal resolution upon frequency, conversely to broadband noise. Gap thresholds for narrowband noise markers were shown to decrease with increasing the markers' center frequency, ranging from 22 ms at 0.2 kHz to 3 ms at 8.0 kHz (Shailer and Moore, 1983, 1985; Eddins et al., 1992). Increasing the noise bandwidth also improved performance. However, with sinusoidal markers at 0.2, 0.4, 1.0, and 2.0 kHz (Shailer and Moore, 1987), gap thresholds were about 5 ms independently of the marker frequency. This contradicts the results with narrowband noise markers.

While the two previous methods attempted to measure temporal resolution in a psychophysical manner, some researchers attempted to model temporal resolution using the system analysis method (Rodenburg, 1977; Viemeister, 1977, 1979). The model of temporal resolution comprises four stages, illustrated in Figure 2.11.

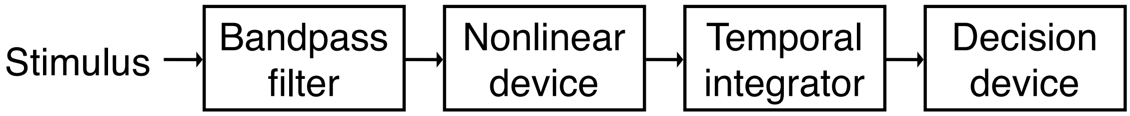


Figure 2.11: Block diagram of the temporal resolution model.

The bandpass filter reflects the action of the peripheral auditory filter (see Sec. 2.3). The nonlinear device reflects the peripheral nonlinearities (see Sec. 2.1.4). The output of the non linear device is linked to a temporal integrator that simulates temporal resolution limitations by removing rapid changes in the envelope of the signal. This third stage can be implemented either as a lowpass filter or as a sliding temporal window. Finally, the decision device is intended to simulate how a listener uses the output of the temporal integrator to make a discrimination in a particular task.

Overall, estimating the time constant of the temporal integrator is equivalent to estimating temporal resolution. Two experimental methods were used to characterize the temporal integrator. In case of a lowpass filter, “temporal modulation transfer functions" (TMTFs, see Viemeister, 1979, for a detailed description of the method) with broadband noise carriers provided resolution estimates of about 3 ms (Rodenburg, 1977; Viemeister, 1977, 1979), consistent with those estimates derived from phase-detection and gap-detection tasks. TMTFs with narrowband noise carriers revealed an improvement in temporal resolution with increasing the noise center frequency (*e.g.*, Viemeister, 1979, obtained resolution estimates of about 6, 5, and 3.5 ms at 0.2, 1.0 and 10.0 kHz, respectively).

In case of a sliding temporal window, Moore et al. (1988) proposed a method to estimate the “shape" of the ear's temporal window. They assumed the temporal integrator to be an intensity weighting function that performs a running average of the incoming auditory stimulus over time. To estimate the shape of this window, threshold was measured for a brief sinusoidal signal presented in a temporal gap between two noise bursts[2]. The window's shape could be best characterized by a sum of two rounded exponential functions. By defining the Equivalent Rectangular Duration (ERD) of the temporal window as the duration of a rectangular window having the same maximum transmission as the estimated window and passing the same total energy of continuous noise, the authors derived an estimation of temporal resolution from each window. The ERD was 8.3 ms at 0.5 kHz and 8.0 ms at 2.0 kHz, indicating no effect of frequency on temporal resolution (see also Plack and Moore, 1990). These estimates are larger than those derived from TMTFs, temporal gap, and phase-detection thresholds (2–4 ms). Moore et al. suggested the nature of the task to account for these discrepancies. As the temporal window slides in time so that its output represents a weighted running average of the energy of “signal plus masker", the listener is assumed to be able to select the “best" temporal window to perform the task. That is, decisions are based on the output of the window at a single instant in time while in other tasks, the outputs of several successive temporal windows may be used to improve performance. This assumption implies the nature of the decision device, the fourth model stage that operates at the output of the temporal integrator (see Fig. 2.11) and whose role falls beyond the scope of the present work. The poorer performance of temporal acuity observed in the Moore et al.'s experiments can also be due to the lesser amount of spectral information available to the listeners to solve the task. Stimuli were indeed chosen so that useful information was available in only one or a small number of auditory filters, while in other experiments using broadband noise or clicks, information could

[2] Note that this method is a replication in the temporal domain of the Patterson's notched-noise method (1976) used to determine the auditory filters' shape (see Sec. 2.3.2).

be combined across several frequency channels.

2.2.1.2 The role of peripheral filtering as a limiting factor of temporal resolution

As described in Section 2.3, the frequency selectivity of the ear can be modeled as a bank of bandpass filters whose bandwidth increases with increasing center frequency (Fletcher, 1940). Considering a bandpass filter, reducing its bandwidth inevitably results in extending its temporal response, or “ringing", so that the filter will continue to oscillate once the input signal has ceased. Thus, the low-frequency filters of the cochlea have a longer temporal response than the high-frequency filters. Shailer and Moore (1983) suggested that this ringing response could explain the improvement of temporal resolution observed at high frequencies in gap detection and TMTFs experiments using narrowband noises as markers or carriers, respectively. However, these results have to be carefully taken in consideration because the use of a narrowband noise introduces inherent random envelope fluctuations. The narrower the bandwidth, the slower the fluctuations. Furthermore, even if the physical bandwidth of the noise is large, the effect of the auditory filter is to filter out a narrow band of noise. This can be very problematic at low frequencies, where the bandwidths of the auditory filters are very small. Thus, in the case of a gap detection task, these fluctuations may have been confounded with the gap to be detected (Shailer and Moore, 1985). In the case of TMTFs, these fluctuations may have affected the detection of the sinusoidal modulation (Viemeister and Plack, 1993).

Summary

Several methods were designed to measure temporal resolution (or temporal acuity) of the human auditory system. Phase-detection and gap-detection methods provided temporal resolution estimates of about 2–3 ms. Comparable estimates were provided by Temporal Modulation Transfer Functions (TMTFs) measurements, while measures of the sliding temporal window provided poorer resolution estimates (8 ms). This discrepancy was accounted for in terms of temporal and spectral information available to the listener to solve the task.

Temporal acuity is independent of stimulus level, at least for levels above 30 dB SL. At lower SLs, temporal resolution tends to deteriorate. Although most of the results revealed that temporal resolution is independent of frequency, it is assumed that resolution mainly depends on two processes: (1) analysis of the time pattern within each frequency channel and (2) comparison of the time patterns across channels.

2.2.2 Temporal integration

Temporal integration refers to the ability of the ear to collect, or integrate, the energy over time in order to improve detection. It is also presented as the "time-intensity trade" (Eddins and Green, 1995). In practice, the integration process is not measured directly. Rather, a decrease in the detection threshold of a sound stimulus as its duration increases is observed. This effect asymptotes at a "critical" stimulus duration, around 500 ms, above which stimulus intensity at threshold is no more dependent of duration. Many authors investigated the relation between hearing threshold and duration for tone pulses or noise bursts, and thereby attempted to model auditory temporal integration as a linear or exponential integration process. More recently, Viemeister and Wakefield (1991) proposed a new approach called "multiple looks". It is based on the idea that threshold decreases with increasing duration because a longer stimulus provides more detection opportunities. This approach is linked to the system analysis model described in Section 2.2.1.

After a brief description of the time-intensity trade, both the linear and exponential models are presented in this section. Then, the "multiple looks" concept is presented and compared to previous models.

2.2.2.1 Short theory on the integration process

The question underlying the time-intensity trade could be the following: why do short signals require more intensity than longer signals to be detected? The frequently evoked reason for such a trade is some kind of accumulation or integration process. Suppose $x(t)$ a real-valued signal as the input of an integration process. The output of the integrator over time is given by the convolution integral according to (see also Eq. (1.6))

$$y(t) = \int_0^{+\infty} x(t-\tau)h(\tau)\,\mathrm{d}\tau \tag{2.1}$$

The integration limits are defined so as to ensure the causality of the system. The choice of the weighting function $h(t)$ depends on the model. For example, consider the simplest form for $h(t)$ associated with the input signal $x(t)$ defined by

$$h(t) = \begin{cases} 1 & 0 < t < \tau \\ 0 & \text{elsewhere} \end{cases} \quad \text{and} \quad x(t) = \begin{cases} S & 0 < t < D \\ 0 & \text{elsewhere} \end{cases} \tag{2.2}$$

where τ is the time constant of the system and the quantity S is proportional to the signal intensity I such that $S = c_0\,I$, c_0 being a constant of proportionality. Let I_0 denote the minimal stimulus intensity required for audibility (*i.e.*, the input x is detected as soon as the model output $y \geq I_0$, see Garner and Miller, 1947). According to Equation (2.1), the integration of $x(t)$ over the time interval τ will produce the quantity $S\tau$. If $D \geq \tau$, the output is $S\tau$ for any signal duration. In this case, the signal intensity at threshold is independent of D because the signal duration has exceeded the time constant of the system. On the other hand, if $D < \tau$, the output is SD. The condition for signal x to be detected then becomes $SD \geq I_0$. Because S is proportional to the signal intensity, as D decreases, the quantity $c_0\,I$ must be increased to achieve a detection threshold. This is the time-intensity trade.

2.2.2.2 Linear integration models

A linear integrator is characterized by the function h in Equation (2.2). Such a model was originally proposed by Garner and Miller (1947), who measured the detection thresholds of pure tones as a function of duration in the presence of a continuous broadband noise[3]. Signal frequencies were 400, 670, 1000, and 1900 Hz. Durations ranged from 12.5 to 2000 ms. Their mean results are displayed in Figure 2.12, who shows the signal-to-noise (S/N) ratio at threshold (in dB) as a function of signal duration (in ms) on a logarithmic scale. For durations below 200 ms, these data clearly show a linear relationship between the S/N ratio and the logarithm of signal duration. Precisely, a doubling in duration results in a -3-dB change in threshold. For durations above 200 ms, the S/N ratio at threshold remains constant. From the least-squares fit to their data (dashed line in Fig. 2.12), Garner and Miller estimated a time constant τ of 200 ms. The authors considered the temporal integration to be complete for durations greater than 1 sec. Note that the data in Figure 2.12 indicate no effect of signal frequency on temporal integration.

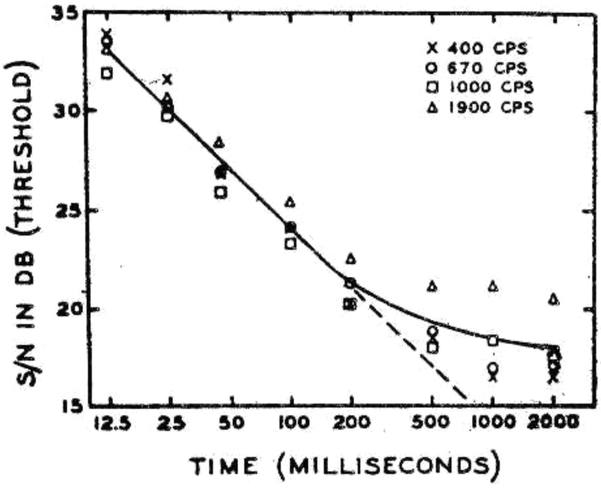


Figure 2.12: Detection thresholds for tone pulses as a function of duration in the presence of a continuous broadband noise. The signal-to-noise (S/N) ratio at threshold (in dB) is plotted as a function of pulse duration (in ms) on a logarithmic scale. Averaged data from four listeners are showed. Each symbol is for a different signal frequency. The solid curve is a visual fit to the data. The dashed line results from a least-squares fit to the 12.5–200 ms data. From Garner and Miller (1947).

Garner (1947) further examined the possible role of frequency on the integration process by measuring detection thresholds in quiet for noise bursts and for 250-, 1000-, and 4000-Hz tone pulses. Durations ranged from 1 to 100 ms (except at 250 Hz where the shortest duration was 4 ms). The author assumed the stimulus energy to be linearly integrated only when all the energy is contained in a narrow band of frequencies. In other words, shortening the signal duration results in a spread of energy over a greater bandwidth, which in turn results in the increase of detection threshold. Because the spectral splatter produced by a very brief tone

[3] Plomp and Bouman (1959) showed that the characterization of the temporal integrator is not affected by the presence of the background noise.

resembles that produced by a noise burst, if the Garner's hypothesis is correct, then the thresholds of noise and tones should coincide at very short durations. This was indeed the case: the thresholds obtained at 1000 (durations < 4 ms) and 4000 Hz (durations < 2 ms) were similar to those obtained with the noise burst. At 250 Hz, the thresholds for the shortest pulses were higher than that for the noise. This could be explained by the spread of energy over a much less sensitive region of the audiogram (see Sec. 3.1). At 1000 Hz, Garner estimated a "critical bandwidth" for perfect integration (*i.e.*, a linear integrator producing a slope of -3 dB per doubling of signal duration, as can be seen in Fig. 2.12) around 175 Hz, which corresponds to 3–4 times the auditory filter's bandwidth centered at 1000 Hz (see Sec. 2.3.3).

In a more recent study, Florentine et al. (1988) measured detection thresholds in quiet for tone pulses of various durations at 250, 4000, and 14000 Hz. Durations ranged from 5 periods to 500 ms. Their results are in good agreement with those from Garner (1947). Because they did not test signal durations greater than 500 ms, Garner (1947) and Florentine et al. (1988) could not estimate the time constant of the linear integrator.

2.2.2.3 Exponential integration models

Based on physiological data on auditory nerve responses (see Lüscher and Zwislocki, 1949; Plomp, 1964; Smith and Zwislocki, 1975 and more recent works by Sumner et al., 2003; Meddis and O'Mard, 2005) showing the exponential decay of the neural excitation produced by an auditory stimulus, Plomp and Bouman (1959) assumed temporal integration data to be better characterized by an exponential integrator than by a linear one. More explicitly, Plomp and Bouman assumed that stimulation by a pulse of intensity I (*e.g.*, signal $x(t)$ defined in Eq. (2.2)) will result in an effect S somewhere in the hearing pathway. This stimulation will decrease exponentially down to an asymptotic value being proportional to I, and signal detection will occur when S exceeds the minimal intensity I_0. An exponential integrator implies the following convolution function h

$$h(t) = \begin{cases} e^{-t/\tau} & t > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{2.3}$$

If signal $x(t)$ is presented at the input of the integrator, the output can be expressed as a function of signal duration D (according to Eq. (2.1))

$$y(D) = S\tau\,(1 - e^{-D/\tau}) \tag{2.4}$$

Reminding that $S = c_0\, I$ (see Eq. (2.2)), the time constant τ in Equation (2.4) can be included in the constant of proportionality such that $S\tau = c_0\, I$. Finally, if we denote by I_∞ the intensity of an infinite pulse, the Plomp and Bouman's hypothesis assumes the asymptotic value of the exponential decay to be equal to $c_0\, I_\infty = I_0$. Therefore, the signal detection occurs when $y(D) = c_0\, I(1 - e^{-D/\tau}) = c_0\, I_\infty$. This leads to the following relation

$$\frac{I}{I_\infty} = \frac{1}{1 - e^{-D/\tau}} \tag{2.5}$$

which gives the intensity at threshold for a pulse of duration D above the intensity threshold for the detection of an infinite pulse (*e.g.*, a continuous tone).

To verify their hypothesis, Plomp and Bouman measured the detection thresholds of tone pulses as a function of duration in the presence of a continuous broadband noise. Signal frequencies ranged from 0.25 to 8.0 kHz. Durations ranged from 4–8 periods of the carrier frequency to 10 sec. For each frequency, the noise level was adjusted so that the threshold of a continuous tone of same frequency were 40 dB SL. The results from one "typical" listener are showed in Figure 2.13. For each frequency, the masked threshold (in dB) above the threshold for the 10-sec tone pulse (see Eq. (2.5)) is plotted as a function of the pulse duration (in ms) on a logarithmic scale.

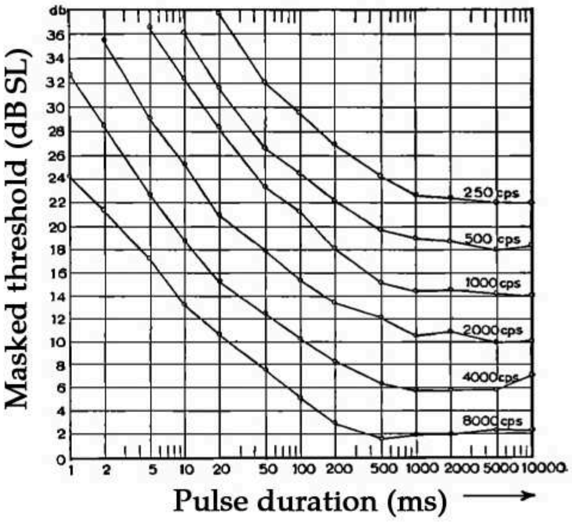


Figure 2.13: Masked thresholds of tone pulses (in dB SL) as a function of duration (in ms) for one "typical" listener. The masker was a continuous broadband noise whose overall level was adjusted at each frequency so as to mask a continuous tone of same frequency by 40 dB. Each curve is for a different frequency. For each frequency, the masked threshold *above the threshold for the 10-sec tone pulse* is indicated (in dB). From Plomp and Bouman (1959).

For durations below 500–1000 ms (depending on signal frequency), all curves in Figure 2.13 show a decrease in threshold with increasing pulse duration. Precisely, a 3-dB decrease in threshold per doubling of duration is observed, consistent with previous data from Garner and Miller (1947). However, the decrease in threshold is more rapid at very short durations (*i.e.*, < 5 periods). This is attributed to the spread of energy over several auditory filters, in reference to the hypothesis by Garner (1947). To discard this short-duration effect when deriving the integration time constants τ from the curves in Figure 2.13, Plomp and Bouman excluded all data points with durations ≤ 10 periods. Estimates ranged from $\tau = 375$ ms at 0.25 kHz to 150 ms at 8.0 kHz.

Further investigations (*e.g.* Zwislocki, 1960; Hempstock et al., 1964) proposed exponential integration models. In the latter study, the time constant τ was frequency-dependent, ranging from about 100 ms at 0.25 kHz to 200 ms at 10 kHz. In the former study, however, τ was estimated to about 200 ms for all the tested signal frequencies. Although the Garner's hypothesis could account for most of the temporal integration data for very short signal durations, the role of frequency on temporal integration for long-duration signals remains a controverted question.

Zwislocki (1960) examined the mechanisms underlying temporal integration. Neural adaptation, which refers to the decrease in the firing rate of neural discharges that occurs during a prolonged stimulation (see, *e.g.*, Lüscher and Zwislocki, 1949; Palmer, 1995; Smith and Zwislocki, 1975), and the functional relationship between stimulus intensity and neural excitation, appeared to be the most important underlying factors.

2.2.2.4 The “multiple looks" approach

Unlike the models described above that dealt with a long-term temporal integration process of the order of hundreds of milliseconds, a totally different approach was proposed by Viemeister and Wakefield (1991). They assumed the decrease in detection threshold with increasing signal duration to not result from a single observation-window integrator with a time constant of 100–300 ms, but rather to a decision process involving multiple short time-constant integration windows. According to this approach, the internal representation of a signal is viewed as a vector of samples, or “looks", that will be used by the listener to solve a task, whatever the latter is. For each look presented at the input, there is a certain probability that the detection threshold be reached. An increase in signal duration increases the number of looks processed. The greater the number of looks, the higher the probability that the threshold be reached, thus the lower the threshold. To provide evidence to their hypothesis, Viemeister and Wakefield conducted two experiments.

In the first experiment, the detection thresholds in quiet for a pair of brief 1-kHz tone pulses were measured as a function of the temporal separation between the pulses and compared to the threshold for a single pulse. For between-pulse separations up to 5 ms, thresholds increased with increasing the between-pulse distance. For larger separations, thresholds varied little. However, for all the tested separations, thresholds for the pulse pair were always lower than that for a single pulse. This indicates that the pulses are processed as if they were independent looks and that the listener combined information from several looks.

To confirm the independence of looks, a second experiment was conducted that investigated the detection thresholds for either a single tone pulse or two tone pulses (both pulses had a duration of 10 ms) separated by 100 ms and presented in a broadband noise. The temporal envelope of the noise contained 10-ms gaps at the times when the pulses might occur. Thresholds were measured as a function of the noise level in the between-pulse interval. The results indicated that, again, thresholds for the pulse pair were systematically lower than that for a single pulse. Furthermore, thresholds were independent of noise level. These results support the multiple-looks hypothesis but are inconsistent with a long-term temporal integration process.

The results from these experiments allowed to estimate the short-time constant of the multiple looks approach to about 3 ms, which would correspond to the critical duration of an independent “time sample". The combination of the available information at the output of each look finally leads to a decision process that will result, or not, in the detection of the auditory stimulus. Making an analogy with the system analysis approach described in Section 2.2.1, the multiple looks model allowed to accurately predict temporal resolution and temporal integration data for various stimulus types.

Summary

The detection threshold of a sound signal decreases as the signal duration increases. This is referred to as temporal integration. This effect asymptotes at a "critical" duration (about 500 ms) above which no further decrease in threshold is observed. Considering the auditory system as an energy integrator, both linear and exponential integration models with time constants in the range 100–300 ms provided good fits to experimental data, at least for signal durations $>$ 10 ms. For shorter durations, the deviation between models and data was attributed to the splatter of spectral energy across several adjacent auditory filters.
More recently, Viemeister and Wakefield (1991) proposed a different approach called "multiple looks". The authors assumed the decrease in detection threshold with increasing signal duration to result from a decision process involving multiple short time-constant (estimated to about 3 ms) integration windows, or "looks". Regarding the incoming sound as a vector of looks, its detection finally depends on the number of available looks.

2.3 Frequency processing

The present section focuses on the frequency processing of sounds by the human auditory system. Frequency processing refers to the ability of the ear to resolve the sinusoidal components of complex sounds, so as to behave like a Fourier analyzer. As mentioned in Section 2.1, frequency selectivity starts in the cochlea with the tonotopic organization of the BM, and follows up to higher levels of the auditory system with the nerve fibers' CF and the associated neural networks. Commonly, the frequency analysis of the peripheral auditory system is considered as a bank of bandpass filters, named "auditory filters". Many studies attempted to estimate the characteristics of these filters using different approaches, presented below. It is important to mention here that frequency resolution plays a major role in many aspects of auditory perception, especially in masking. Indeed, as shown in Chapter 3 auditory frequency masking reflects the limits of frequency selectivity.

2.3.1 The critical band concept

The first experimental measurement of frequency analysis was proposed by Fletcher (1940) who measured the detection threshold of a sinusoidal target as a function of the bandwidth of a noise masker. The noise was always centered at the target frequency and the noise power density was held constant so that the total noise power increased as its bandwidth broadened. The results, plotted in Figure 2.14, indicate that the target detection threshold (expressed as the ratio of target power at threshold P_T to noise power spectral density N_0) first increases as the noise bandwidth increases, up to a critical value above which threshold asymptotes so that a further increase in the noise bandwidth does no more affect threshold. The critical noise bandwidths, indicated by the intersections between the horizontal and the 45-degree lines, increase with increasing target frequency.

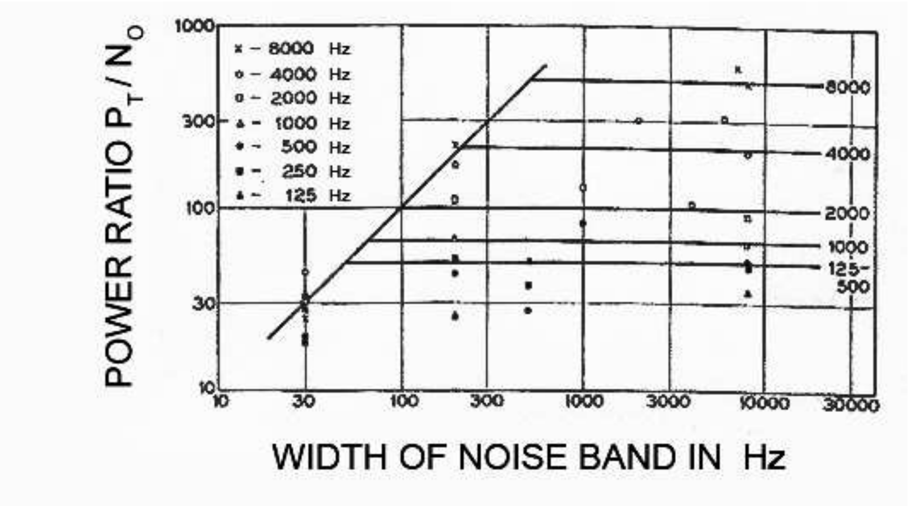


Figure 2.14: Ratio of target power at threshold (P_T) to noise power spectral density (N_0) as a function of noise bandwidth, for different frequency regions. The target frequency corresponds to the noise center frequency. For each frequency region, the power ratio increases as the noise bandwidth broadens until the critical bandwidth is reached. The intersections between the horizontal and the 45-degree lines provide some approximations of the critical bandwidths. From Fletcher (1940).

To account for these results, the author suggested that the peripheral auditory system can be modeled by a bank of bandpass filters whose bandwidth increases as a function of their center frequency. These filters are commonly named “auditory filters”. The listener is assumed to detect the target by attending to the output of a single auditory filter centered at the target frequency, such that the detection threshold correspond to a certain target-to-noise ratio at the output of the filter. Only the noise components passing through that filter will have an effect in masking the target. Once the noise bandwidth exceeds that of the auditory filter centered at the target frequency, a further broadening of the bandwidth will not result in an increase of the target detection threshold. Fletcher called the bandwidth at which the target threshold ceased to increase the “Critical Band” (CB). Because the stimuli are represented by their long-term spectra (the relative phases of the components and the fluctuations of the noise are ignored), the model proposed by Fletcher is commonly referred to as the “power spectrum model” (Patterson and Moore, 1986).

By relating the CB to the auditory filters concept, Fletcher simply assumed that the shape of the auditory filters could be approximated by simple rectangles with a flat top, vertical edges, and a bandwidth equal to that of the CB. Given the assumptions of the power spectrum model, the value of the CB can be estimated indirectly by measuring the required power of a sinusoidal target for being detected in the presence of a broadband noise. N_0 (expressed in power unity per Hz) is indeed constant and independent of frequency for a white noise. The total noise power passing through a CB-Hz-wide critical band is then $N_0 \times CB$. As the model assumes that P_T is directly proportional to the total noise power passing through the CB, a constant of proportionality K can be introduced such that

$$P_T = K \times N_0 \times CB \tag{2.6}$$

By measuring P_T and N_0 and estimating K, the value of CB can be evaluated. Fletcher simply fixed $K = 1$ so that the critical target-to-noise ratio P_T/N_0 numerically equals the CB. From the data in Figure 2.14, the CB width was estimated to about 40 Hz at 250 Hz, 60 Hz at 1000 Hz, and 150 Hz at 4000 Hz. However, further experiments showed that K rather equals 0.4 (Scharf, 1970) and varies with center frequency (Greenwood, 1961a; Patterson and Moore, 1986).

2.3.2 Estimating the shape of the auditory filters

When Fletcher introduced the CB concept (1940), he was fully aware that his simple assumption of the rectangular-shaped filter was not realistic. Nevertheless, he provided an experimental method for measuring frequency selectivity which was replicated in several other experiments (see, *e.g.*, Greenwood, 1961a; Swets et al., 1962; Scharf, 1970; Zwicker and Feldtkeller, 1999, Chap. 6). Those experiments led to more accurate estimates of the CBs and their parameters. Furthermore, the power spectrum model is based on several assumptions that turned out to be inaccurate. First, listeners can combine information from more than one auditory filter in order to enhance target detection, so that masker components falling outside the filter centered at the target frequency can facilitate the detection of the target. This ability to detect a signal through a filter that is not centered at the signal frequency is commonly referred to as "off-frequency listening" (Leshowitz and Wightman, 1971). Second, the slow fluctuations inherent to the use of narrow bands of noise may also affect target detection (*e.g.*, Shailer and Moore, 1985). Thus, on the basis of the Fletcher's model, some other methods had to be designed in an attempt to estimate a more realistic and more accurate shape for the auditory filters.

2.3.2.1 Psychophysical tuning curves

A standard physiological method for measuring frequency selectivity is the tuning curve, presented in Section 2.1.3 (see Fig. 2.7). It measures the response of a single auditory nerve fiber as a function of the frequency and intensity level of a tone. The frequency requiring the lowest tone level to produce a response of the fiber corresponds to the CF of that fiber.

Based on the physiological method of tuning curves, a psychophysical method was developed to estimate the shape of the auditory filters: the psychophysical tuning curves (PTCs). This method consists in presenting to the listener a target tone of fixed level and frequency and measuring the power that a second tone must have to mask the target as a function of the frequency of the masking tone. It is assumed that if the target has a sufficiently low level (*e.g.*, 10 dB SL), it will excite only a small group of neurons with similar CFs so that the PTCs and the physiological tuning curves have basically the same shape. Moreover, all maskers (*i.e.*, different signal frequency and level combinations) that produce the same amount of masking are assumed to produce the same amount of activity in the target frequency region. Then, by inverting the shape of the PTCs, one would get a simple estimation of the auditory filters shape. Examples of PTCs are given in Figure 2.15. These curves were obtained in simultaneous masking conditions with sinusoidal stimuli. Although the tuning curves are markedly asymmetric when plotted on a logarithmic frequency scale, they are much more symmetric when replotted on a linear frequency scale. The symmetry allows to consider each curve as a first-order notch filter, and then facilitates the estimation of the -3-dB bandwidth.

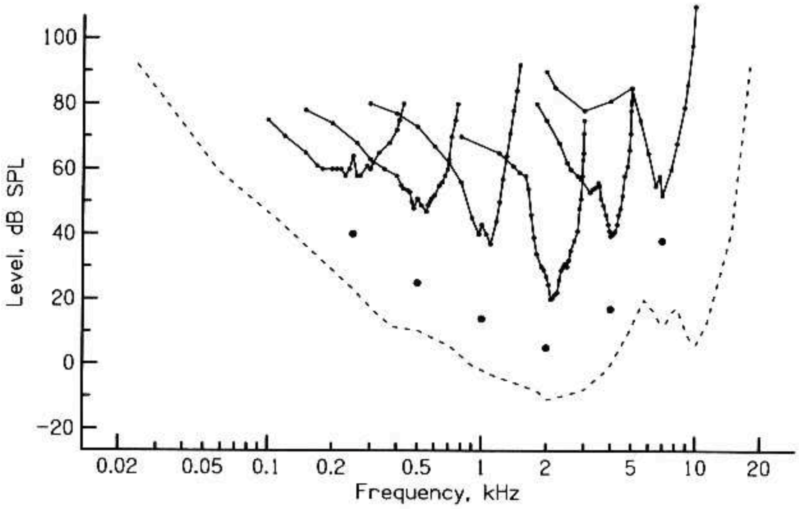


Figure 2.15: Psychophysical tuning curves (PTCs) determined in simultaneous masking with sinusoidal stimuli. Targets had a fixed level of 10 dB SL. The masker level required for threshold is plotted as a function of masker frequency on a logarithmic scale. The filled circles below each curve indicate the frequency and the SPL of the target at threshold. The dashed line shows the threshold of the target in quiet. Adapted from Vogten (1978a).

The PTCs plotted in Figure 2.15 are very similar to the physiological curves presented in Figure 2.7, which suggests that (1) the frequency selectivity of the auditory system is established at the level of the auditory nerve, and (2) the shape of the auditory filters mimics that of the tuning curves. However, the analogy between the PTCs and the auditory filters' shape cannot be established so straightforwardly for two reasons. The first is that even if PTCs are measured at a very low target level, several primary neurons with close CFs are excited. Instead, physiological tuning curves are truly derived from the responses of single neurons. Thus, in the method of PTCs, the listener may focus his attention on the output of an auditory filter that is not centered at the target frequency (off-frequency listening). The second reason concerns the stimulus presentation. Because masker and target are presented simultaneously, they may interact so as to produce beats or combination tones that the listener can use as a cue to the presence of the target (Ehmer, 1959b; Goldstein, 1967; Moore et al., 1998; Ashihara, 2006). The simultaneous presentation of masker and target also implies the suppression effect described in Section 2.1.4.

Overall, PTCs do not provide an accurate estimation of the auditory filters' shape, unless some experimental restrictions are used to minimize the effects of masker-target interactions and off-frequency listening. Some recommendations are given in Patterson and Moore (1986, Chap. 3, Sec. 8).

2.3.2.2 The notched-noise method

In order to prevent off-frequency listening and possible masker-target interactions from biasing the estimation of the auditory filters' shape, Patterson (1976) designed a new method, illustrated in Figure 2.16. The threshold of a sinusoidal target is determined as a function of the width of a spectral notch inserted in a broadband noise masker. The target frequency (f_0) is fixed and the notch band is $2\Delta f$ wide and centered at f_0. Because the notch is placed symmetrically around f_0, the method cannot reveal asymmetries in the auditory filters.

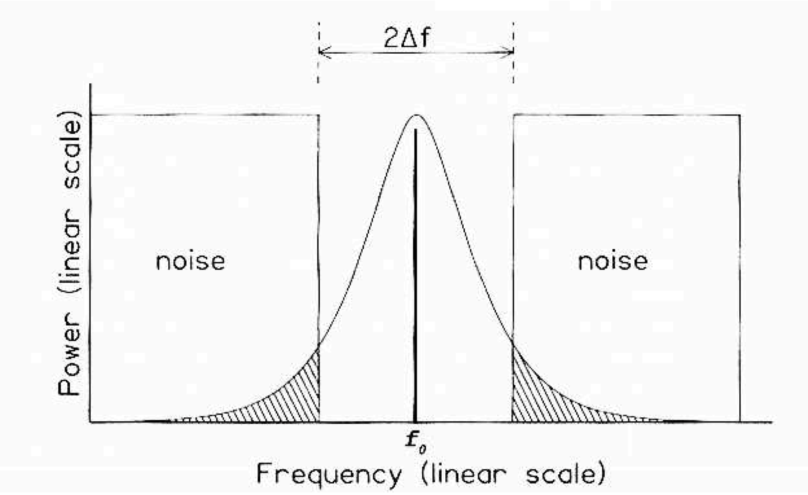


Figure 2.16: Schematic illustration of the notched-noise method used by Patterson (1976) to estimate the auditory filters' shape. The threshold of a sinusoidal target of fixed frequency (f_0) is measured as a function of the width of a spectral notch in the noise masker. The amount of noise passing through the auditory filter centered at f_0 is proportional to the hatched portions.

The shape of the auditory filters is estimated based on the power spectrum model described in Equation (2.6). The model becomes

$$P_T = \underbrace{K N_0 \int_{-\infty}^{f_0-\Delta f} |W(f)|^2 \, \mathrm{d}f}_{\mathbf{A}} + \underbrace{K N_0 \int_{f_0+\Delta f}^{+\infty} |W(f)^2 \, \mathrm{d}f}_{\mathbf{B}} \tag{2.7}$$

where $W(f)$ is the transfer function of the auditory filter and the two integrals **A** and **B** represent the hatched portions in Figure 2.16. Because the notch is placed symmetrically around f_0, the terms **A** and **B** are equal. Equation (2.7) is then reduced to

$$P_T = 2\,K N_0\, f_0 \int_{g}^{+\infty} |W(g)|^2 \, \mathrm{d}g \tag{2.8}$$

where g is the deviation in frequency from f_0, defined by $g = \frac{|f-f_0|}{f_0}$. Equation (2.8) shows that the function relating the target power at threshold to the notch width provides a measure of the integral of the auditory filter. At any value of g, the height of the filter is given by the slope of the threshold curve at the corresponding notch width. That is, an analytic expression of the shape of the auditory filter can be obtained by deriving Equation (2.8) with respect to g

$$|W(g)|^2 = -\left(\frac{1}{2KN_0f_0}\frac{\mathrm{d}P_T}{\mathrm{d}g}\right) \tag{2.9}$$

A typical auditory filter derived using this method with $f_0 = 1$ kHz and $N_0 = 40$ dB/Hz is shown in Figure 2.17. The relative response — or gain (defined as output level - input level) — of the filter is plotted as a function of target frequency. Unlike a rectangular filter, the filter has a rounded top and shallow slopes. Because such a shape cannot be characterized with a single number like the CB, the filters are characterized with their -3-dB bandwidths, which typically vary between 10 and 15% of the center frequency.

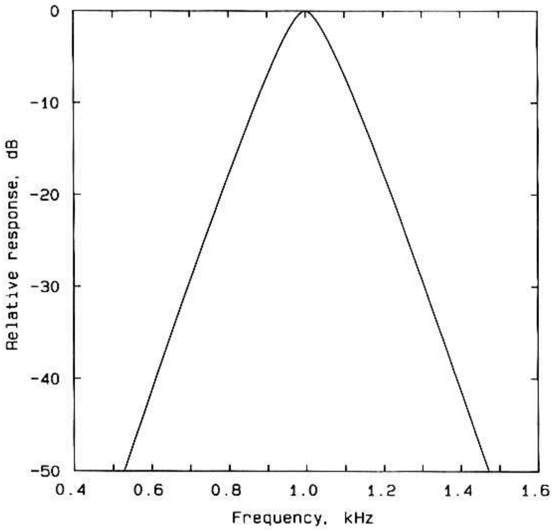


Figure 2.17: A typical auditory filter's shape determined using the notched-noise method (Patterson, 1976) with $f_0 = 1$ kHz and $N_0 = 40$ dB/Hz. The relative response (or gain) of the filter, defined as the output level minus the input level (in dB), is plotted as a function of frequency (in kHz). The gain of the filter is assumed to be 0 dB at its tip.

Patterson et al. (1982) suggested a rounded-exponential function to describe the auditory filters' shape such as that presented in Figure 2.17. Any filter can be computed by $W(g) = (1 - r)(1 + pg)e^{-pg} + r$, where p determines the shape and width of the filter, and r controls the shallow tail section of the filter outside its passband. This function is commonly referred to as the "*Roex(p,r)* filter".

2.3.3 Characteristics and origin of the auditory filters

The notched-noise method revealed to be the most accurate and convenient method to estimate the auditory filters' shape, provided the auditory filters can be assumed to be symmetric (Patterson and Moore, 1986). This method was reproduced several times by varying the target frequency and the notched-noise masker level so as to evaluate their influence on the filters' shape. The potential filters' asymmetry was also tested experimentally. The influence of these parameters on the auditory filters' shape are discussed below. The present section concludes with a discussion on the possible origin of the auditory filters.

2.3.3.1 Effect of center frequency

The original notched-noise experiment of Patterson (1976) provided bandwidth estimates for the filters centered at 0.5, 1.0, and 2.0 kHz. Since then, the filters' bandwidths have been estimated for a large variety of center frequencies using the symmetric notched-noise method (Patterson et al., 1982; Moore and Glasberg, 1983b; Shailer and Moore, 1983; Lutfi and Patterson, 1984; Glasberg and Moore, 2000). All bandwidth estimates were in agreement with the first Fletcher's observation: the bandwidth increases with increasing center frequency. In order to characterize the filters with a single value like in the CB concept by Fletcher (1940), an alternative measure of bandwidth was proposed: the equivalent rectangular bandwidth (ERB). The ERB of a given filter is equal to the bandwidth of a perfect rectangular filter which has a transmission in its passband equal to the maximum transmission of the specified filter and transmits the same power of white noise as the specified filter. The ERB of the auditory filter centered at frequency F, "ERB_F", is (Glasberg and Moore, 1990)

$$ERB_F = 24.7\,(4.37F + 1) \tag{2.10}$$

with F specified in kHz. The resulting value of ERB_F is in Hz. The stars in Figure 2.18 represent the ERB_F values resulting from notched-noise data (Moore and Glasberg, 1983b). The latter are well fitted with the solid curve, which shows the ERB_F function according to Equation (2.10). The dashed curved in Figure 2.18 represents the CB function proposed by Zwicker (1961).

As it is sometimes useful to plot psychoacoustical data on a frequency scale related to CBs, Glasberg and Moore (1990) defined an ERB scale based on the original Bark scale proposed by Zwicker (1961). The ERB scale relates an ERB_F *number* to the auditory filter centered at F according to

$$ERB_F\ \text{number} = 21.4\ \log(4.37F + 1) \tag{2.11}$$

with F in kHz. For example, the auditory filter centered at $F = 4$ kHz has a width of 0.456 kHz and its ERB_4 number is 21.1. Thus, an increase in frequency from 3772 Hz to 4228 Hz represents a step of one ERB.

2.3.3.2 Effect of level

If the auditory filters were linear, then their shapes would not vary with the level of the noise used to measure them, as it is the case. In fact, at low input levels ($\leq$ 30 dB SPL), the filters are sharply tuned at their tips. With increasing input level, the filters become broader (Lutfi and Patterson, 1984; Patterson and Moore, 1986; Glasberg and Moore, 1990, 2000). For input levels above 50 dB SPL, the filters become asymmetric: the left-hand side of the filters' response (relative to the center frequency) becomes shallower while the right-hand side becomes slightly

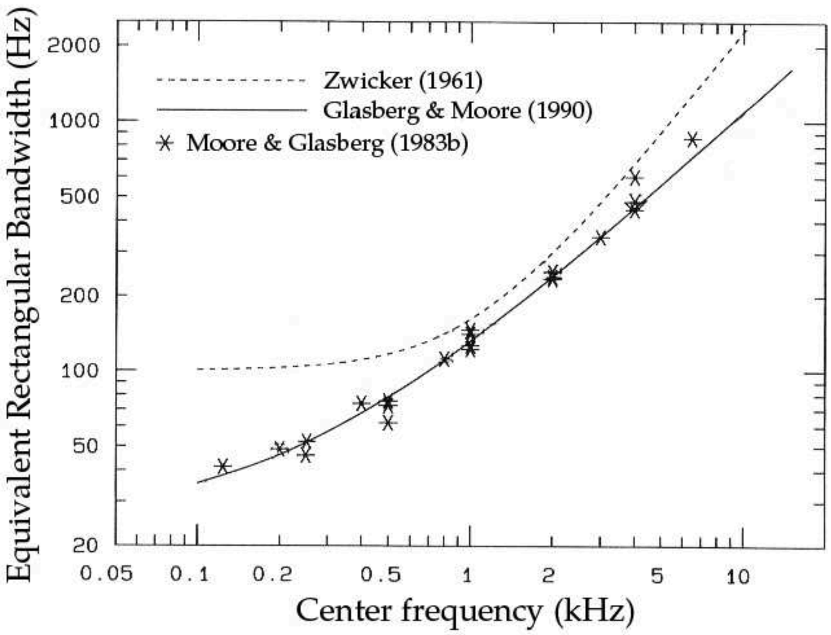


Figure 2.18: ERB_F (in Hz) as a function of center frequency F (in kHz). The solid curve shows the ERB_F function according to Equation (2.10). The stars represent the ERB_F values derived from notched-noise data (Moore and Glasberg, 1983b). The dashed curve shows the CB function proposed by Zwicker (1961).

steeper. Figure 2.19 shows how the shape of the auditory filter centered at 4 kHz varies with input level. These filter shapes were derived from notched-noise data (Glasberg and Moore, 2000). Input levels ranged from 30 to 80 dB SPL in 10-dB steps. The filter responses are plotted as normalized gain (*i.e.*, to have a 0-dB gain at the tip for all input levels). Figure 2.19 clearly exhibits the filters' asymmetry that appears for levels above 50 dB.

2.3.3.3 Origin of the auditory filters

Although the physiological basis of the auditory filters is still uncertain, the frequency analysis observed on the BM is most likely involved (see Sec. 2.1.2). There are indeed many similarities between the frequency selectivity measured on the BM and the frequency selectivity measured psychophysically. The ERBs of the auditory filters roughly correspond to a constant distance along the BM. In humans, each ERB corresponds to about 0.9 mm, regardless of the center frequency (Greenwood, 1961b, and Moore, 2003, Chap. 3, Sec. 6).

Furthermore, the nonlinearities observed in the filters' shape when measured as a function of input level resemble the level dependence of BM filtering (Glasberg and Moore, 2000). Reminding that the input-output function of the BM is linear for frequencies well below the CF but compressive for frequencies close to the CF (see Sec. 2.1.4), the data presented in Figure 2.19 can be reasonably explained in terms of BM compression. The gain for input frequencies well below the center frequency is invariant with level (linear portion on the BM input-output function), while the gain at the center frequency decreases with increasing input level (compressive portion).

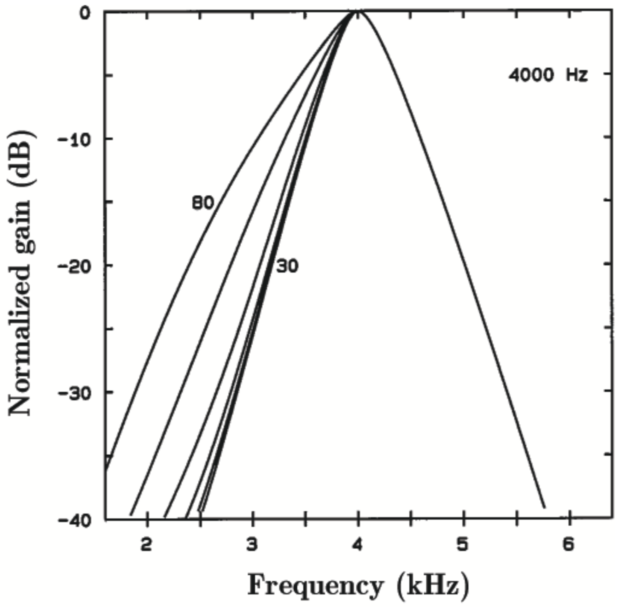


Figure 2.19: The shape of the auditory filter centered at 4 kHz obtained for input levels ranging from 30 to 80 dB SPL, in 10-dB steps. The output of the filter is expressed as normalized gain (*i.e.*, to have a 0-dB gain at the tip for all input levels). From Glasberg and Moore (2000).

Summary

The peripheral auditory system can be modeled as a bank of bandpass filters, named the “auditory filters” or “critical bands” (CBs), whose equivalent rectangular bandwidth (ERB) increases with increasing center frequency.
Based on the power spectrum model, which assumes that the detection threshold of a target signal in a band of noise is proportional to the target-to-noise ratio at the output of the auditory filter centered at the target frequency, the notched-noise method allowed to best approximate the auditory filters shape as a rounded exponential function. The filters shape is sharply tuned and symmetric at low input levels (< 40 dB SPL), but tends to broaden on the low-frequency side (relative to the center frequency) at higher input levels so as to become asymmetric.
The similarities between the frequency selectivity measured on the BM and the frequency selectivity measured psychophysically support the idea that the auditory filters have their origin in the BM.

2.4 Summary

The human auditory system and its related temporal *and* frequency processing of sound signals were described in this chapter. This description mainly concerned the peripheral auditory system, which is composed of the outer, middle, and inner ears. The outer and middle ears act like an amplifier and an impedance matcher, respectively, of the incoming pressure variation. The inner ear contains the specialized receptors of hearing, located in the cochlea. Along the cochlea runs the basilar membrane (BM). Incoming sounds produce traveling waves along the BM. The resulting vibration patterns contain peaks localized at specific places on the BM, related to the frequency content of the input signal. High frequencies produce maximum vibration close to the base, whereas low frequencies produce maximum vibration close to the apex. The BM thus behaves like a Fourier analyzer, or a filter bank. Movements of the BM cause some displacements of the stereocilia at the top of the hair cells, the sensory receptors located in the organ of Corti on the BM. These displacements initiate action potentials within the auditory nerve fibers. The mechanical vibrations of the BM are then coded into a series of nervous impulses to be processed by the central auditory system. This complex functioning of the auditory system underlies several nonlinearities such as BM compression, combination tones, and suppression.

By temporal processing, one must distinguish *temporal resolution*, which refers to the ability to detect rapid changes in sounds over time, from *temporal integration*, which refers to the ability to improve the detection of a sound signal as its duration increases. Temporal resolution is commonly measured using temporal-gap detection tasks. Most of the investigations provided estimates of about 2–3 ms. Comparable estimates were provided with other methods such as phase detection or temporal modulation transfer functions (TMTFs). In an attempt to estimate the shape of the ear's temporal window, Moore et al. (1988) provided a resolution estimate of about 8 ms. The poorer temporal acuity observed in the latter case was explained by the lesser amounts of temporal and spectral informations available to the listener to solve the task.

Temporal integration is usually characterized by measuring the detection threshold of a tone pulse as a function of its duration. Increasing the pulse duration results in a decrease of its detection threshold. However, this effect asymptotes at a "critical" duration (around 500 ms) above which threshold is no more dependent of signal duration. Both linear and exponential integration models with time constants ranging from 100 to 300 ms were proposed. A different approach was proposed by Viemeister and Wakefield (1991), called "multiple looks". The authors assumed the decrease in detection threshold with increasing signal duration to result from a decision process involving multiple short time-constant (estimated to about 3 ms) integration windows, or "looks". Regarding the incoming sound as a vector of looks, its detection finally depends on the number of looks available, *i.e.*, on its duration.

Frequency processing refers to the ability of the ear to resolve the sinusoidal components of complex sounds. The early experiments by Fletcher (1940) allowed to characterize the peripheral auditory system as a bank of bandpass filters, named auditory filters, whose equivalent rectangular bandwidth (ERB) increases with increasing center frequency. Different methods were proposed to estimate the shape of these filters, such as the psychophysical tuning curves (PTCs) or notched-noise methods. The latter allowed to describe the auditory filters' shape with a rounded exponential function, commonly referred to as the *Roex(p,r)* filter. Given a center frequency, the parameters p and r determine the width of the filter and the shallow tail section outside its passband, respectively. The filters are non linear; their shapes vary with the input level. At low input levels (< 30 dB SPL), the filters are narrow and symmetric. With increasing level, the filters broaden and tend to become asymmetric (levels $>$ 50 dB SPL). The similarities between the frequency selectivity measured on the BM and the frequency selectivity measured psychophysically support the idea that the auditory filters have their origins on the BM. Furthermore, the nonlinearities observed in the filters' shape when measured as a function of input level resemble the level dependence of filtering observed on the BM.

In this chapter, the temporal and frequency processing have been treated separately. However, the psychoacoustical results presented above showed that the temporal and frequency processes strongly interact. *What about the joint time-frequency analysis of the human auditory system?*

A few authors attempted to characterize the *time-frequency* processing of sounds by the ear. Van den Brink and Houtgast (1990) studied the influence of the spectro-temporal pattern of a Gaussian signal on its detection threshold so as to derive an "efficient spectro-temporal integration for signal detection". They found that efficient temporal integration requires that the total signal bandwidth should be confined to a narrow "critical frequency window" of about one CB. Interestingly, this finding supports the Garner's hypothesis (1947, see Sec. 2.2.2). Similarly, the efficient spectral integration requires that the total signal duration should be confined to a narrow "critical time window", whose length depends on the signal spectrum. For a Gaussian signal with a bandwidth of 3 octaves centered around 1.6 kHz, van den Brink and Houtgast estimated a critical length of 30 ms. More recently, van Schijndel et al. (1999) attempted to derive the shape of the "internal" time-frequency window of the auditory system by assessing just-noticeable differences in intensity for Gaussian stimuli with various spectro-temporal shapes. They hypothesized that the intensity discrimination threshold should reach a maximum when the spectro-temporal pattern of the signal covers the smallest numbers of internal time-frequency windows. They found that the spectral width of one window is roughly a CB. The temporal width approximately corresponded to four periods of the carrier frequency, *e.g.*, 4 ms at 1 kHz and 1 ms at 4 kHz. These results served as a basis for the design of the stimuli used in the present PhD work. Therefore, the study by van Schijndel et al. is more deeply considered in Chapter 5.

Chapter 3

Auditory masking

Contents

Auditory masking refers to the fact that the presence of sound signal (referred to as the "masker" in psychoacoustics) impedes or decreases the detection threshold of another sound (the "target"). When masker and target are presented simultaneously, auditory masking is characterized in the frequency domain ("frequency masking"). When masker and target are presented non simultaneously, masking is characterized in the temporal domain ("temporal masking"). In this chapter, the main psychoacoustical results obtained for each domain are reviewed, and the physiological mechanisms of masking are evoked.

3.1 Definition of quiet and masked thresholds

The detection threshold in quiet, or "absolute threshold" of a sound is the minimum detectable level (in dB) of that sound when presented alone. This can be measured with headphones or loudspeakers. When loudspeakers are used, the measurement usually takes place in an anechoic chamber and the threshold is taken as the sound level at the theoretical center of the listener's head. A threshold determined in this way is called the Minimum Audible Field (MAF). When headphones are used, the threshold is taken as the sound level measured inside the auditory canal, very close to the eardrum (see Fig. 2.1). A threshold determined in this way is called the Minimum Audible Pressure (MAP). Ideally, the measurement of the sound pressure level (SPL) close to the eardrum is done using a small "probe" microphone. However, since this method is difficult to perform and inconvenient for the listener, the SPL at the eardrum can be estimated using calibrated headphones, or can be deduced from MAF data (see Killion, 1978). Examples of MAF (solid curve) and MAP (dashed curve) data obtained with a long-duration (> 200 ms) sinusoidal signal are shown in Figure 3.1. Both curves represent the average data for young listeners with normal hearing[1]. Note that the MAF and MAP curves have different shapes. This reflects how the torso, head and pinna influence the sound field. Also, for both the MAF and MAP, thresholds increase drastically at high (> 10 kHz) and low (< 1 kHz) frequencies. This partly reflects the transmission characteristic of the middle ear (see Sec. 2.1.1).

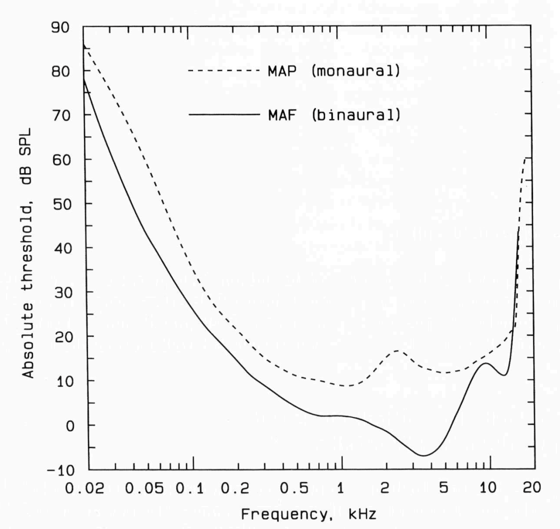


Figure 3.1: Absolute threshold (in dB SPL) as a function of signal frequency (in kHz). The solid curve shows the minimum audible field (MAF) for binaural listening. The dashed curve shows the minimum audible pressure (MAP) for monaural listening. From Moore (2003).

Measuring the minimum detectable level of a sound in the presence of one of more background sounds results in masked thresholds. Many examples of masked thresholds measurements were presented in Chapter 2 (see, *e.g.*, Fig. 2.12). Auditory masking is measured by quantifying the degree to which the detection threshold of a sound increases in the presence of a masker. Subtracting the absolute threshold of the target from the masked threshold provides the "amount of masking".

[1] The hearing capacities of a listener are determined by measuring is/her absolute thresholds in a clinical situation. In this case, the sound level is specified relative to standardized values produced by specific headphones (ANSI S3.6, 1996). Thresholds determined in this way have units dB HL ("hearing level") and help to determine eventual hearing disorders.

3.2 Frequency masking

3.2.1 Definition

To study frequency masking, masker and target are presented simultaneously and the frequency separation (ΔF) between masker and target is varied. In the common method of masking patterns, the masker has fixed frequency (F_M) and level (L_M) and the amount of masking is measured for various target frequencies (F_T).

In Section 2.3, various frequency masking experiments were presented. These experiments were specifically designed to characterize the frequency processing of the auditory system. Indeed, remind that frequency masking *per se* reflects the frequency selectivity of the ear.

3.2.2 Main psychoacoustical results

The first investigation of frequency masking was done by Wegel and Lane (1924). They measured masking patterns with pure tones as both masker and target. However, their results were complicated by the occurrence of beats for F_T very close to F_M, and combination tones (CTs) for F_T above F_M (see Sec. 2.1.4). The use of narrowband noise as either masker or target in later experiments helped to reduce these problematic masker-target interactions (*e.g.*, Egan and Hake, 1950; Greenwood, 1961a; Moore et al., 1998; Zwicker and Feldtkeller, 1999; Savel and Bacon, 2002). Typical masking patterns are represented in Figure 3.2. The elevation in threshold (amount of masking, in dB) of a pure tone is plotted as function of the tone frequency (in Hz). The masker was a continuous 90-Hz-wide band of noise centered at 410 Hz. Each curve is for an overall noise level.

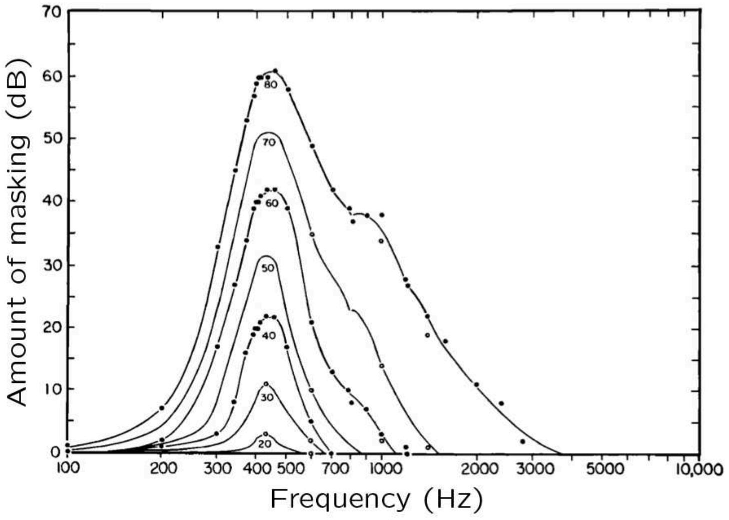


Figure 3.2: Masking patterns for a continuous narrowband noise (90-Hz wide) masker centered at 410 Hz at overall levels ranging from 20 to 80 dB SPL. The target was a continuous sinusoid whose frequency is indicated on the abscissa. Adapted from Egan and Hake (1950).

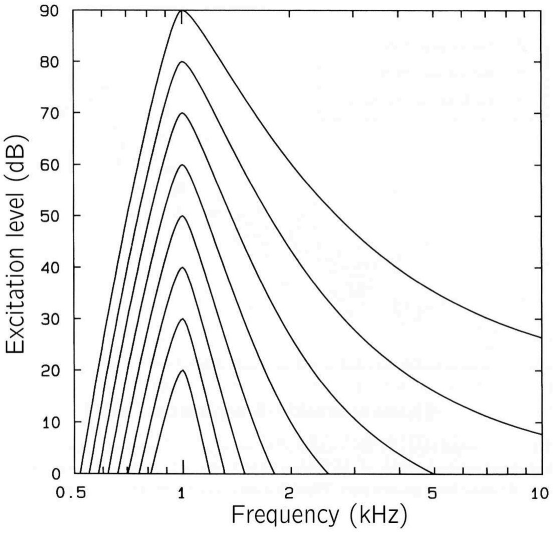


Figure 3.3: Excitation patterns calculated for a 1-kHz sinusoid at levels ranging from 20 to 90 dB SPL by 10-dB steps. From Moore (2003).

For all levels, masking is greatest when the target frequency equals the masker frequency, and decreases as ΔF increases. The slope of masking patterns is shallower on the high-frequency side ($F_T > F_M$). With increasing level ($L_M > 50$ dB SPL), this asymmetry increases in a nonlinear way. This nonlinear growth of masking on the high-frequency side has been called the “upward spread of masking”.

Masking patterns are thought to reflect the spread of masker-induced excitation on the BM (see Sec. 2.1.2). If one assumes that the target is detected when the excitation it produces on the BM is some constant proportion of the excitation produced by the masker at CFs close to F_T, then the target level at threshold is proportional to the masker excitation level (see Sec. 2.3.1 and Patterson and Moore, 1986). Hence, masking patterns and excitation patterns should have similar shapes. Moore and Glasberg (1983b) proposed a method for deriving the shape of excitation patterns from that of the auditory filters. They suggested that the excitation pattern of a given sound can be thought of as the output of the auditory filters adjacent to the signal frequency, plotted as a function of their center frequency[2]. As an example, Figure 3.3 shows excitation patterns for a 1-kHz sinusoid at levels from 20 to 90 dB SPL. The excitation patterns are asymmetric because ERB_F increases with increasing center frequency (see Eq. (2.10), Sec. 2.3.3). Also, the asymmetry of excitation patterns increases nonlinearly with increasing level. This is due to the left-hand side asymmetry that appears at high levels in the auditory filters' response (see Fig. 2.19). Overall, excitation patterns (Fig. 3.3) and masking patterns (Fig. 3.2) seem to have similar shapes. Nevertheless, this is not the case in practice since the masking patterns' shape is influenced by stimulus parameters, masker-target interactions, and off-frequency listening (Leshowitz and Wightman, 1971). The influence of masker-target interactions and off-frequency listening on masking results has been addressed in Sections 2.1.4 and 2.3.2, respectively. The main stimulus parameters affecting the masking patterns' shape are examined below.

[2] Illustration of the method and computer programs can be found at http://hearing.psychol.cam.ac.uk/Demos/demos.html (last viewed: **2011/01/06**).

In the literature, masking patterns were measured with various signal types as either masker or target: bands of noise (*e.g.*, Egan and Hake, 1950; Greenwood, 1961a), sinusoids (Wegel and Lane, 1924; Bacon and Viemeister, 1985), or complex tones (Zwicker and Feldtkeller, 1999; Oh and Lutfi, 1997). To minimize the problematic interactions as well as possible confusions between masker and target (Neff, 1985), studies generally involved masker and target signals with different spectro-temporal characteristics, namely, that differed either in their spectra (broadband masker *vs* narrowband target), in their durations (long-duration masker *vs* short-duration target), or both. Moreover, a recent study by Moore et al. (1998) in which they compared masking patterns for sinusoids and narrow[3] bands of noise as either masker or target (they tested all four possible masker-target combinations), showed that the shape of masking patterns for target frequencies below and above the masker frequency ($\Delta F \neq 0$) is determined by the characteristics of the masker rather than by the characteristics of the target. When the target frequency coincides with the masker frequency ($\Delta F = 0$), thresholds are mainly determined by the ability of the listener to detect the change in overall level caused by adding the target to the masker[4] (Moore et al., 1998). These statements remain valid as long as the target bandwidth is less than or equal to the masker bandwidth. When the target bandwidth exceeds that of the masker, some additional cues (such as off-frequency listening and the temporal structure of the stimuli) contribute to the target detection so that lower amounts of masking are obtained (Hall, 1997; Moore et al., 1998). Increasing the duration of the target may also facilitate its detection, due to temporal integration (Jeffress, 1975; Wright and Dai, 1994). However, the frequency masking studies considered below all used target signals whose bandwidths were less than or equal to the maskers' bandwidths (typically, gated or continuous sinusoids). Thus, the discussions mainly deal with the maskers' characteristics.

3.2.3 Effect of masker bandwidth

Considering that masking patterns reflect the excitation produced by the masker on the BM (Wegel and Lane, 1924; Egan and Hake, 1950; Zwicker and Feldtkeller, 1999), the spectral width of masking patterns should therefore be proportional to the masker bandwidth. In Figure 3.4, the masking patterns for maskers with comparable levels (70–80 dB SPL) and frequencies (4 kHz) but with various bandwidths are plotted together: (a) a continuous broadband noise (Bilger and Hirsh, 1956) (b) a 200-ms narrowband noise (Moore et al., 1998), (c) a gated sinusoid temporally shaped with a Hamming window that allows restricting the spectral broadening (Bacon and Viemeister, 1985), and (d) a continuous sinusoid whose spectral energy is optimally concentrated at F_M (Ehmer, 1959b).

To qualify the spectral width of the patterns in Figure 3.4, their quality factors at the 3-dB bandwidth, Q_{3dB} (see Sec. 1.3 and, *e.g.*, Tooley, 2006, pp. 77–78), were

[3] In psychoacoustics, a band of noise can be considered as "narrow" if its bandwidth is less than or equal to the ERB of the auditory filter at the noise center frequency.

[4] In the particular case when masker and target are sinusoids with same frequency, the threshold is determined by the differential threshold for intensity at that frequency. Note that the intensity increase caused by adding the target to the masker roughly depends on the relative phase between the signals (Egan and Hake, 1950; Grantham and Yost, 1982; Bacon and Viemeister, 1985; Moore et al., 1998).

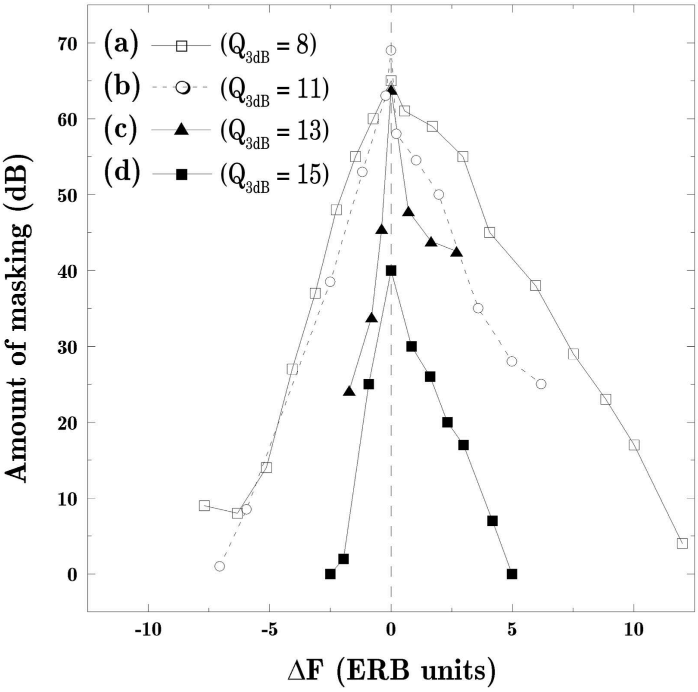


Figure 3.4: Comparison of masking patterns for maskers with equivalent frequencies and levels but with various bandwidths: (**a**) a continuous, 420-Hz-wide band of noise centered at 1210 Hz (Bilger and Hirsh, 1956), (**b**) a 200-ms (incl. 10-ms Hamming rise/fall times), 80-Hz-wide band of noise centered at 4 kHz (Moore et al., 1998), (**c**) a 50-ms (incl. 10-ms Hamming rise/fall times) sinusoid (Bacon and Viemeister, 1985), and (**d**) a continuous sinusoid (Ehmer, 1959b). The values of Q_{3dB} estimated for each masking pattern are specified in the legend.

obtained by (1) computing the linear regression lines to each side of the masking patterns[5], (2) calculating the intersection of the two regression lines, and (3) dividing the frequency (in Hz) of the intersection point by the bandwidth 3 dB below the amount of masking at the intersection point. The values of Q_{3dB} estimated for each masking pattern are specified in the legend. It can be seen that Q_{3dB} increases (and hence, patterns become narrower) as the masker bandwidth narrows. The broadband noise masker indeed produces the broadest pattern ($Q_{3dB} = 8$) while the continuous sinusoid produces the narrowest pattern ($Q_{3dB} = 15$). The narrowband noise and the gated sinusoid result in patterns with intermediate widths ($Q_{3dB} =$ 11–13). It has to be mentioned, however, that the use of gated narrowband signals (such as narrow bands of noise or sinusoids) requires specific attention regarding the choice of the temporal envelope (Bacon and Viemeister, 1985; Hartmann and Wolf, 2009). Shortening the duration of a signal with inappropriate window shapes (*e.g.*, rectangular) can result in a large spectral broadening, thereby producing broad masking patterns (see, *e.g.*, Green, 1969). Otherwise, all maskers produce the greatest masking at $\Delta F = 0$ and asymmetrical patterns, with steeper slopes for F_T below F_M than for F_T above.

[5] Because they represent a special condition of masking, the conditions $\Delta F = 0$ were excluded from the fits

3.2.4 Effect of masker center frequency

Given that the CB width increases with increasing center frequency (see Sec. 2.3), it was suggested that masking patterns measured in the lower portion of the audible spectrum should have different shapes than those measured in the higher portion. This is actually the case. Most of the studies with tone and noise maskers reported broader masking patterns at low frequencies than at high frequencies (Bilger and Hirsh, 1956; Ehmer, 1959b,a; Zwicker and Jaroszewski, 1982; Moore et al., 1998; Zwicker and Feldtkeller, 1999). In Figure 3.5, the results from Ehmer (1959a) obtained with sinusoidal (straight lines) and narrowband noise (dashed lines) maskers at 0.5, 1.0, 2.0 and 4.0 kHz are reported. At each center frequency, the noise bandwidth was fixed to $\frac{1}{3}$ octave (*i.e.*, approximately one CB). The level of each masker was adjusted so that both masker types were presented at 60 dB SL. For both tone and noise maskers, the masking patterns seem to broaden as F_M decreases from 4.0 to 0.5 kHz. Computing the quality factors (Q_{3dB}) for each of the patterns in Figure 3.5 indeed showed that Q_{3dB} decreases from about 14 at $F_M = 4$ kHz to 6 at $F_M = 0.5$ kHz for the tone masker. Similarly, Q_{3dB} decreases from about 21 at $F_M = 4$ kHz to 8 at $F_M = 0.5$ kHz for the noise masker. This clearly indicates that low-frequency patterns are broader than high-frequency ones. In other terms, high-frequency maskers are effective over a relatively narrow frequency region around F_M, while low-frequency maskers tend to be effective over a wider range of frequencies.

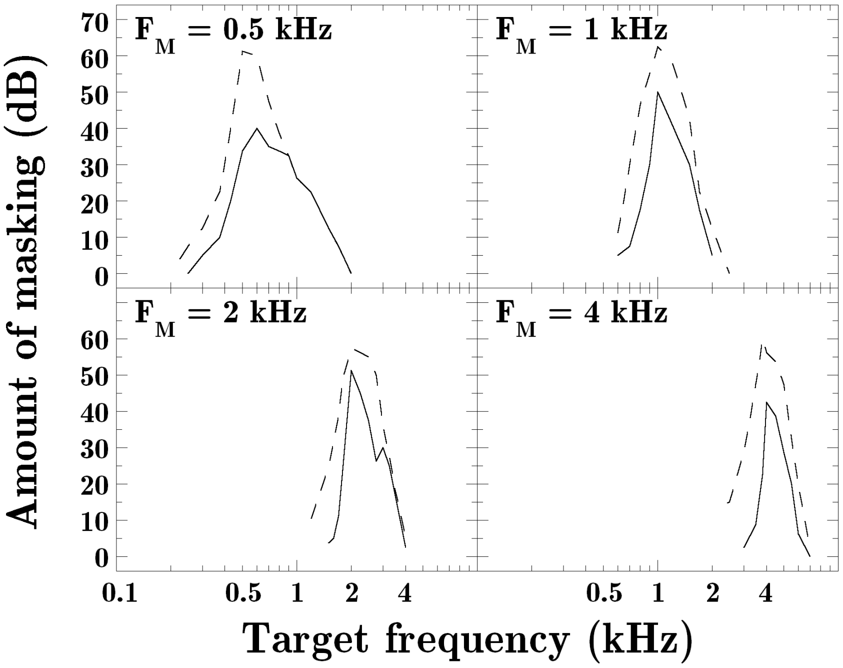


Figure 3.5: Masking patterns for sinusoidal (straight lines) and narrowband noise (dashed lines) maskers at $F_M = 0.5$, 1.0, 2.0 and 4.0 kHz. At each center frequency, the noise bandwidth was fixed to $\frac{1}{3}$ octave. Both masker types were presented at 60 dB SL. The results for one typical listener are reported. Adapted from Ehmer (1959a).

3.2.5 Effect of masker level

The masking patterns in Figure 3.2 are asymmetric at high masker levels ($L_M > 50$ dB SPL): the slopes for F_T above F_M become shallower with increasing masker level. Besides the investigation by Egan and Hake, several studies with tone and noise maskers examined the dependence of the masking patterns asymmetry on masker level (Bilger and Hirsh, 1956; Ehmer, 1959b; Vogten, 1978b,a; Zwicker and Jaroszewski, 1982; Lutfi and Patterson, 1984; Moore et al., 1998). All the cited studies observed the right-hand asymmetry at high levels. Moreover, most of these studies reported a reversal of the masking patterns asymmetry at low levels ($L_M < 40$ dB SPL), *i.e.*, shallower slopes for F_T below F_M than for F_T above (Ehmer, 1959b; Vogten, 1978a; Zwicker and Jaroszewski, 1982; Lutfi and Patterson, 1984). Although it is not well apparent in Figure 3.2, the masking patterns of Egan and Hake actually represent this low-level asymmetry (Zwicker and Jaroszewski, 1982).

Figure 3.6 presents results obtained by Zwicker and Jaroszewski (1982) with a continuous sinusoidal masker at 4 kHz at levels ranging from 20 to 60 dB SPL. The target was a 250-ms sinusoid with variable frequency. The amount of masking (in dB) is plotted as a function of ΔF (in ERB units). Each masking pattern is for a masker level. These data clearly show the reversal of the patterns asymmetry that appears at low masker levels ($L_M < 40$ dB SPL). At these levels, masking is greater for F_T below F_M than for F_T above. At intermediate levels ($L_M = 40$–50 dB SPL), the patterns are approximately symmetrical. Then, as L_M increases to 60 dB SPL, the expected right-hand asymmetry is observed with greater masking for F_T above F_M than for F_T below. The interpretation of the level-dependency of these asymmetries is given below.

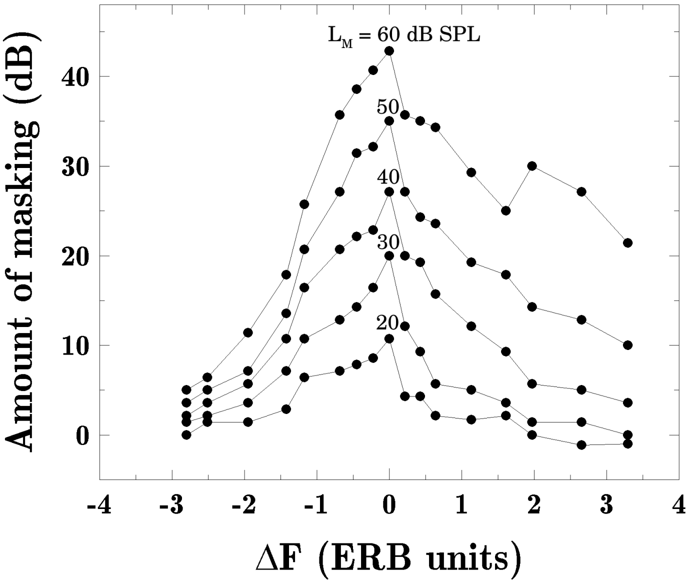


Figure 3.6: Masking patterns for a continuous sinusoidal masker at 4 kHz at levels ranging from 20 to 60 dB SPL. The abscissa is scaled in ERB units relative to the ERB centered at 4 kHz. The median data from four listeners are reported. Adapted from Zwicker and Jaroszewski (1982).

Another method for examining the effect of masker level on masking is to plot "Growth-Of-Masking" (GOM) functions, otherwise called "masking functions". These are plots of the target level at threshold (or the amount of masking) as a function of masker level. Typically, such functions are monotonic and can be described by one or more straight-line segments depending on the range of tested levels. In simultaneous masking and for masker levels in the range 40–80 dB SPL, the slopes of the GOM functions roughly depend on ΔF. When F_T is well above F_M, the function is very steep, with a slope usually greater than unity. This non linear growth of masking for $F_T > F_M$ reflects the upward spread of masking observed in the masking patterns (Figs. 3.2, 3.4, 3.6). When F_T is close to or below F_M, the function is shallower, with a slope less than 1.0 (Wegel and Lane, 1924; Egan and Hake, 1950; Bacon and Viemeister, 1985; Bacon et al., 1999). An example of GOM function is presented in Figure 3.7 (adapted from Bacon et al., 1999). These functions were obtained with a 20-ms sinusoidal target at 1944 Hz presented in the temporal center of a 100-ms sinusoidal masker at 1350 Hz (*i.e.*, $F_T \gg F_M$). L_M ranged from 40 to 95 dB SPL. The function was fitted with two straight-line segments whose slopes are indicated in the figure. As expected, the slope of the first segment (60–85 dB) is about 2.0, indicating a non linear GOM in this region. At higher levels, the slope becomes closer to the unity.

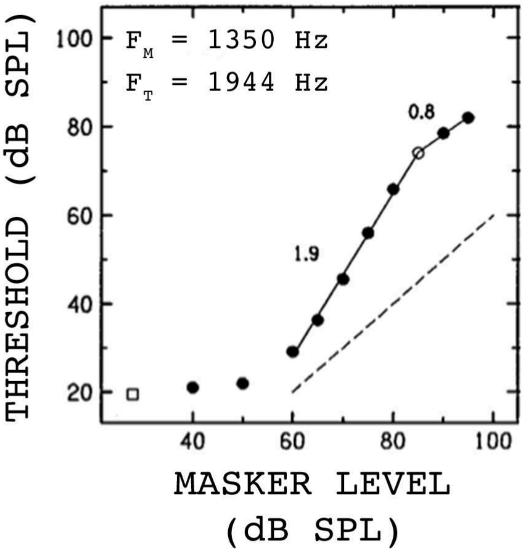


Figure 3.7: Growth-of-masking (GOM) function for a 20-ms sinusoidal target at 1944 Hz presented in the temporal center of a 100-ms sinusoidal masker at 1350 Hz ($F_T = 1.44 \times F_M$). The target level at threshold (in dB SPL) is plotted as a function of masker level (in dB SPL). The empty square in the bottom left corner indicates the absolute threshold of the target. The function was fitted with two straight-line segments whose slopes are indicated close to each segment (points with less than 10 dB of masking were excluded from the fits). The dashed line represents a linear reference (slope = 1.0). The results averaged across four listeners are reported. Adapted from Bacon et al. (1999).

Note that an analogy can be made between GOM functions and the input-output function of the BM (see Fig. 2.9, Sec. 2.1). In fact, GOM functions are used as a tool for estimating BM nonlinearities, namely BM compression and suppression (see Sec. 2.1.4 and Ruggero et al., 1997; Bacon et al., 1999; Savel and Bacon, 2002; Oxenham and Bacon, 2003; Yasin and Plack, 2005; Rodriguez et al., 2010).

3.2.6 Physiological mechanisms underlying frequency masking

Frequency masking is commonly attributed to two different physiological mechanisms (Delgutte, 1990; Moore, 2003, Sec. 3.6). The first is excitatory masking. It is caused by the swamping of target-induced excitation from masker-induced excitation on the BM. More precisely, reminding that the stimulation of an auditory nerve fiber results in the increase of its discharge rate (see Sec. 2.1.3), excitatory masking occurs when the masker produces an increase in the discharge rates of the fibers that would normally respond to the target, *i.e.*, at CFs close to F_T. Stimulation by the target will then produce a further increase in the discharge rates. If this further increase is too small, *i.e.*, *not detected*, then the target is masked. The second mechanism is suppressive masking. It refers to the reduction of target-induced excitation on the BM resulting from the addition of another signal which itself does not produce excitation at the CF of the target. The added signal thus has the effect of masking the target. In that case, the masker does not increase the firing rates of the fibers around F_T, but shifts their thresholds towards higher intensities. Note that an analogy can be made between suppressive masking and "two-tone suppression" described in Section 2.1.4.

These two mechanisms of frequency masking are not mutually exclusive. Masking sounds might produce both excitation and suppression in the channels (*i.e.*, auditory filters or CBs) responding to the target. In fact, the relative contributions of excitatory and suppressive masking depend on the frequency and level relationships between masker and target (Delgutte, 1990; Yasin and Plack, 2005; Rodriguez et al., 2010). Most of the studies on suppression found that this effect is dominant for $F_T > F_M$ and for levels greater then 50 dB SPL (Vogten, 1978b,a; Delgutte, 1990; Yasin and Plack, 2005; Rodriguez et al., 2010). At lower levels and for $F_T < F_M$, suppression still occurs but masking is mainly excitatory. Overall, the consideration that the masking patterns' shape reflects the masker-induced excitation on the BM is challenged by the fact that suppression increases the masking effect in some cases. Because the target cannot be suppressed by the masker when both signals do not overlap in time, it was proposed that forward masking patterns (*i.e.*, masking patterns measured with a brief temporal gap, of the order of 1–2 ms, between masker and target) might better reflect the actual pattern of excitation produced by the masker than simultaneous masking patterns (Vogten, 1978b,a; Delgutte, 1990; Yasin and Plack, 2005; Rodriguez et al., 2010). Accordingly, psychophysical tuning curves (PTCs, se Sec. 2.3.2) measured in forward masking are sharper than those measured in simultaneous masking (Vogten, 1978a; Moore and Glasberg, 1982; Yasin and Plack, 2005; Rodriguez et al., 2010).

The masking patterns asymmetry is mainly attributed to the increase in the auditory filters bandwidth with increasing center frequency (Moore, 2003). The level-dependency of this asymmetry is commonly attributed to the contribution of suppression effects (Vogten, 1978b,a; Delgutte, 1990; Yasin and Plack, 2005), and to the level-dependent changes in the shape of auditory filters (see Lutfi and Patterson, 1984 and Sec. 2.3.3). Note that at very high masker levels ($>$ 90 dB SPL), the activation of the middle-ear reflex can result in a reduction of the upward spread of masking (Bilger and Hirsh, 1956; Gelfand, 1998). As described in Section 2.1.1, the middle-ear reflex reduces the transmission of sounds through the middle ear, and the degree of attenuation roughly depends on the sound frequency, low-frequency sounds ($<$ 1 kHz) being more attenuated than higher ones. This frequency-selective attenuation has the effect of reducing the suppressive masking of high-CF fibers in the presence of high-level, low-frequency masking sounds (for a more detailed description of these effects, see Pang and Guinan, 1997; Liberman and Guinan, 1998), which overall results in a reduction of the upward spread of masking.

To account for the broad masking patterns obtained at low frequencies ($F_M < 1$ kHz), it was suggested that low-frequency maskers excite a wider portion of the BM than high-frequency ones (Ehmer, 1959b; Gelfand, 1998; Moore et al., 1998). Remind from Chapter 2 that high-frequency sounds produce maximum displacements of the BM near the base with little movement on the rest of the membrane. Low-frequency sounds conversely produce displacements along most of the BM but with a maximum near the apex. High-frequency sounds are thus most likely affected by the displacement patterns caused by low-frequency sounds. Moreover, the auditory filters at low frequencies are very narrow (*e.g.*, $ERB_F < 50$ Hz for $F < 0.2$ kHz, see Fig. 2.18). It is therefore expected that even narrowband maskers excite several CBs around F_M, thus producing broad masking patterns (Moore et al., 1998). Finally, the elevated threshold in quiet in the low-frequency portion of the audible spectrum (see Fig. 3.1) can affect the shape of masking patterns. Zwicker and Jaroszewski (1982) indeed showed that masking patterns measured with a 250-Hz sinusoidal masker are almost symmetrical at low masker levels, *i.e.*, the reversal of the asymmetry is not observed at low frequencies.

Summary

To study frequency (simultaneous) masking, the frequency separation between target and masker (ΔF) is varied. In the common method of masking patterns, the masker frequency (F_M) is fixed and the amount of masking is measured for various target frequencies (F_T). Simultaneous masking can be attributed to two different physiological phenomena. The first is excitatory masking. It is caused by the swamping of target-induced excitation from masker-induced excitation on the basilar membrane (BM). The second is suppressive masking. It refers to the reduction of target-induced excitation on the BM resulting from the addition of another signal which itself does not produce excitation at the tonotopic place associated with the target. The relative contributions of excitatory and suppressive masking depend on the frequency and level relationships between masker and target.

Masking is greatest when the target frequency equals the masker frequency, and decreases as ΔF increases. The slope of masking patterns is shallower on the high-frequency side ($F_T > F_M$), which is attributed to the increase of the auditory filters bandwidth with increasing center frequency. With increasing signal level (> 50 dB SPL), this asymmetry of masking patterns increases in a non linear way. This non linear growth of masking on the high-frequency side has been called the “upward spread of masking”, and is commonly attributed to the contribution of suppression effects. Irregularities (*i.e.*, local dips) were observed in the masking patterns obtained with sinusoidal maskers for F_T above F_M. They were attributed to other interactions between masker and target such as cochlear distortion products and beats.

3.3 Temporal masking

3.3.1 Definition

The previous section presented *simultaneous* masking, *i.e.*, situations where target and masker are presented simultaneously. In such a paradigm, varying the frequency separation (ΔF) between both signals allows to study frequency masking. Masking can also occur when the target is presented before or after the masker. This is referred to as *non-simultaneous*, or temporal masking. Depending on the temporal sequence of the stimuli, two kinds of non-simultaneous masking can be distinguished. The first is *backward* masking (also referred to as pre-masking), in which the target temporally precedes the masker. The second is *forward* masking (post-masking), in which the masker temporally precedes the target. In Figure 3.8, the two types of non-simultaneous masking and the simultaneous masking paradigms are illustrated.

To study temporal masking, F_M and F_T are identical ($\Delta F = 0$) and the temporal separation (ΔT) between masker and target is varied. Depending on the stimulus durations and whether backward or forward masking is investigated, the definition of the ΔT parameter can greatly vary across studies. In the literature, ΔT is commonly defined as (1) masker onset-to-target onset (*e.g.*, Duifhuis, 1973), (2) masker offset-to-target onset (Elliott, 1962), (3) masker offset-to-target offset

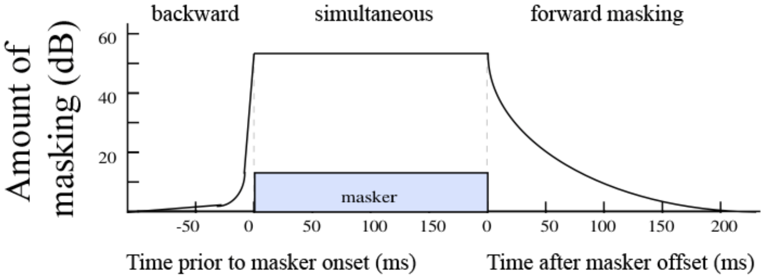


Figure 3.8: Illustration of the backward, simultaneous, and forward masking paradigms.

(Zwislocki et al., 1959), or (4) masker peak-to-target peak (Nizami and Schneider, 1999). Usually, ΔT is positive for forward and negative for backward masking. Because of these various definitions, comparing temporal masking data from diverse studies should be done cautiously. The ΔT definition hardly affects the temporal masking results for large values of ΔT. However, specific attention is required with small temporal separations because definitions (1) and (4) of ΔT might result in a physical overlap of masker and target (see Nizami and Schneider, 2000).

In this section, typical forward and backward masking results are reported, and their underlying physiological mechanisms are evoked.

3.3.2 Forward masking results

Figure 3.9 shows a typical forward masking function obtained by Fastl (1976) with a 500-ms impulse of uniformly masking noise (bandwidth = 0–16 kHz) at 60 dB SPL overall. The target was a 1-ms sinusoid (including 0.5-ms Gaussian rise/fall times) at 8 kHz. The target level at threshold (in dB) is plotted against ΔT (defined here as maker offset-to-target offset, in ms) on a logarithmic scale. Masking is greatest when the target is presented shortly after the masker offset ($\Delta T < 5$ ms). Then, masking roughly decreases as ΔT increases up to about 150 ms. This decay of forward masking is a linear function of $\log(\Delta T)$.

Forward masking is largely influenced by the masker duration. Precisely, the amount of masking at a given ΔT increases with increasing masker duration for durations up to 200 ms (Penner, 1974; Kidd Jr. and Feth, 1982; Zwicker, 1984). The studies by Fastl (1976, 1977, 1979) revealed an effect of masker duration for durations up to 50 ms only. The effect of target duration is limited to small ΔT values (typically, masker offset-to-target onset intervals < 20 ms). In these conditions, increasing the target duration results in a decrease of the amount of masking (Thornton, 1972; Fastl, 1976, 1977; Oxenham, 2001). For larger ΔTs, increasing the target duration has no effect on thresholds.

Figure 3.10 illustrates the effect of masker level (L_M) on forward masking (data from Jesteadt et al., 1982). The amounts of masking (in dB) for various levels of a 300-ms sinusoidal masker (L_M = 20–80 dB SPL) are plotted (a) as a function of ΔT (defined as masker offset-to-target onset, in ms) on a logarithmic scale, with L_M as the parameter, and (b) as a function of L_M (in dB SPL), with ΔT as the parameter. The target was a 20-ms sinusoid. Both masker and target had a carrier frequency

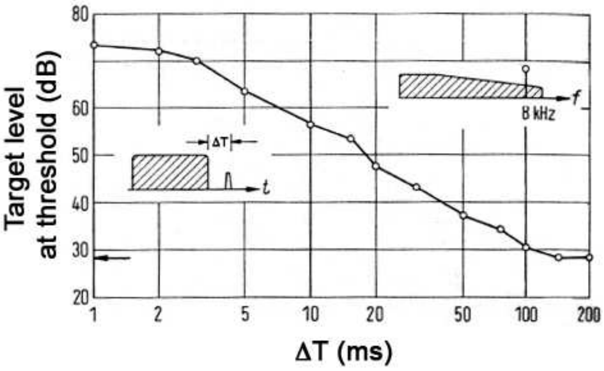


Figure 3.9: Forward masking of a 1-ms sinusoid by a 500-ms impulse of uniformly masking noise. The target level at threshold (in dB) is plotted as a function of ΔT (defined as masker offset-to-target offset, in ms) on a logarithmic scale. The temporal and spectral relations between stimuli are indicated in inserts. The arrow in the bottom left corner indicates the absolute threshold of the target. Adapted from Fastl (1976).

of 4 kHz, and were gated on and off with 10-ms raised-cosine ramps.

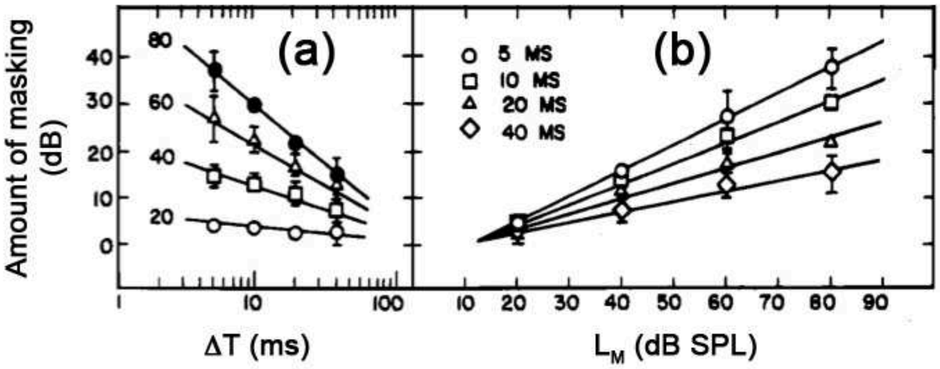


Figure 3.10: Forward masking of a 20-ms sinusoidal target by a 300-ms sinusoidal masker whose level varied from 20 to 80 dB SPL in 20-dB steps. Masker and target had a carrier frequency of 4 kHz. (a) amount of masking (in dB) as a function of ΔT (defined as masker offset-to-target onset, in ms) on a logarithmic scale, with L_M as the parameter. (b) same data plotted as a function of L_M (in dB SPL), with ΔT as the parameter. The data averaged across 4 listeners are shown. From Jesteadt et al. (1982).

First, Figure 3.10(a) indicates that for each masker level the decay of forward masking is a linear function of $\log(\Delta T)$. All functions intersect at a common point on the abscissa. In other words, the rate of decay of masking is more rapid for the highest masker levels (Jesteadt et al., 1982; Moore and Glasberg, 1983a). Second, Figure 3.10(a) indicates that increasing L_M results in an increase of the amount of forward masking. Note, however, that a given increment in masker intensity does not produce an equal increment in amount of masking. This is best seen in Figure 3.10(b) showing GOM functions. In simultaneous masking and when F_T is close to or below F_M, such functions have slopes close to the unity (see Sec. 3.2.5). In forward masking, the functions have slopes less than one, and the slopes decrease with increasing ΔT (Widin and Viemeister, 1979a; Moore and Glasberg, 1983a). The slopes of the GOM functions in Figure 3.10(b) range from 0.55 for $\Delta T = 5$ ms to 0.20 for $\Delta T = 40$ ms.

Finally, the frequency content of the masker can also affect forward masking results at small ΔT values (< 20 ms). Broadband maskers (such as broadband noises or short sinusoids temporally controlled with windows producing a large amount of spectral spread) produce more masking than narrowband maskers (narrowband noises or long-lasting sinusoids) (Duifhuis, 1973; Widin and Viemeister, 1979b; Weber and Moore, 1981; Moore, 1981). Moreover, Weber and Moore (1981) showed than for short targets (duration $\leq$ 35 ms), noise maskers are more effective than sinusoidal maskers for offset-onset intervals of 0–50 ms.

A few studies measured forward masking functions for various frequency separations between masker and target (*i.e.*, $\Delta F \neq 0$, see Fastl, 1979; Kidd Jr. and Feth, 1981; Soderquist et al., 1981). Those studies involved long-lasting (duration $\geq$ 250 ms) sinusoidal maskers and short sinusoidal targets. Their results indicated that (1) varying ΔF does not affect the linear decay of forward masking as a function of $\log(\Delta T)$, (2) the slope of this decay decreases as ΔF increases (for both positive and negative ΔF values), and (3) forward masking decays to 0 dB at approximately the same ΔT value (which is listener-dependent) for all ΔFs. Furthermore, Munson and Gardner (1950) and Widin and Viemeister (1979a) measured GOM functions in forward masking with $\Delta F \neq 0$. They reported a non linear growth of masking for $\Delta F \neq 0$.

3.3.3 Backward masking results

Figure 3.11 shows backward masking results for 2-ms sinusoidal targets (including 1-ms Gaussian rise/fall times) with $F_T = 3.5$, 4.0, and 5.5 kHz masked by a 300-ms, 4-kHz sinusoidal masker at 70 dB SPL (Fastl, 1979). The target level at threshold (in dB) is plotted as a function of ΔT (specified as the maker onset-to-target onset distance, in ms). Masking is greatest when the target is presented within 2–4 ms prior to the masker onset (note that thresholds obtained at $\Delta T = 0$ correspond to simultaneous masking, see Fastl 1979, Fig. 2). Masking drastically decreases as ΔT increases. For all target frequencies, the amount of backward masking for $\Delta T \leq$ -10 ms is less than 10 dB. This observation is in line with most of the backward masking studies using broadband or narrowband maskers with durations $\geq$ 50 ms (Elliott, 1962; Penner, 1974; Fastl, 1976, 1977, 1979; Soderquist et al., 1981; Dolan and Small, 1984). Backward masking is indeed influenced by the masker duration, such that short (duration $\leq$ 10–25 ms, depending on the study) masker impulses elicit less backward masking than longer ones (Duifhuis, 1973; Penner, 1974; Fastl, 1976, 1977, 1979). Extending or shortening the target duration hardly affects backward masking results (Duifhuis, 1973; Fastl, 1976, 1977, 1979).

As for forward masking, backward masking is largely influenced by the masker level (L_M). Increasing L_M results in an increase of the amount of backward masking (Penner, 1974; Fastl, 1976, 1977, 1979). The GOM functions deduced from the

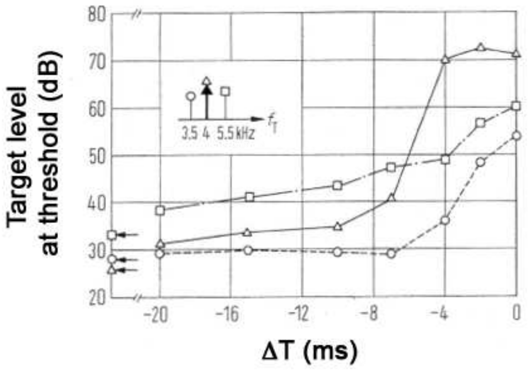


Figure 3.11: Backward masking of 2-ms sinusoids with frequencies F_T (indicated in the insert) by a 300-ms, 4-kHz sinusoidal masker at 70 dB SPL. The target level at threshold (in dB) is plotted as a function of ΔT (defined as masker onset-to-target onset, in ms). The arrows in the bottom left corner indicate absolute thresholds of the targets. Adapted from Fastl (1979).

cited studies suggest a non linear relationship between the increase of L_M and the resulting increase in the amount of masking. However, it has to be noted that the data reported were collected on a few listeners but that inter-listener differences were considerably large.

Finally, the amount of backward masking depends on the spectral relationships between masker and target. For broadband maskers *versus* narrowband targets, smaller amounts of backward masking are observed for F_Ts in the high-frequency portion of the audible spectrum (*i.e.*, $>$ 4 kHz) than for F_Ts in the mid- or low-frequency portions (*i.e.* $<$ 2 kHz, see Duifhuis, 1973; Dolan and Small, 1984). For others combinations of masker-target types (*e.g.*, broadband masker *vs.* broadband target, or narrowband masker *vs.* narrowband target), the amount of backward masking decreases when a frequency separation between masker and target is introduced (*i.e.*, $\Delta F \neq 0$, see Duifhuis, 1973; Fastl, 1976, 1977, 1979; Dolan and Small, 1984).

3.3.4 Physiological mechanisms underlying temporal masking

Although the mechanisms underlying backward masking remain unclear, the most advanced explanation is of peripheral origin. Backward masking would be caused by the temporal overlap of the BM responses to masker and target at the outputs of the auditory filters (Duifhuis, 1973; Dolan and Small, 1984). The amount of overlap depends on the “ringing" time of the BM, *i.e.*, the length of the impulse response of the BM, which itself depends on signal frequency. Given that the auditory filters' bandwidth increases with increasing center frequency (see Sec. 2.3), low-frequency filters have longer impulse responses than high-frequency filters[6]. Consequently, low-frequency components are more likely to produce overlapping

[6] An estimate of the ringing time at a given frequency can be obtained by taking the inverse of the CB at that frequency. For example, the ERB of the auditory filter centered at 4 kHz is 456 Hz (see Eq. (2.10)). Thus, the ringing time at 4 kHz is estimated to $456^{-1} \approx 2.2$ ms (Carlyon, 1988).

responses than high-frequency components, and hence more backward masking. This explanation is in agreement with the observation above that broadband maskers produce less backward masking at high frequencies. This explanation further implies that backward masking strongly depends on the spectro-temporal properties of the stimuli, namely the amount of interaction the signals are likely to produce at the filters outputs (see Duifhuis, 1973, Sec. I and Fig. 3). In other words, if short *and* narrowband masker and target are used, backward masking is expected to occur only if there is a physical overlap of the two signals. This was actually verified by Duifhuis. Note, however, that backward masking studies often reported large inter-listener differences (Penner, 1974; Fastl, 1976, 1977, 1979; Dolan and Small, 1984), and that trained listeners often show little or no backward masking. Thus, backward masking effects may also reflect some confusion effects between masker and target (Moore, 2003, Sec. 3.9).

Forward masking can be attributed to four mechanisms. The first is the temporal overlap of the BM responses to masker and target at the outputs of the auditory filters as a consequence of the filters ringing (Duifhuis, 1973; Carlyon, 1988; Nizami and Schneider, 2000). This phenomenon is more likely to be involved with small values of ΔT and at low frequencies where ringing times are the longest (see footnote 6). Second, the exponential decay of masker-induced excitation over time in the cochlea and in the auditory nerve can reduce the response to a target signal presented shortly after the extinction of the masker (Zwislocki et al., 1959; Elliott, 1962; Plomp, 1964; Thornton, 1972; Soderquist et al., 1981; Jesteadt et al., 1982; Kidd Jr. and Feth, 1982). This effect is commonly referred to as short-term adaptation or “fatigue”, and is largely dependent on the intensity and duration of the stimulation evoked by the masker (Lüscher and Zwislocki, 1949; Zwislocki et al., 1959; Plomp, 1964; Smith and Zwislocki, 1975; Palmer, 1995; Sumner et al., 2003; Meddis and O’Mard, 2005). However, because physiological measures of neural adaptation made in some auditory nerve fibers could not account for all the forward masking effects observed psychophysically, a third higher-level mechanism was suggested: more central effects of persistence of the neural activity induced by the masker (Oxenham and Plack, 2000; Oxenham, 2001). Based on this idea, the cited authors proposed a model of temporal integration, the fourth suggested mechanism, that could fairly well predict forward and backward masking data for various stimulus configurations (see the system analysis approach described in Fig. 2.11). This model was an attempt to distinguish between neural adaptation and temporal integration as the most probable explanation to forward masking. To date, however, this problem is not clearly solved (Moore, 2003).

3.3.5 A particular case: Temporal course of simultaneous masking (overshoot)

The threshold of a brief target (duration $\leq$ 20 ms) masked by a long-duration ($>$ 200 ms) masker can be 10–20 dB higher when that target is presented near the onset of the masker than when it is presented at the temporal center or offset of the masker (Elliott, 1965; Zwicker, 1965; Fastl, 1976, 1977, 1979; Soderquist et al., 1981; Wright, 1997; Strickland, 2001; Savel and Bacon, 2004). This effect has been called “overshoot”. Greatest overshoot is observed at high frequencies (little or no overshoot effect was measured for $F_T < 1$ kHz), when the masker bandwidth covers a large frequency range, and when the masker is presented at moderate levels (*i.e.*, 40–80 dB SPL). Although the overshoot effect is still not well understood, several possible explanations were proposed. One is that overshoot represents a sharpening of the CB with time (Elliott, 1965, 1967; Green, 1969; Bacon et al., 2002). Another explanation is that overshoot represents short-term adaptation (Lüscher and Zwislocki, 1949; Smith and Zwislocki, 1975). More recently, it has been suggested that overshoot is related to the locations of masker and target signals on different portions of the compressive input-output function of the BM. Because the BM mechanics *per se* do not change over time, it is suggested that the role of BM compression in overshoot is mediated by active processing in the auditory system. This active processing is thought to be due to the action of the OHCs (see Sec. 2.1.2) on the amount of BM compression via an efferent feedback loop (Strickland, 2001; Savel and Bacon, 2004). Accordingly, McFadden and Champlin (1990) showed that ingestion of aspirin, an agent that selectively affects the activity of the OHCs (Ruggero and Rich, 1991), reduces or cancels overshoot.

3.4 Additivity of masking

All the experiments described above measured the amount of masking produced by a single masker on a target separated either in frequency (simultaneous masking) or in time (non-simultaneous masking) from that masker. Others studies, however, examined the effects of combining several maskers distributed either along the frequency (*e.g.*, Moore, 1985; Humes et al., 1992) or time (*e.g.*, Penner, 1980; Cokely and Humes, 1993) axis. Their results allowed deriving laws on the additivity of masking. Although the present dissertation is mainly concerned with the spread of masking evoked by a single masker, this section briefly relates the main results on the additivity of masking for the sake of the modeling approach treated in Chapter 13.

If the additivity of masking were considered as linear, then the masked threshold of a target in the presence of two equally effective maskers (*i.e.*, each masker alone causes the same masked threshold) would be 3 dB higher than the masked threshold of the target in the presence of each masker alone. In other words, a linear summation would result in a doubling of each masker energy. In fact, it has been found that combining two maskers often results in higher masked thresholds than those predicted by linear additivity. The difference between the linear prediction and the actually measured threshold is referred to as "excess masking". The amount of excess masking strongly depends on the stimulus configurations. Large amounts of excess masking (10–17 dB) were obtained for *temporally non-overlapping* stimuli in case of non-simultaneous masking and *spectrally non-overlapping* stimuli in case of simultaneous masking. However, no excess masking occurs (*i.e.*, linear additivity of masking) in conditions involving a temporal *and* spectral overlap of maskers and target.

The commonly accepted origin of excess masking is that the individual maskers are subjected to a peripheral compressive non linearity before their effects are combined linearly at higher stage (Penner, 1980; Humes and Jesteadt, 1989). Based on that explanation, Humes and Jesteadt (1989) proposed a model of the additivity of masking that could accurately predict most of the data for combined simultaneous maskers and combined non-simultaneous maskers. The model, called "modified power-law model", has the form

$$i_x = (I_{MT_x})^p - (I_{QT})^p \tag{3.1}$$

where i_x represents the internal effect produced by masker 'X' at the target frequency, I_{MT_x} is the intensity of the target at masked threshold for masker X, and I_{QT} is the target intensity at threshold in quiet (constant term). The exponent p defines the amount of compression, its value ranges from 0 to 1. The model predicts linear additivity of masking for $p = 1$ and excess masking for lower values of p.

When maskers and target fall into the same peripheral frequency channel and temporally overlap, no excess masking occurs because the stimuli are added *before* they are compressed.

Summary

To investigate temporal (non-simultaneous) masking, F_T and F_M are identical and the temporal separation (ΔT) between masker and target is varied. Backward masking (the target temporally precedes the masker) is weaker than forward masking (the masker precedes the target), independently of masker duration. The amounts of backward and forward masking depend on masker duration.

Backward masking is thought to be caused by the temporal overlap of the BM responses to masker and target at the outputs of the auditory filters. The amount of overlap depends on the ringing time of the BM. Forward masking is thought to be caused by the decay of masker-induced excitation in the auditory periphery, by neural adaptation, and by a more central persistence of masker-induced activity. With small values of ΔT inducing a temporal overlap of the BM responses to masker and target, forward masking involves the same mechanisms as those underlying backward (or simultaneous) masking.

Chapter 4

Why studying auditory time-frequency masking with Gaussian signals?

Contents

Chapter 3 presented psychoacoustical masking data obtained with simultaneous (frequency masking) and non-simultaneous (temporal masking) presentation of masker and target. A well-known application of psychoacoustical masking data in the field of sound signal processing is audio coding. To reduce the digital size of audio files, audio codecs (such as MPEG-1 Layer III or MPEG AAC) use time-frequency (TF) analysis schemes (such as those described in Chap. 1) to decompose arbitrary complex sounds into TF segments, and exploit the properties of auditory masking to reduce the bit rates in segments which are subject to masking. To determine the perceptual relevance of TF components, audio codecs – and, to a larger extent, a great variety of audio applications that are concerned with TF representations (like sound analysis-synthesis tools or sound characterization techniques) – require a model of TF masking for "elementary" sounds (*i.e.*, with maximal concentration in the TF plane).

To date, however, most of the perceptual audio coding algorithms mainly use a simple model of *frequency* masking, based on psychoacoustical results with sinusoids or noise bands as masker and/or target (see Sec. 3.2). Masking models were developed to exploit both frequency and temporal masking effects in audio codecs, but most of them simply assume a linear combination of frequency and temporal masking effects. Given the highly nonlinear behavior of the TF processing by the human auditory system (see Chap. 2), such a simple combination of temporal and frequency masking data is unlikely to predict accurately the masking effects between TF components.

Overall, psychoacoustical data on TF masking for elementary sounds are currently missing. This leads to the main purpose of the present research, namely to collect data on the TF spread of masking produced by a single TF component. Such measurements may allow new TF masking model to be implemented in sound analysis-synthesis tools such as perceptual audio codecs.

In this chapter, the most frequently implemented masking models in perceptual audio codecs are presented, and the limits of these models are evoked. An improvement of these models is then proposed based on the measurements of TF masking data for stimuli with maximal concentration in the TF plane.

4.1 Modeling of masking data for implementation in audio codecs

In this section, the current models of frequency and temporal masking are presented, and their implementations in audio codecs (specifically, the MPEG-1 Layer I) are shortly described. Note that other audio applications like “irrelevance filters” (Balazs et al., 2010) or speech processing algorithms (*e.g.*, Skoglund and Kleijn, 2000) use those models. Depending on the applications, the parameters and the implementations of these models in signal processing tools can greatly vary.

4.1.1 Presentation of current masking models

To reduce the digital size of audio files, audio codecs allocate low bit rates in TF segments which are subject to masking. Reducing the bit rates results in introducing quantization noise in these segments. Therefore, the role of the masking models is to deliver a “global” masking threshold indicating how much quantization noise can be introduced in each TF segment without the noise becoming audible. In psychoacoustics, two types of models can be used to predict a masking threshold. The first type are excitation pattern-based models. These models transform the short-term spectrum of the input signal into an excitation pattern reflecting the spread of excitation induced by the signal on the BM (see Fig. 3.3). This approach is based on the power-spectrum model of masking in which the auditory periphery is conceived as a bank of bandpass filters (see Sec. 2.3.1). Masking is then determined by the target-to-masker ratio at the output of each filter, or “frequency channel”. This is the most frequently employed technique in audio codecs, and the one considered in this section.

The second type of models attempts to simulate the effective signal processing in the auditory system (Patterson et al., 1995; O’Donovan and Furlong, 2005; Jepsen et al., 2008). In particular, the model presented in Jepsen et al. (2008) is based on the modeling of both spectral and temporal masking effects. An attempt was made by van de Par et al. (2008) to incorporate this model in an audio coding algorithm. The resulting coder was shown to outperform (in terms of both coding efficiency and audio quality) the more conventional coders based on frequency masking models only. However, the computational complexity required by this type of models renders their implementations problematic, at least for audio coding techniques. However, this is beyond the scope of the present study. The remaining of this section focuses on the first type of models, the excitation pattern-based ones.

4.1.1.1 Frequency masking model

Frequency masking is modeled by the "spreading function" (SF) of masking. This function is assumed to describe the spread of excitation induced by a single sinusoid on the BM. SF was designed based on an analogy with the masking patterns' shape (see Sec. 3.2). An analytic expression of SF in the Bark scale is

$$SF(z) = 15.81 + 7.5\,(z + 0.474) - 17.5\,\sqrt{1 + (z + 0.474)^2} \tag{4.1}$$

where z is the CB number in Bark units (see Sec. 2.3.3 and Zwicker, 1961) and $SF(z)$ is expressed in dB (Painter and Spanias, 2000). $SF(z)$ is a triangle-like function characterized by lower and upper slopes of about +25 and -10 dB/Bark, respectively. To take into account the decreasing slopes with increasing masker level (see Sec. 3.2.5), an offset parameter can be included in Equation (4.1) (see, *e.g.*, Lincoln, 1998; Painter and Spanias, 2000; Vafin et al., 2000). In Figure 4.1, typical SFs are represented for offset values ranging from 0 to 2 dB (from Lincoln, 1998).

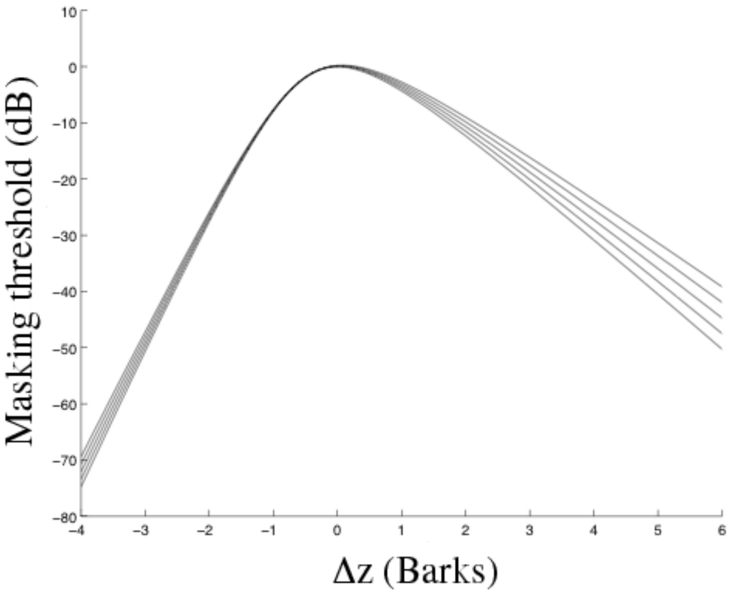


Figure 4.1: Spreading function of masking $SF(\Delta z)$ (see Eq. (4.1)) for various levels of a single sinusoidal component. The masking threshold (in dB) is plotted as a function of Δz, the CB distance between masker and target (in Barks). The offset parameter accounting for the level effect was varied from 0 (bottom curve) to 2 dB (top curve) by 0.5-dB steps. From Lincoln (1998).

4.1.1.2 Combined time-frequency masking model

Although most of the audio coding algorithms use the SF only to predict frequency masking effects, some algorithms were developed to exploit both frequency and temporal masking effects (Lincoln, 1998; Vafin et al., 2000; Huang and Chiueh, 2002; Najaf-Zadeh et al., 2003; He and Scordilis, 2008). Because backward masking effects are very weak compared to forward masking effects (see Sec. 3.3), only forward masking was considered. Forward masking can be modeled in three different ways. The simplest way is to use a linear function of time

$$M(t) = 0.85\, M(t-1) + 0.15\, M(t) \tag{4.2}$$

where $M(t)$ and $M(t-1)$ are the amounts of masking (in dB) in the current and previous time segments, respectively (Lincoln, 1998; He and Scordilis, 2008). The coefficients were adjusted heuristically so that a short-duration masker does not affect $M(t)$ too much, while a long-duration masker sustains masking over several time segments. The temporal masking function in Equation (4.2) is applied to the global masking threshold, that is, after the power addition of the individual masking thresholds due to each component (see Sec. 4.1.2). The individual masking threshold of a component in the z^{th} CB is given by $SF(z)$ (see Eq. (4.1)).

Some algorithms use a linear function of the logarithm of time to exploit forward masking

$$M(\Delta t) = s\, \log(\Delta t) \tag{4.3}$$

where Δt is the time difference between two time segments (depends on the sampling rate), and s is a slope factor that depends on masker level (Vafin et al., 2000; Najaf-Zadeh et al., 2003). Such a model was designed based on the study by Jesteadt et al. (1982). In Vafin et al. (2000), the forward masking model in Equation (4.3) was applied to the global masking threshold in each time segment, while in Najaf-Zadeh et al. (2003) the frequency and forward masking models were combined using a power-law model (see Eq. (3.1)).

Finally, Huang and Chiueh (2002) modeled forward masking with an exponential function and the combined masking model in each CB is given by

$$M(t,z) = \max\ \left\{SF(t,z);\ M(t-1,z)\, e^{-\Delta t/\sigma\tau(z)}\right\} \tag{4.4}$$

where $M(t,z)$ and $M(t-1,z)$ are the amounts of masking (in dB) in the current and previous time segments in the z^{th} CB, respectively, and Δt is the time difference between two time segments. $SF(t,z)$ is the spreading function in the current time segment, $\tau(z)$ is a time constant that characterizes the temporal decay of masking in the z^{th} CB, and σ is the total loudness in the current time segment. The loudness estimation takes into account the absolute threshold in each CB, and the level and duration of the masker (see Eq. (3) and Figs. 2–3 in Huang and Chiueh, 2002). σ is normalized by the total loudness of a 60-dB uniformly masking noise so that the σ value lies between zero and one. When $\sigma = 1$, forward masking decays with the maximum time constant $\tau(z)$. When $\sigma = 0$, there is no forward masking.

4.1.2 Implementations of models in audio codecs

The present section describes the global architecture of an audio codec, namely the MPEG-1 standard, and how the masking models presented above are implemented in such a system. For more detailed descriptions and examples of audio coding algorithms, see, *e.g.*, Painter and Spanias (2000).

The basic structure of an audio codec is schematized in Figure 4.2. The input signal is segmented into quasi-stationary frames ranging from 2 to 50 ms in duration. Then, the TF analysis block estimates the temporal and spectral components in each frame. The TF analysis scheme can consist of a filter bank (*e.g.*, cosine modulated filter banks) or a TF transform (*e.g.*, modified discrete cosine transform, wavelet transform). The TF analysis block works in conjunction with the psychoacoustical model so as to adapt the TF parameters to the human auditory perception. The quantization and encoding block finally compares the original TF parameters to the masking thresholds, and allocates low bit rates in the perceptually irrelevant TF segments.

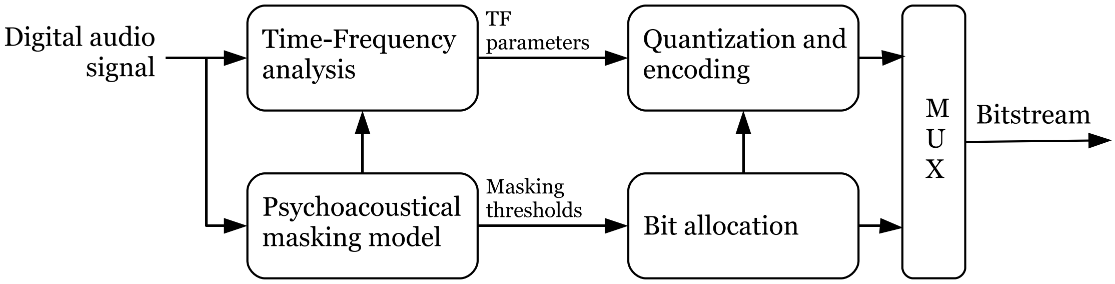


Figure 4.2: Basic structure of an audio codec.

In the MPEG-1 standard, the input signal is first segmented into 12-ms frames. Then, the fast Fourier transform (FFT) of each frame is computed (the FFT resolution depends on the coder but is generally 512 or 1024 points) and scaled in dB SPL. Because the actual playback level remains unknown during the entire signal processing, it is assumed that a 4-kHz signal with ±1 bit amplitude is associated with a SPL of 0 dB (see the normal-hearing threshold curve in Fig. 3.1), while a full-scale sinusoid is associated with a SPL close to 90 dB. To match the spectral analysis of the signal to the spectral analysis by the auditory system (see Sec. 2.3), the whole spectrum is divided into 32 sub-bands simulating the CBs (in the Bark scale).

A third step consists in the identification of tonal and noise maskers in the spectrum of each frame. Local maxima in the power spectrum that exceed neighboring components within a certain Bark distance (Δz) by at least 7 dB are classified as tonal. The selection criterion Δz depends on the frequency region. Δz ranges from 2 Barks in the 0.2–5.5 kHz region to 6 Barks in the 11–20 kHz region. Once tonal maskers have been selected, the remaining spectral peaks are combined (components are linearly added in power) to form a single noise masker in each CB. Then, the total number of maskers is reduced using two criteria. First, any tonal or noise masker whose SPL falls below the absolute threshold is discarded. The absolute threshold at any frequency is well approximated by the function

$$T_q(f) = 3.64\, f^{-0.8} - 6.5\, e^{-0.6(f-3.3)^2} + \frac{f^4}{1000} \tag{4.5}$$

where $T_q(f)$ represents the threshold in quiet of a normal-hearing listener in dB SPL, and f is frequency in kHz (Terhardt, 1979, see the dashed curve in Fig. 4.3). The second criterion to reduce the number of maskers consists in replacing any pair of maskers occurring within a distance of 0.5 Bark by the most prominent of the two.

The next step consists in the application of the psychoacoustical masking models in Section 4.1.1 to compute the individual masking thresholds of the identified maskers. This is achieved by computing the spreading function $SF(z)$ for each masker. Then, a global masking threshold is estimated by combining all the individual masking thresholds. A linear additivity of masking effects is assumed. The global masking threshold is thus computed by adding the individual thresholds in power. An example of global masking threshold estimation on a pop music extract is showed in Figure 4.3.

For algorithms incorporating both frequency and temporal masking, the forward masking model is either applied to the global masking threshold (Lincoln, 1998; Vafin et al., 2000; He and Scordilis, 2008) or combined with the SF in each CB (Huang and Chiueh, 2002; Najaf-Zadeh et al., 2003).

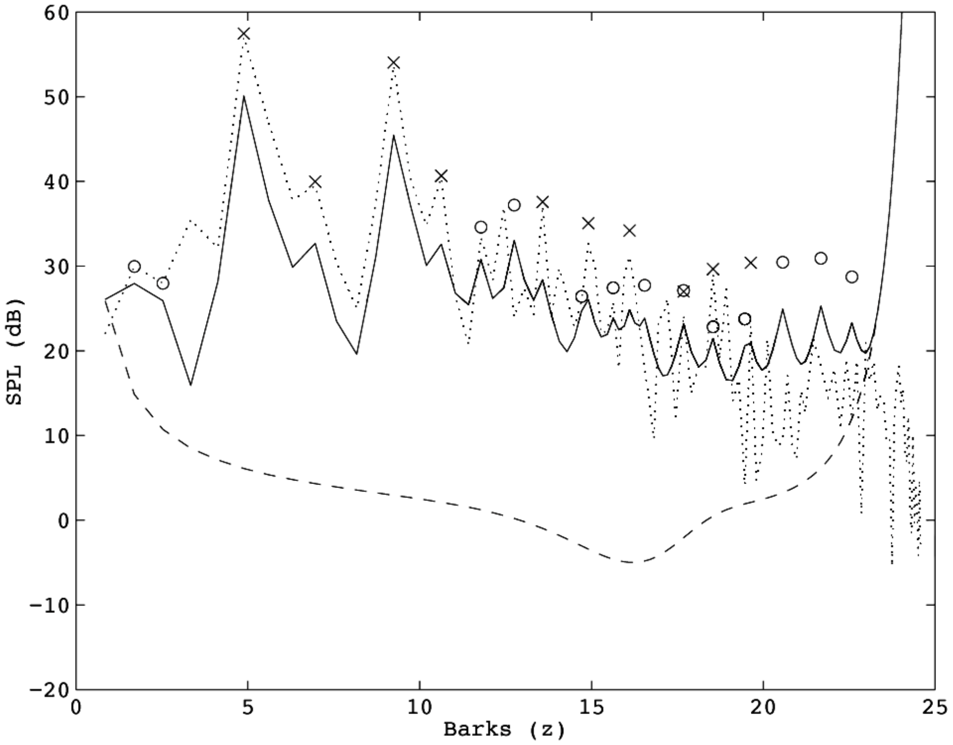


Figure 4.3: Example of a global masking threshold estimation on a pop music selection in the MPEG-1 Layer I codec. The global masking threshold (straight-line curve) is obtained by combining the individual masking thresholds of the maskers (tonal maskers are denoted by 'x'; noise maskers are denoted by '∘') identified in the FFT of the input time frame (dotted curve). The absolute threshold function $T_q(f)$ (see Eq. (4.5)) is superimposed (dashed curve). From Painter and Spanias (2000).

Finally, the global masking threshold allows determining the maximum level of quantization noise that can be introduced in each frame without the noise becoming audible. From the admissible noise level in each frame is determined the bit rate that can be allocated in this frame. The bit rates in successive time frames are combined at the end to form the output audio bitstream.

4.1.3 Limits of current masking models

As described in Chapter 1, TF analysis tools allow the decomposition of any real-world sound into a set of elementary functions or “atoms" well localized in the TF plane. Furthermore, they allow perfect reconstruction of the input signal. Therefore, TF analysis tools play an important role in the signal processing chain (see Fig. 4.2) of many audio applications dealing with the analysis, processing and re-synthesis of non stationary signals. Such applications are, for instance, sound analysis-synthesis systems or *audio codecs*, the main concern of this chapter. For many of these applications, it is important to determine the perceptual relevance of TF components and, in other words, the mutual masking effects between atoms. To address this issue, a TF representation of masking effects for such atoms is required. Nevertheless, it appears that the masking models currently implemented in audio codecs present several limitations as to accurately predict the masking effects between TF components. These limitations are of three types: (1) the spectro-temporal characteristics of the stimuli used in psychoacoustical studies from which masking models were developed, (2) the combination of frequency and temporal masking effects to achieve a combined TF masking model, and (3) the linear additivity of masking effects. These three aspects are revised below.

First, it has to be considered that the masking models presented in Section 4.1.1 are based on psychoacoustical measurements obtained with stimuli which are not maximally compact in the TF domain. Studies of frequency masking mostly used relatively long-duration maskers to keep the spectrum narrow (typically, long-lasting sinusoids). Studies of temporal masking mostly used relatively broadband maskers (typically, broadband noise or clicks), allowing the precise control of the temporal properties of the stimuli. The use of such stimuli is not compatible with the atomic decomposition offered by TF analysis (see Chap. 1). For that reason, a TF representation of masking effects cannot be achieved by combining the results from these experiments in a straightforward approach.

Second, most of the implemented TF masking models are based on a linear combination of frequency and temporal masking effects by superposition of the frequency and forward masking models (except in Najaf-Zadeh et al., 2003). Given the highly non linear behavior of the TF processing by the human auditory system (see Chap. 2), such a simple combination of frequency and temporal masking models is unlikely to predict accurately the masking effects between TF components[1]

Third, most of the implemented models assume a linear additivity of masking effects to compute the global masking threshold in each time frame. The linear additivity of masking effects is not verified in practice, except in particular cases when masker and target fall into the same frequency channel *and* temporally overlap

[1] The deviation between a linear combination of frequency and temporal masking and actual TF masking data is examined in Section 10.4.

(see Sec. 3.4). Instead, the power-law model proposed by Humes and Jesteadt (1989) is generally preferred to take into account the additivity of masking (see Eq. (3.1)). Among the audio coding algorithms presented above, only thatby Najaf-Zadeh et al. (2003) incorporates the power-law model. Nonetheless, the power-law model is also based on psychoacoustical experiments with stimuli which are not maximally compact in the TF domain. In non-simultaneous masking studies, maskers were usually broadband and at least one of the maskers had a long duration. In simultaneous masking studies, maskers were usually narrowband and long. Overall, psychoacoustical data on the additivity of masking for narrowband and short maskers are missing.

It is therefore important to examine how the masking effects produced by a single atom spread across the TF plane, and how the masking effects arising from multiple atoms shifted in time and frequency add up. These issues are the topic of the present study and a collaborative study (Laback et al., 2008), respectively.

4.2 Improving models by measuring time-frequency masking with Gaussian signals

The rationale behind the present study can be stated as follows: given that any real-world sound can be decomposed into a TF matrix of elementary atoms well localized in the TF domain, acquiring knowledge on the “basic” spread of TF masking produced by a single atom would constitute an important advance towards describing and predicting the effective auditory masking in complex sounds. The concept is illustrated in Figure 4.4 where the TF representation of a snare drums sound is represented (Fig. 4.4a) along with the schematic decomposition of this signal into elementary atoms (Fig. 4.4b). The arrows symbolize the basic spread of TF masking produced by a single atom, that is, *the currently missing data*. The obtention of these data may allow predicting the masking interactions between atoms, and thus identifying which components of the matrix in Figure 4.4b are perceptually relevant. In the context of perceptual audio codecs, these data may allow overcoming the limitations of current masking models with the development of a new TF masking model based on actual TF masking data.

Therefore, in the present study, we investigated the TF spread of masking for stimuli with maximal concentration in the TF plane, namely Gaussian-shaped sinusoids. The choice of using Gaussians to measure TF masking and the outline of the study are detailed below.

4.2.1 Advantages of Gaussian signals for estimating the basic spread of time-frequency masking

The accurate prediction of masking effects in the atomic decomposition of an arbitrary sound requires that the spectro-temporal characteristics of the elementary function used for signal decomposition match the spectro-temporal characteristics of the masker used in psychoacoustical experiments. Therefore, a necessary condition for estimating the basic spread of TF masking is to use stimuli (both masker *and*

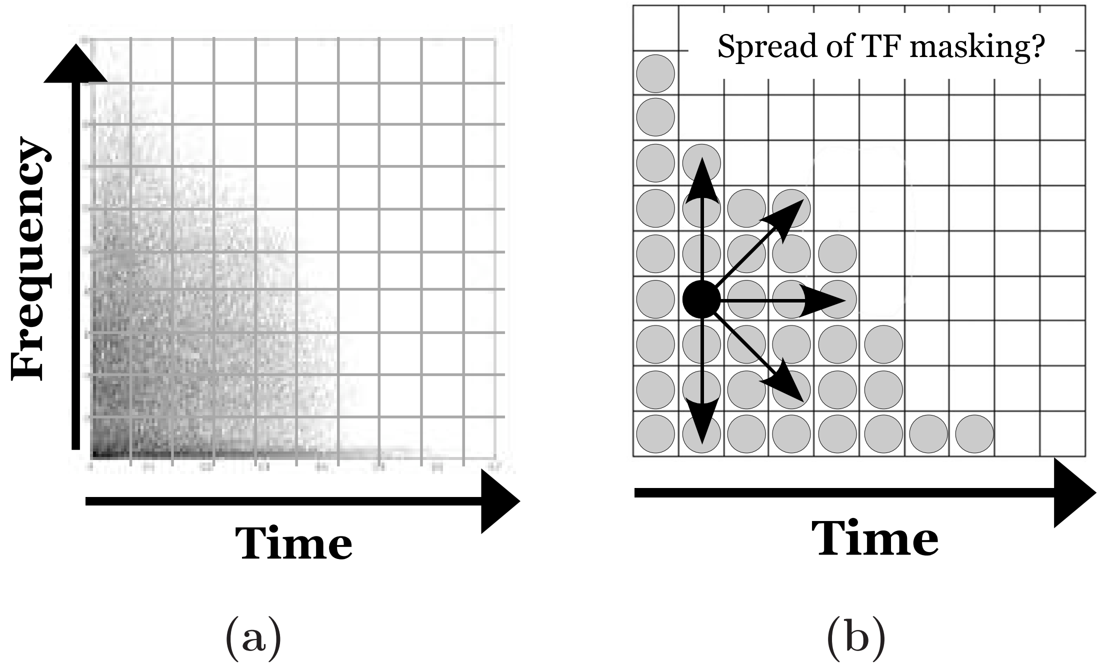


Figure 4.4: (a) TF representation (Gabor transform) of a snare drums sound and (b) schematic decomposition of this signal into elementary atoms well localized in the TF plane. Arrows symbolize the spread of TF masking produced by a single atom.

target) with maximal concentration in the TF domain, *i.e.*, narrowband *and* short signals.

There are few data on the TF spread of masking for narrowband stimuli (Fastl, 1979; Kidd Jr. and Feth, 1981; Soderquist et al., 1981; Lopez-Poveda et al., 2003; Yasin and Plack, 2005). The cited studies involved long (duration $\geq$ 100 ms) sinusoidal maskers *versus* short (duration $\leq$ 20 ms) sinusoidal targets. Because the spectro-temporal characteristics of these maskers do not fulfill the requirement of maximum concentration in the TF domain (see Fig. 4.5), the results from these studies are not suitable for prediction of masking effects between TF atoms.

The signal that has the best localization in the TF plane is the Gaussian. It has Gaussian shapes in both domains and minimizes Heisenberg's uncertainty principle (see Chap. 1). Additionally, narrowband and very short Gaussians are assumed to excite a limited number of spectro-temporal observation windows of the auditory system (van Schijndel et al., 1999) compared to broadband and/or long signals. As an example, Figure 4.5 shows the schematic representation of the concentration in the TF domain of three stimuli with various spectro-temporal characteristics (adapted from van Schijndel et al., 1999). It can be seen that brief Gaussian-shaped sinusoids (middle) provide the best TF localization compared to long-lasting sinusoids (left) and short noise bursts or "clicks" (right).

We used Gaussian-shaped sinusoids (referred to as Gaussians) as masker and target to investigate TF masking. The work by van Schijndel et al. (1999) served as a basis for the design of the stimuli, as detailed in Chapter 5.

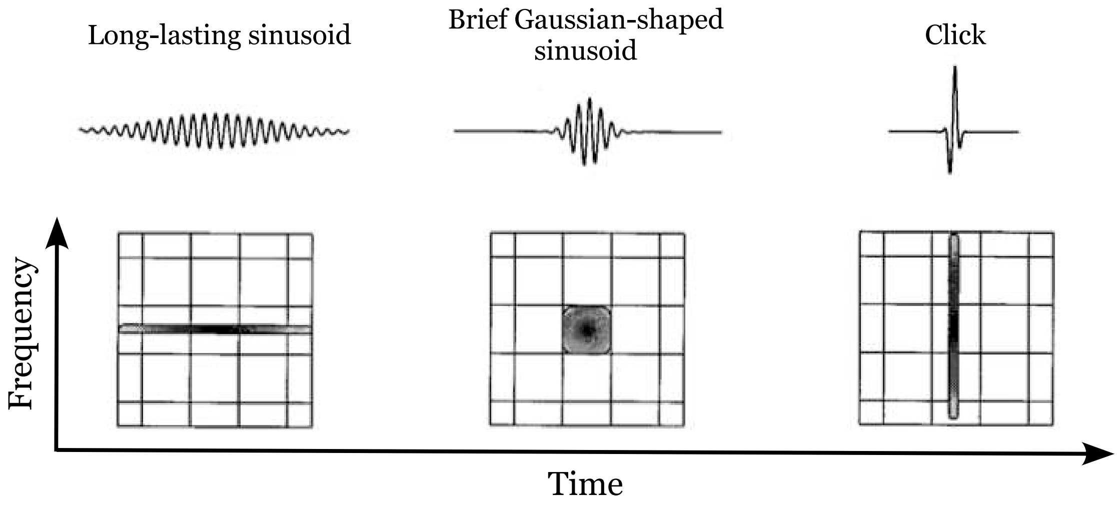


Figure 4.5: Schematic representation of the concentration in the TF domain of three stimuli with various spectro-temporal characteristics: a long-lasting sinusoid (left), a brief Gaussian-shaped sinusoid (middle) and a click (right). The grid schematizes the "internal" partitioning of the auditory system into TF observation windows. Adapted from van Schijndel et al. (1999).

4.2.2 Outline of the study

Six experiments were conducted. In Experiment 1, absolute thresholds were measured for 11 carrier frequencies of the Gaussian target. We examined how absolute thresholds for short Gaussian-shaped sinusoids compare to those for long, steady-state ones. In Experiment 2, masked thresholds were measured as a function of the frequency separation (ΔF) between masker and target, which were presented simultaneously ($\Delta T = 0$). The masker had a carrier frequency of 4 kHz and a level of 60 dB SL. In Experiments 3 and 4, we examined the effects of masker level and masker frequency on simultaneous masking, respectively. The conditions from Experiment 2 were replicated with masker levels of 30, 45 and 60 dB SL (Exp. 3) or with a masker frequency of 0.75 kHz (Exp. 4). In Experiment 5, masked thresholds were measured as a function of the temporal separation (ΔT) between masker and target, which had the same frequency ($\Delta F = 0$). In Experiment 6, both ΔF and ΔT were varied.

In a collaborative study (Laback et al., 2008), we focused on the additivity of masking effects arising from up to four "equally effective" Gaussian maskers distributed in the time or frequency plane.

A first attempt was made to model the gathered masking data for implementation in a sound signal processing algorithm allowing to remove the perceptually irrelevant atoms in the TF representations of audio signals.

Summary

A well-known application of psychoacoustical masking data in the field of sound signal processing is audio coding. To reduce the digital size of audio files, audio codecs (like MPEG-1 Layer III) use TF analysis tools to decompose arbitrary complex sounds into TF matrices of elementary atoms well localized in the TF plane, and exploit the properties of auditory masking to allocate low bit rates in TF channels which are subject to masking. To date, most of the audio coding algorithms use simple models of frequency masking only to predict the masking effects between TF components. These models are based on psychoacoustical measurements obtained with stimuli which do not have a maximal concentration in the TF domain. The use of such stimuli is not compatible with the atomic decomposition offered by TF analysis. Overall, audio codecs require psychoacoustical data on the actual spread of masking produced by a single atom to accurately predict the masking effects between TF components.

We therefore propose a new approach, namely to obtain a measure of the "basic" spread of TF masking produced by a single atom. This approach requires that the stimuli used in masking measurements (both masker and target) have a maximal concentration in the TF domain. Because existing studies on TF masking mostly used narrowband and long maskers, their results are not suitable for prediction of masking effects between atoms.

The signal that has the maximal concentration in the TF plane is the Gaussian. Therefore, the present study involved Gaussian-shaped sinusoids as both masker and target to investigate TF masking. Six experiments were conducted which examined the effects of the frequency and/or temporal separation between masker and target, and the effects of masker level and frequency on masking. These data may allow accurately predicting the masking interactions between TF components, and thus improving the current masking models with the development of a new TF model.

Part II

EXPERIMENTAL CONTRIBUTION

Contents of the Second Part

Chapter 5

Definition of the Gaussian signals used in psychoacoustical experiments

Contents

5.1 Formula used for stimuli generation

Throughout the experiments, both the masker and target were Gaussian-shaped sinusoids (referred to as Gaussians) defined by

$$s(t) = \sqrt{\Gamma} \sin\left(2\pi f_0 t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2} \tag{5.1}$$

where f_0 is the carrier frequency and $\Gamma = \alpha f_0$ (van Schijndel et al., 1999). For a given f_0, the shape factor of the Gaussian window, α, allows controlling both the duration and bandwidth of $s(t)$. The equivalent rectangular bandwidth of the Gaussian window, “ERB_{GW}", is Γ. Its equivalent rectangular duration, “ERD_{GW}", is Γ^{-1}. By introducing the $\pi/4$ phase shift, the energy of the signal is independent of f_0. In Figure 5.1, the signal $s(t)$ with $f_0 = 4$ kHz and $\alpha = 0.15$ ($\Gamma = 600$) is represented in the time (Fig. 5.1a), frequency (Fig. 5.1b), and TF (Fig. 5.1c) domains.

5.2 Stimuli parameters

The shape factor α of the Gaussian was chosen based on the study by van Schijndel et al. (1999), who used Gaussians to measure spectro-temporal integration. On the basis of the “multiple looks” model (see Sec. 2.2.2.4 and Viemeister and Wakefield 1991), van Schijndel et al. (1999) made the assumption that the auditory system performs a TF analysis through its own TF windows. They attempted to characterize the shape of these elementary TF observation windows by assessing just-noticeable differences in intensity (“jnd(I)”) for Gaussians with

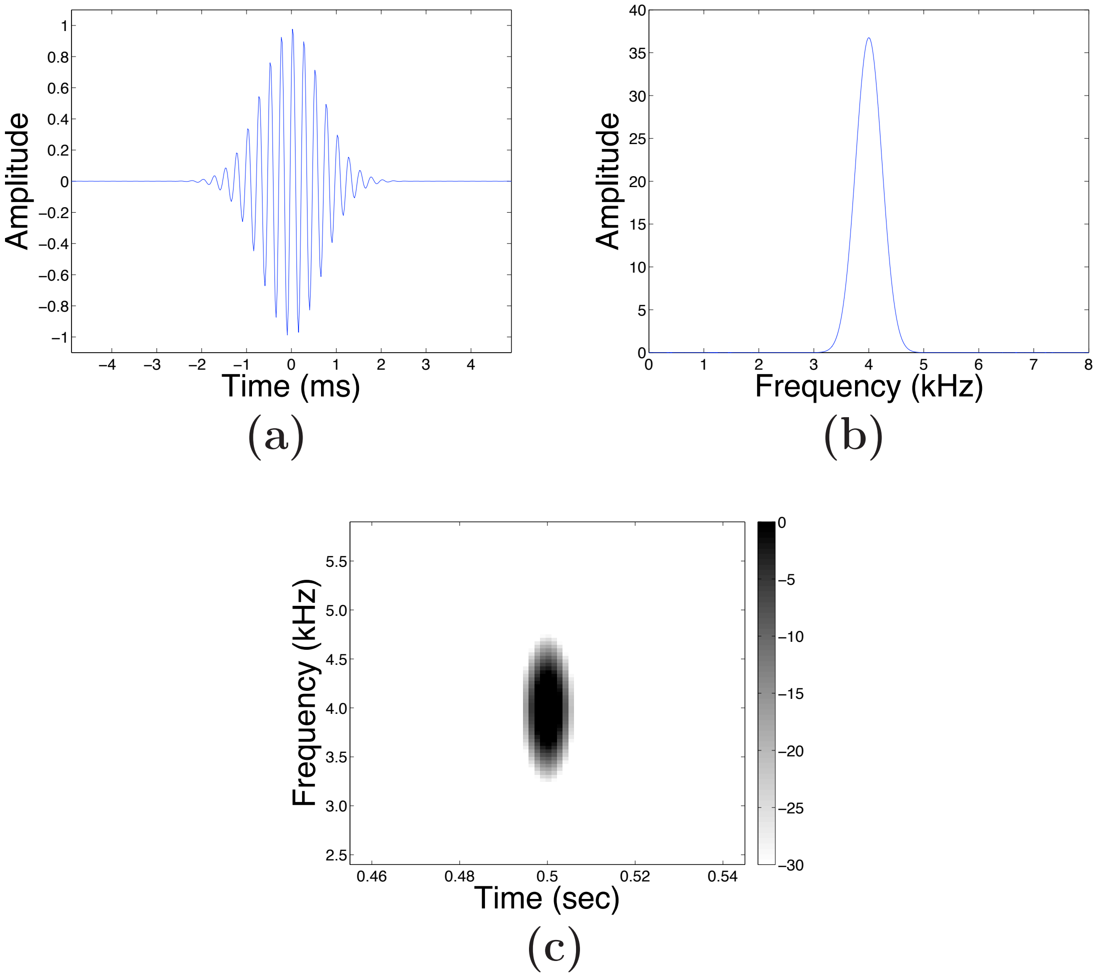


Figure 5.1: A Gaussian-shaped sinusoid (signal $s(t)$ in Eq. (5.1)) with $f_0 = 4$ kHz and $\alpha = 0.15$ ($\Gamma = 600$) is plotted (a) as a function of time, (b) as a function of frequency (modulus of the Fourier transform), and (c) in the TF domain (module of the Gabor transform, in dB).

various spectro-temporal shapes (*i.e.*, for various α values) at carrier frequencies of 1 and 4 kHz. Their model predicted that intensity discrimination thresholds should depend upon the number of "internal" windows covered by a signal such that jnd(I) be highest for stimuli covering one window. Table 5.1 lists the spectro-temporal characteristics of the Gaussians for each value of α tested by van Schijndel et al. (1999) at 1 and 4 kHz. The last column labeled "#TF windows" provides the estimated number of internal TF windows covered by each Gaussian. It was determined by assuming that a Gaussian window with a shape factor α of 0.23 has a bandwidth of $\frac{1}{3}$ octave and an ERD_{GW} of 4 ms at 1 kHz and 1 ms at 4 kHz: such a window covers about one CB along the frequency axis and about one "look" on the time axis.

α	f_0 (Hz)	ERD_{GW} (ms)	ERB_{GW} (Hz)	#TF windows
0.0375	1000	27.0	37.5	7
	4000	6.7	150.0	7
0.075	1000	13.0	75.0	3
	4000	3.3	300.0	3
0.15	1000	6.7	150.0	2
	4000	1.7	600.0	2
0.3	1000	3.3	300.0	1
	4000	0.83	1200.0	1
0.6	1000	1.7	600.0	3
	4000	0.42	2400.0	3
1.2	1000	0.83	1200.0	6
	4000	0.21	4800.0	6

Table 5.1: Set of Gaussians used by van Schijndel et al. (1999) to estimate the shape of the auditory TF observation windows. For each value of α are given (from left to right) the carrier frequency f_0, the equivalent rectangular duration ERD_{GW}, and the equivalent rectangular bandwidth ERB_{GW} of the Gaussian-window. The last column indicates the estimated number of TF windows covered by each Gaussian.

At both carrier frequencies, van Schijndel et al. (1999) obtained the highest jnd(I) for a "critical" α value comprised between 0.15 and 0.30. These results suggest that a Gaussian window with an α value of 0.15 approximates the shape of the "elementary" TF observation window of the auditory system. This window has a spectral width of roughly one CB and a temporal width that approximately corresponds to four periods of the carrier frequency: 4 ms at 1 kHz and 1 ms at 4 kHz.

In the present study, f_0 varied depending on the frequency separation (ΔF) between masker and target. By keeping α constant, Γ would have varied with f_0, and the stimulus bandwidth would have been proportional to the CB. However, because this would have caused the durations of the stimuli to vary, we decided instead to fix the Γ value so as to keep ERB_{GW} and ERD_{GW} constant. In Experiment 2, 3, 5 and 6, the carrier frequency of the masker (F_M) was fixed at 4 kHz. ERB_{GW} was set to 600 Hz, corresponding to an ERD_{GW} of 1.7 ms. In Experiment 4, F_M was fixed to 0.75 kHz. In this case, ERB_{GW} was 112.5 Hz, corresponding to an ERD_{GW} of 8.9 ms.

5.3 Temporal windowing of signals

Because a Gaussian window has an infinite support (*i.e.*, $s(t)$ is of infinite duration), a strategy had to be found to temporally window the signal $s(t)$ while preserving the properties of the Gaussian. This was achieved using a numerical optimization procedure developed by Depalle and Hélie (1997), who designed a family of windows with no spectral sidelobes based on Gaussian functions.

Although a Gaussian window is not time limited, it has no spectral sidelobes and quickly tends towards zero. Multiplying a Gaussian function $GW(t) = e^{-\pi(\Gamma t)^2}$ with any window w in the temporal domain results in the convolution product of the respective Fourier transforms of GW and w (see Eq. (1.7)). When the spectral width of the Gaussian is large enough, the shallow decay removes the sidelobes of w after convolution. The no-sidelobe windows designed by Depalle and Hélie (1997) used the power of a triangular window as w. However, in the present study, the properties of the Gaussian window had to be retained, which was not possible using a triangular function. Then, w was defined as the two-parameter Tukey window, $w(N, r)$, whose sample number k is computed according to

$$w(N,r)(k+1) = \begin{cases} 1.0 & 0 \leqslant |k| \leqslant \frac{N}{2}(1+r) \\ 0.5\left[1.0 + \cos\left[\pi\,\frac{k-\frac{N}{2}(1+r)}{N(1-r)}\right]\right] & \frac{N}{2}(1+r) \leqslant |k| \leqslant N \end{cases} \tag{5.2}$$

$$N, k \in \mathbb{Z}, \;\; r \in [0,1]$$

where N is the number of samples, and r is the ratio of cosine-tapered to constant sections, comprised between 0 and 1 (Harris, 1978). If $r = 0$, w is a rectangular window with abrupt transitions. If $r = 1$, w is a Hanning window with no constant section. In the present study, the bandwidth of the Gaussian window was fixed by Γ and r was set to 0.1. The procedure therefore computed the smallest value of N (thus the shortest duration of w) corresponding to Γ which avoided the sidelobes. At a sampling rate of 48 kHz, the optimal estimates of N were 472 samples for $\Gamma = 600$ ($f_0 = 4$ kHz), and 2476 samples for $\Gamma = 112.5$ ($f_0 = 0.75$ kHz). The corresponding "effective durations" (defined as the 0-amplitude points durations) of the stimuli were 9.6 ms and 51 ms, respectively.

Figure 5.2a shows a 9.6-ms Tukey window ($w(472, 0.1)$, in red) and a Gaussian function $GW(t)$ with $\Gamma = 600$ multiplied by w (in blue) as a function of time. Figure 5.2b shows the respective Fourier transforms of w (in red) and $w\,GW(t)$ (in blue). Specifically, Figure 5.2b shows how the sidelobes of w have been considerably smoothed by the convolution with the Fourier transform of GW. Furthermore, it can be seen that the main lobe of the Gaussian function (in blue) emerges from the asymptotic level by about 250 dB.

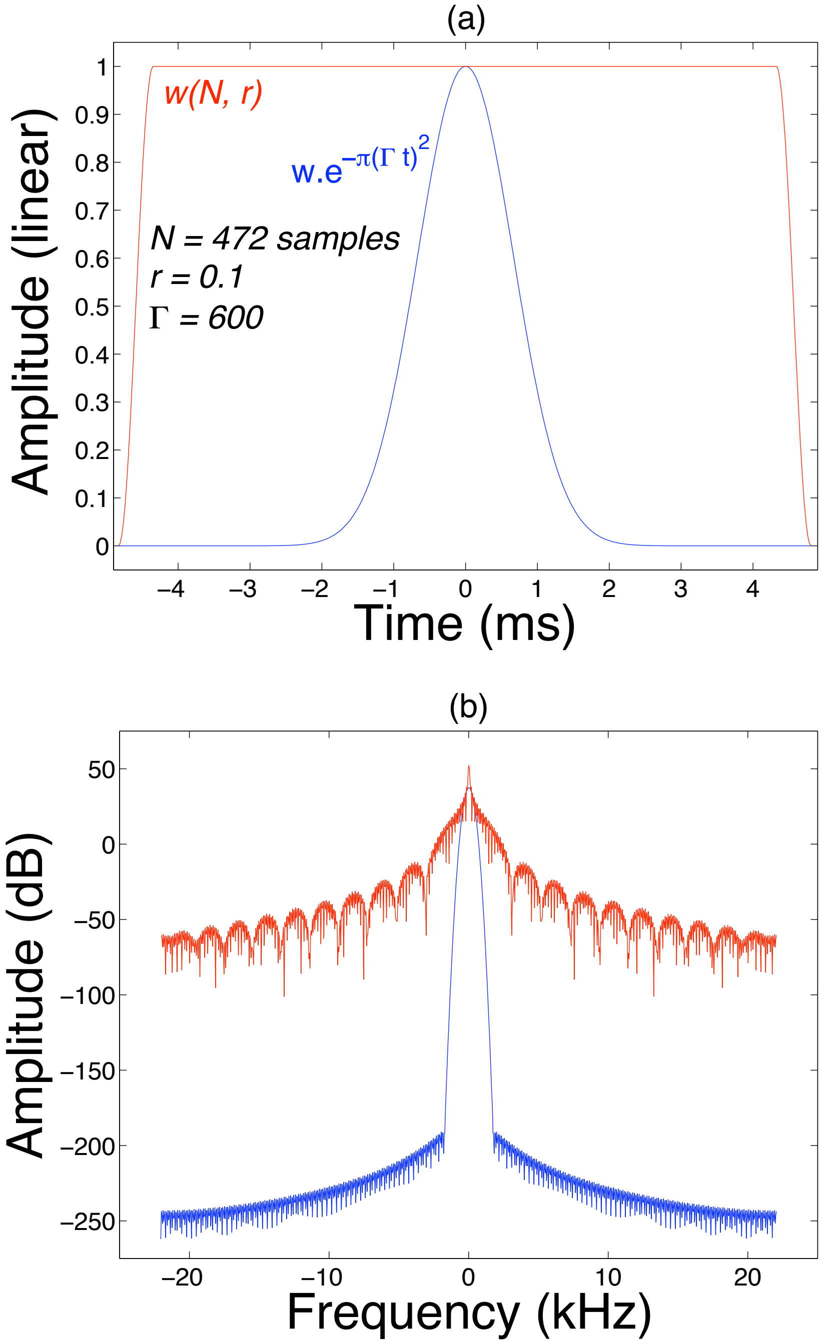


Figure 5.2: Temporal windowing of a Gaussian function $GW(t) = e^{-\pi(\Gamma t)^2}$ with $\Gamma = 600$ and $t = k/F_S$ ($F_S = 48$ kHz) by a 472-sample Tukey window (effective duration $= 9.6$ ms). (a) The Tukey window $w(472, 0.1)$ (in red) and $w\,GW(t)$ (in blue) are plotted as a function of time. (b) Fourier transforms of w (red) and $w\,GW(t)$ (blue).

Summary

The masker and target signals used throughout the experiments were Gaussian-shaped sinusoids defined by

$$s(t) = \sqrt{\Gamma} \sin\left(2\pi f_0 t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2}$$

where f_0 is the carrier frequency and Γ is a parameter allowing to control both the duration and bandwidth of $s(t)$. In our study, f_0 varied depending on ΔF. The Γ value was fixed so as to keep the equivalent rectangular bandwidth ($ERB_{GW} = \Gamma$) and the equivalent rectangular duration ($ERD_{GW} = \Gamma^{-1}$) constant. By introducing the $\pi/4$ phase shift, the energy of the signal was independent of f_0. Since a Gaussian window has an infinite support, a strategy had to be found to temporally window the signal $s(t)$ while preserving the properties of the Gaussian. This was achieved using a numerical optimization procedure developed by Depalle and Hélie (1997), who designed a family of windows with no spectral sidelobes based on Gaussian functions. The signals were windowed in the time domain using a Tuckey window.

The table below summarizes the spectro-temporal characteristics of the stimuli for each masker frequency (F_M) and corresponding Γ value used in the experiments. The last column indicates the effective duration (*i.e.*, 0-amplitude points duration) of the "optimal" Tukey window.

F_M (kHz)	Γ	ERB_{GW} (Hz)	ERD_{GW} (ms)	Effective duration (ms)
4.0	600.0	600.0	1.7	9.6
0.75	112.5	112.5	8.9	51.0

Chapter 6

General procedure used in psychoacoustical experiments

Contents

When an acoustic signal reaches the ear of a human listener, it results in an auditory sensation in that listener. Varying the physical parameters of the acoustic signal inevitably affects the auditory perception of that signal. Because there is not a linear relationship between the physical attributes of sounds and the auditory sensation, it is important to describe the perceptual correlates of the physical parameters of sounds, and attempt to explain the underlying mechanisms of auditory processing. This is the concern of psychoacoustics. Precisely, psychoacoustics considers the auditory system as a complex process that converts a physical *input* (the sound stimulus) into a perceptual *output* (the auditory sensation), and attempts to establish a link between both. This is schematized in Figure 6.1.

Sound stimulus
input
Auditory system
Perceptual sensation
output

Figure 6.1: Schematic illustration of the basic concept of psychoacoustics, in which a sharp distinction is made between the physical input and the perceptual response to it.

Because the system output in Figure 6.1 cannot be measured directly, psychophysical methods were developed to evaluate how a listener perceives the variation of one of the physical attributes of a sound stimulus. In an experimental situation, the experimenter must always establish a clear relationship between the sound presented and how the listener actually perceives it. For examples, an increase in signal level should be perceived as an increase in loudness, or a decrease in signal frequency should be perceived as a decrease in pitch. Furthermore, to minimize the effects of response bias, the experimenter must always inform the listener about the stimulus he will hear, the procedure used and the response he must provide.

In Section 3.1, the notions of absolute and masked thresholds were presented. To determine an absolute threshold, the detection of a target signal in quiet is measured. This detection is characterized by the notion of threshold, which is the limit between “audible” and “inaudible”. To determine a masked threshold, the detection of the target is measured in the presence of a masker. Several psychophysical methods were designed for determining detection thresholds. The most frequently employed methods in psychoacoustics are shortly presented below. Then, the method selected for the masking measurements in the present study is detailed.

6.1 Classical psychophysical methods of thresholds determination

There are four types of psychophysical methods classically employed to estimate thresholds: (1) the method of limits, (2) the method of adjustment, (3) the method of constant stimuli and (4) the adaptive method. These methods are presented below. For a detailed review on the psychophysical measurement methods, see, *e.g.*, Robinson and Watson (1973), or Gelfand (1998, Chap. 7).

6.1.1 The method of limits

In this method, the stimulus level is under the experimenter’s control and the listener simply responds after each presentation to indicate whether he did hear (+) or not (-) the target stimulus. The method consists in presenting to the listener several stimulus series in which the target level is *ascending* or *descending*. An equal number of ascending and descending series is generally achieved, with a minimum of three. In a descending series, the target level starts well above the expected threshold and decreases by a discrete amount of dB (usually 2 dB) as long as the listener detects it. The series stops when the listener provides an incorrect response (-), and the resulting threshold corresponds to the mean target level between the last (+) and the first (-). In an ascending series, the process is reversed. Finally, the listener’s threshold is obtained by averaging the threshold levels across series.

The method of limits has the advantage of being the least time-consuming one. However, it comprises two response biases. First, as a series either ascends or descends and is terminated by a single change in response, the listener may “anticipate" the level at which his response should change from (-) to (+) in an ascending series, resulting in a lower threshold estimate, or from (+) to (-) in a descending one, thus resulting in a higher threshold. The second bias is the opposite effect. In an ascending series, the listener can persist in responding (-) although his actual threshold has been exceeded for a few trials, which raises the measured threshold. Similarly, in a descending series, the listener can persist in responding (+) for a few trials after the sound became actually inaudible, which lowers the measured threshold. This effect is usually called “habituation". However, these biases may be minimized by alternatively or randomly presenting a large number of ascending and descending series to the listener, and by varying the starting levels across series.

6.1.2 The method of adjustment

In the method of adjustment, the stimulus level is controlled by the listener. The target level is varied continuously via a potentiometer rather than in discrete steps. Otherwise, the procedure is similar to that in the method of limits: the target level is decreased from above threshold until the target is not detectable, or increased from below threshold until the target is just detectable. The final threshold is taken as the average of the just detectable and not detectable levels.

To prevent response bias due to the potentiometer, the knob must be unlabeled and should provide no tactile cue. Moreover, the experimenter usually has a second control by varying the starting level across series. As for the method of limits, the anticipation and habituation effects are still present and may be minimized by achieving several ascending and descending series.

6.1.3 The method of constant stimuli

Unlike the methods of limits and adjustment, the method of constant stimuli is a non sequential procedure. Specifically, the stimuli are not presented in an ascending or descending manner but in a random order. A range of levels is tested, which is supposed to encompass the threshold level based on a pilot experiment. Then, the stimuli are presented in random order to the listener, who indicates on each trial whether he has detected the target stimulus or not. At each selected level, an equal number of stimuli is presented. The resulting psychometric function, which shows the percentages of correct responses as a function of stimulus level, is then plotted, and the listener's threshold is taken as the 50-% point. An example of absolute thresholds determination using the method of constant stimuli is illustrated below. The collected data are displayed in Table 6.1 and the resulting psychometric function is plotted in Figure 6.2 (from Gelfand, 1998, Chap. 7).

To avoid response biases, the experimenter checks if the listener is fully concentrated on the task by presenting what is called “catch" trials. These are intervals during which the listener is asked whether a tone was heard when no tone was actually presented.

Compared to the two methods described above, the method of constant stimuli has the advantage of being more precise and avoiding (but not completely removing) the response biases. However, it has the disadvantage of being highly time consuming. Indeed, a very large number of trials is needed (and these trials are multiplied by the number of levels tested), so that this method increases the effects of listener's fatigue and the difficulty of keeping him fully concentrated on the task.

Stimulus level (dB)	Number of (+) responses	Percentage of (+) responses
11	50	100
10	50	100
9	47	94
8	35	70
7	17	34
6	3	6
5	0	0
4	0	0

Table 6.1: Absolute thresholds determination using the method of constant stimuli with a step size of 1 dB. The middle and right columns indicate the number and the percentage of correct (+) responses, respectively.

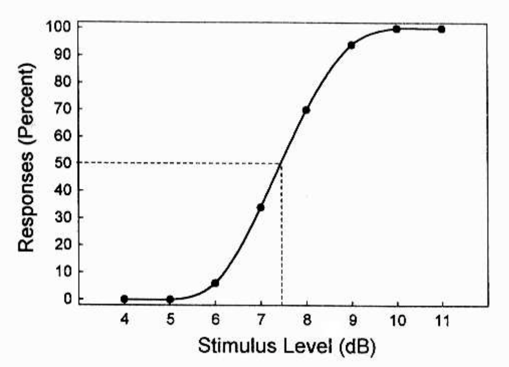


Figure 6.2: Psychometric function based on the data from Tab. 6.1. The threshold, taken as the 50-% point, is 7.5 dB.

6.1.4 The adaptive methods

In adaptive procedures, the level at which a particular stimulus is presented to the listener depends upon how he responded to the previous trials. Such methods allow to quickly converge upon the listener’s threshold by placing most of the trials close to it, and are independent of the starting level. Efficiency and precision are thus maximized.

6.1.4.1 The Békésy’s tracking method

The automated von Békésy’s tracking method (1960) was the first adaptive procedure designed. This method profits from both the limits and adjustment methods, and is currently implemented in most of the professional audiometers to evaluate hearing thresholds. Because the von Békésy’s tracking method is not very appropriate for measurements with very brief stimuli, it is not detailed further in this document. More appropriate methods are the simple and transformed “up-down" procedures (Levitt, 1971), presented below.

6.1.4.2 The simple up-down procedure

The simple up-down procedure is similar to the method of limits in that the stimulus level is decreased (or increased) after a correct (or incorrect) response by a certain step size. The difference lies in the fact that a series does not stop after the first reversal (from (+) to (-) or from (-) to (+)), but continues up to at least six reversals. The listener’s threshold is then taken as the average of the stimulus levels at those reversals[1]. Because each correct response leads to a decrease in stimulus level and each incorrect response leads to a level increase, the simple up-down procedure converges upon the 50-% point on the psychometric function. This means that (+) and (-) have the same probabilities of response. To increase both the precision and convergence quickness of the method, a large step size is usually chosen at the beginning of the series (*e.g.*, 5 dB up to the second reversal). This initial step size is then decreased (*e.g.*, to 2 dB) for the remaining of the series.

6.1.4.3 The transformed up-down procedures

The simple up-down procedure estimates the threshold at the 50-% point on the psychometric function. To obtain a more reliable threshold estimate, the simple up-down rule can be modified to converge upon another point on the psychometric function, where the probability of correct responses is greater than 50%. This is achieved by changing the stimulus level only if a certain sequence of correct or incorrect responses has been provided by the listener (Levitt, 1971). The criterion for decreasing the stimulus level is called the “down rule", and that for increasing the stimulus level the “up rule". Examples of “transformed up-down" strategies are given in Table 6.2. Each entry in Table 6.2 corresponds to an estimation point on the psychometric function. For a detailed computation example of the probability of correct responses see, *e.g.*, Gelfand (1998, Chap. 7). Figure 6.3 illustrates how the “1 up - 3 down" procedure (Entry 3 of Tab. 6.2) converges upon the 79.4-% point on the psychometric function. As for simple up-down procedures, transformed procedures are run up to at least six reversals and the average of the stimulus levels at those reversals is taken as the listener’s threshold.

There are two major aspects that were not specified in the description of the simple and transformed up-down procedures. These are (1) how stimuli are presented and (2) how responses are collected. This can be accomplished by using the “yes/no" task described above. In this task, the listener simply reports after each trial whether he did detect (+) the target stimulus or not (-). However, the most frequently used paradigm used in psychoacoustics is the forced-choice method. It consists in presenting two or more alternatives from which the listener must choose a response. For example, in the determination of a masked threshold using a two-alternative forced choice (2-AFC) task, each trial consists in two intervals presented successively. One of the intervals, randomly selected across trials, contains the masker *and* the target, while the other interval contains the masker alone. The listener must then indicate in which interval the target was presented. If the listener selects the correct interval, then a correct (+) response is provided. These procedures are often called 2- (or more) interval forced-choice methods (2-IFC, 3-IFC, etc. . .).

[1] If the total number of reversals is even, then the first two – or four – reversals are omitted from the average computation. If the total number is odd, then only the first reversal is omitted.

	Response sequences		
Entry	"Up rule" Increase level after:	"Down rule" Decrease level after:	Probability of correct responses at convergence (%)
1	-	+	50.0
2	+ - or -	+ +	70.7
3	+ + - or + - or -	+ + +	79.4
4	+ + + - or + + - or + - or -	+ + + +	84.1

Table 6.2: Examples of transformed up-down strategies. For each entry are given the possible response sequences for the "up" and "down" rules, and the corresponding percent of probability of correct responses at convergence. Adapted from Levitt (1971).

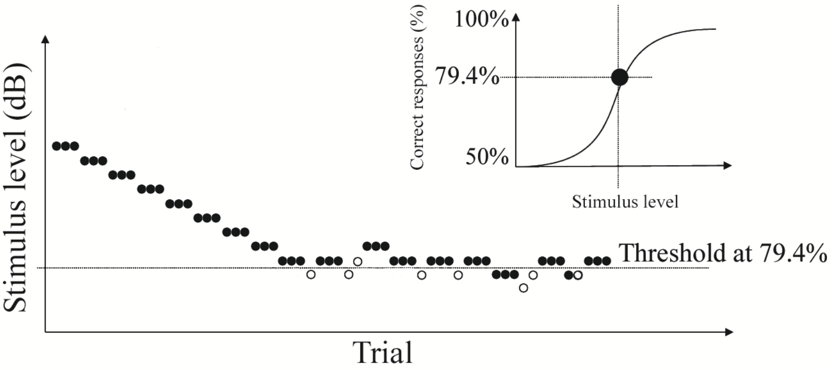


Figure 6.3: Illustration of how the "1 up - 3 down" procedure (Entry 3 of Tab. 6.2) converges upon the 79.4-% point on the psychometric function. Each trial is indicated by a circle. Full circles are correct responses (+) while empty circles are incorrect responses (-). In this example, the series ran up to twelve reversals. The corresponding psychometric function is given in the insert.

Summary

The table below summarizes the properties of the common psychophysical methods used in psychoacoustics for thresholds determination. The precision of each method is indicated by the percentage of correct responses. Because the duration of an entire series roughly depends on the stimulus duration (and on the up-down rule in case of IFC tasks), the time consumption of a given method cannot be simply quantified. Therefore, the time consumption of each method is specified here with a qualitative adjective. The magnitude of the time scale varies from “weak" to “huge".

Method	Percentage of correct responses	Time consumption	Biases
Limits	50.0	Weak	Anticipation, habituation
Adjustment	50.0	Weak	Anticipation, habituation, labeled knob, tactile cues
Constant stimuli	50.0	Huge	Fatigue
Simple “up-down"	50.0	Moderated	Anticipation habituation
Transformed “up-down"	70.7 and above	Large	Fatigue

6.2 Details of the adaptive procedure chosen for masking measurements

A transformed up-down procedure with a “1 up - 3 down" rule (see Entry 3 of Tab. 6.2 and Fig. 6.3) was elected for the masking measurements carried out in this PhD work. This method converges upon the 79.4-% point on the psychometric function and thus provides reliable threshold estimates. It was also chosen because (1) it minimizes response biases and (2) fatigue effects can be moderated by splitting the total number of series (one threshold estimate took about 4–6 min) into sessions and running 1–2 sessions per day.

Masked thresholds were initially estimated using a 2-IFC task. Each trial consisted of two 200-ms observation intervals indicated by lights on the response box. The two intervals were separated by a 800-ms gap. The masker was presented in the two intervals and the target was presented with the masker in one of those intervals, chosen randomly. The listener indicated which interval was thought to contain the target by pressing one of two buttons on a response box. Immediate feedback was provided by lights on the response box. A series started with a target level about 10–15 dB above the expected threshold value, as determined by practice trials.

The target level varied adaptively (*i.e.*, decreased after three correct responses, and increased after one incorrect response) by initial steps of 5 dB and 2 dB following the second reversal. Twelve reversals were obtained. The threshold estimate was the mean of the target levels at the last 10 reversals. A threshold estimate was discarded when the standard deviation of these 10 reversals exceeded 5 dB. Before data collection, practice series were performed for a number of conditions until the threshold estimates become stable (*i.e.*, stopped decreasing or fluctuating by more than 5 dB with repetition of the series). Then, two threshold estimates were obtained for each condition. If the standard deviation of these two estimates exceeded 3 dB, up to four additional series were completed until the standard deviation of the total number of estimates (maximum = 6) be less than 3 dB. If the standard deviation still exceeded 3 dB after six repetitions, the threshold was conserved as the average of the six estimates. In the experimental data reported below, the few data points for which the standard deviation exceeded 3 dB are precised. Absolute thresholds for the Gaussian targets were estimated using the same 2-IFC task. Silent intervals were marked by lights on the response box and no practice series was performed.

However, pilot tests with the 2-IFC task on four highly trained listeners revealed large within-listener variability (5–10 dB) in some experimental conditions, although several hours of testing had been completed. This can be explained by some possible confusion effects between masker and target (Neff, 1985). Remind that the present masking experiments involved masker and target signals with identical spectro-temporal characteristics (see Chap. 5). Thus, in some critical conditions where masker and target were presented simultaneously (or with a short inter-stimulus delay) *and/or* with little or no frequency difference, listeners could hardly distinguish the target from the masker (or, which was the target). To avoid such confusions between masker and target, studies on masking generally involve masker and target signals that differ either in the spectral domain (broadband masker *vs.* narrowband target), in the temporal domain (long-duration masker *vs.* short-duration target), or both (see Chap. 3 and Weber and Moore 1981; Moore and Glasberg 1983a). Therefore, to provide the listeners an additional “cue”, the initial 2-IFC task was changed into a 3-IFC task. The general procedure described above remained the same except that the masker was presented in three observation intervals instead of two. Hence, listeners could hear the masker or “reference interval" twice so as to discern the interval containing the target more easily than with the 2-IFC task. The use of the 3-IFC task indeed reduced within-listener variability, even for naive listeners.

Overall, both absolute and masked thresholds were estimated using a 3-IFC task with a “1 up - 3 down" rule. In all experiments, listeners completed 1–2 sessions of 30 min per day. A session consisted of six threshold estimates.

6.3 Apparatus

All stimuli were digitally generated at a 48-kHz sampling rate and a 24-bit resolution using a Tucker-Davis Technologies System III including:

– A real-time processor (RP2.1)
– Two programmable attenuators (PA5)
– A signal mixer (SM3)

– A headphone buffer (HB7)
– A response box (Ebox)

The entire procedure was piloted by a Delphi program running on a PC. Masker and target were computed in Delphi and routed to two different channels of the processor and two digital-to-analog converters (DAC). When a continuous noise was needed to mask cochlear distortion products, a white noise was generated in real time, lowpass-filtered (16th order digital Butterworth filter) and released through the masker's channel. The outputs of the two DACs were attenuated, added, and routed to the headphone buffer and to the right ear-pad of a circumaural headphone (Sennheiser HD545). The headphones were calibrated so that levels were defined as SPL close to the eardrum. Listeners were tested individually in a double-walled, sound-attenuated booth.

6.4 Remarks on the stimulus level and SPL definition

To minimize quantization effects, stimuli were generated at the maximum voltage range of the processor outputs (±10 Volts peak for TDT RP2.1), attenuated and passed to the headphone buffer. The SPL (close to the eardrum) was deduced from the RMS (Root Mean Square) voltage of the signals using a headphone-specific calibration file[2]. The attenuation value necessary to obtain the desired SPL was then computed and sent to the attenuation device (TDT PA5).

For example, consider $s(k)$ as a N-sample Gaussian signal defined in Equation (5.1). Its RMS voltage is

$$V_{S_{RMS}} = \sqrt{\frac{1}{N}\sum_{k=1}^{N} s(k)^2} \tag{6.1}$$

with k being the sample number. The SPL corresponding to $V_{S_{RMS}}$ is then

$$L_{S_{MAX}} = X_{f_0} + 20\log_{10}(V_{S_{RMS}})$$

where X_{f_0}, provided by the calibration file, is the SPL of a 1-Volt RMS stationary sinusoid with frequency f_0. Hence, to present a signal with a SPL of L_S dB to the listener, the required attenuation level (in dB) is $L_{S_{MAX}} - L_S$. The same operation was used to fix the background noise level except that the headphones response was averaged across the noise bandwidth.

Determining the SPL from the RMS voltage of the signals can be problematic in some cases, however. The RMS voltage implies that the SPL depends on the signal duration (see Eq. (6.1)). Such a SPL definition is not appropriate when various signal durations are compared, as it is the case in some discussions below. Therefore, when the stimulus duration is implicated in the discussion of experimental results

[2] This file is issued from a calibration procedure which was designed and conducted at the LMA by A. Marchioni and G. Rabau. It provides the headphones frequency response (in dB SPL close to the eardrum) measured in a 10-Hz step with a 1-Volt RMS sinusoidal signal at the input (within the 0–20 kHz range).

presented below, we report SPLs defined as the SPLs of the stationary carriers, *i.e.*, long-lasting sinusoids with the same frequency and amplitude as the Gaussians. The level conversion was done by shifting all thresholds obtained for the Gaussians up by 9.13 dB. Indeed, given that all stimuli were generated with a 10-Volt peak amplitude, the RMS voltage of a 4-kHz, 9.6-ms Gaussian signal defined in Equation (5.1) and sampled at 48 kHz was $V_{S_{RMS}} = 2.47$ Volts. Because we used a $\pi/4$ phase shift, all Gaussian targets (*i.e.*, with variable f_0) had the same energy. They also had the same duration, thus the same RMS voltage. The RMS voltage of a long-lasting sinusoid with the same frequency and amplitude is $10/\sqrt{2} = 7.07$ Volts. Hence, the difference between the SPL of the stationary carrier and the SPL determined from the RMS voltage of the Gaussian signal is $20 \log(7.07/2.47) = 9.13$ dB, independently of the carrier frequency.

Unless otherwise stated, SPLs determined from the RMS voltage are reported below.

Chapter 7

Detection of Gaussian targets in quiet (Experiment 1)

Contents

In the masking experiments described below, the carrier frequency of the masker (F_M) was fixed while the carrier frequency of the target (F_T) was varied as to control the frequency separation (ΔF) between masker and target. To determine the amount of masking at each ΔF, in Experiment 1, absolute thresholds of Gaussian targets were measured for 11 carrier frequencies. Furthermore, to examine how absolute thresholds for very short Gaussian-shaped sinusoids compare to those for long, steady-state ones, absolute thresholds for 300-ms steady-state sinusoids with same frequencies were measured on a subset of listeners. It was expected that the detection of brief Gaussians be poorer than that of long-lasting sinusoids, due to temporal integration.

7.1 Methods

7.1.1 Stimuli

Stimuli were either Gaussian-shaped sinusoids (defined in Eq. (5.1)) with a duration of 9.6 ms ($\Gamma = 600$, $ERD_{GW} = 1.7$ ms) or steady-state sinusoids with a duration of 300 ms, including 10-ms raised-cosine rise/decay times. Eleven target frequencies ($F_T = 2521$, 2833, 3181, 3568, 4000, 4480, 5015, 5611, 6274, 7012, and 7835 Hz) were tested. These frequencies correspond to the different ΔFs involved in Experiment 2 and 6.

In Experiment 1, we compare SPLs for signals of different durations. Therefore, in this chapter, we report SPLs defined as the SPLs of the stationary carriers (see Sec. 6.4).

7.1.2 Listeners

Six normal-hearing listeners participated in the experiments. All had thresholds of 15 dB HL or lower for octave frequencies from 125 Hz to 8 kHz (ANSI S3.6, 1996) and had no indications of hearing disorders. Two of them (L2 and L3) were highly experienced in psychoacoustical tasks.

Four listeners (L1–L4) performed absolute threshold measurements with both Gaussian and steady-state targets. The other two listeners performed measurements with Gaussian targets only. Because of time constraints, listener L6 did not test $F_T = 2833$, 5611 and 7012 Hz.

7.1.3 Thresholds determination

All thresholds were measured using the 3-IFC adaptive procedure described in Section 6.2. Within a session, the target frequencies were chosen at random. For L1–L4 who were tested with both Gaussian and steady-state targets, each session consisted of three threshold estimates for each target type.

7.2 Results and discussion

Figure 7.1 presents individual and mean absolute thresholds as a function of target frequency, obtained with Gaussian (filled squares) and steady-state (empty triangles) targets. First, absolute thresholds obtained with the Gaussian targets at 4 kHz can be compared to those obtained in a collaborative study (Laback et al., 2008) using identical stimuli and procedure but with different listeners and equipment. The mean absolute threshold for the 4-kHz Gaussian target in Laback et al. (see Tab. 1) was 24.7 dB SPL[1], thus nicely comparable to the mean threshold of 23.9 dB SPL in our study.

Second, thresholds varied with frequency, but this variation was independent of target type. Thresholds were consistently higher with Gaussian than with steady-state targets (mean difference = 18.3 dB). Repeated measures analyses of variance (ANOVAs) performed on absolute thresholds by target frequency (11) and target type (Gaussian, steady-state) indicated a significant main effect of target frequency ($F_{10,66} = 25.62$; $p < 0.001$), a significant main effect of target type ($F_{1,66} = 901.33$; $p < 0.001$), and no interaction between target frequency and type ($F_{10,66} = 0.58$, $p = 0.83$). The threshold difference between the two targets is very likely due to the different durations (9.6 *vs.* 300 ms) and reflects temporal integration (see Sec. 2.2.2). In Figure 7.2, this difference (black bar) is compared to mean differences in threshold reported earlier between 2-ms ($\approx ERD_{GW} = 1.7$ ms) and 300-ms sinusoids (gray bars), and between 10-ms (≈ 9.6-ms effective duration of the Gaussians) and 300-ms sinusoids (white bars). The stimuli used in Hempstock et al. (1964) and Florentine et al. (1988) were temporally controlled with Hamming windows as to limit the spectral spread caused by shortening the signals'

[1] The mean absolute threshold deduced from Table 1 in Laback et al. (2008) is 16.7 dB SPL. However, because a headphones calibration error occurred during the data collection, this value is wrong. Thresholds were remeasured after the Acoustics'08 conference and the actual value is 24.7 dB SPL.

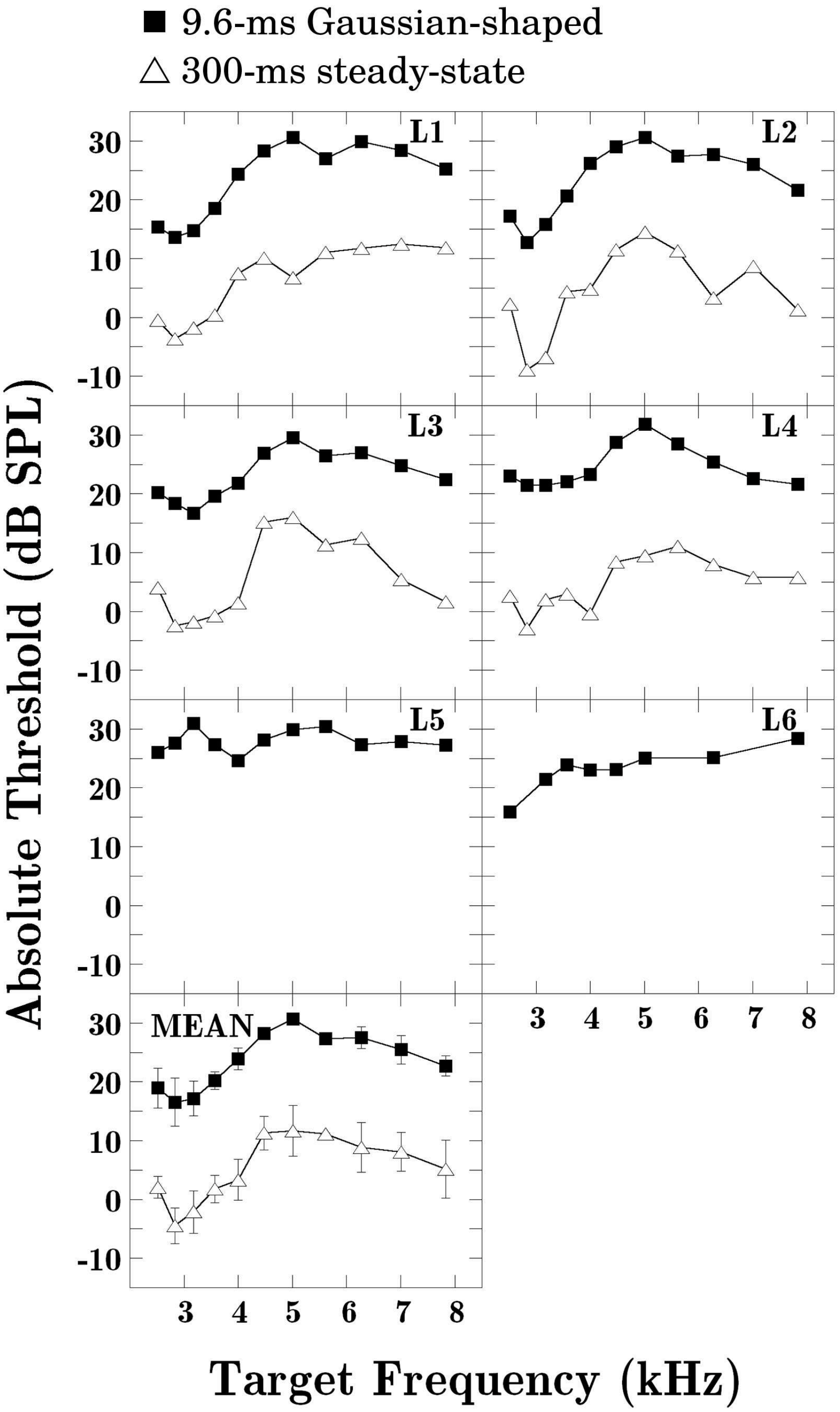


Figure 7.1: Individual absolute thresholds (in dB SPL at the eardrum, defined as the SPL of the stationary carrier) obtained with 9.6-ms Gaussian-shaped (■) and 300-ms steady-state sinusoids (△) are plotted as a function of target frequency (in kHz). The bottom panel shows the mean data for L1–L4 with ±1 standard deviation bars.

duration. Plomp and Bouman (1959) used rectangular windows. It can be seen that the mean difference of about 18 dB in the present study (black bar) resembles the differences of 17–21 dB reported earlier between 2-ms and 300-ms sinusoids (gray bars), but is above the difference of 11 dB reported in studies between 10-ms and 300-ms sinusoids (white bars). This suggest that our data for brief Gaussian stimuli should be compared with literature data for stimuli with comparable ERDs rather than for stimuli with comparable effective durations.

Finally, larger across-listener variability was observed with 300-ms targets than with 9.6-ms ones (see error bars in Fig. 7.1). We propose individual differences in integration time constants as a possible explanation. Both linear and exponential temporal integration models indeed provided individual time constant estimates ranging from 100 to 300 ms (see Sec. 2.2.2).

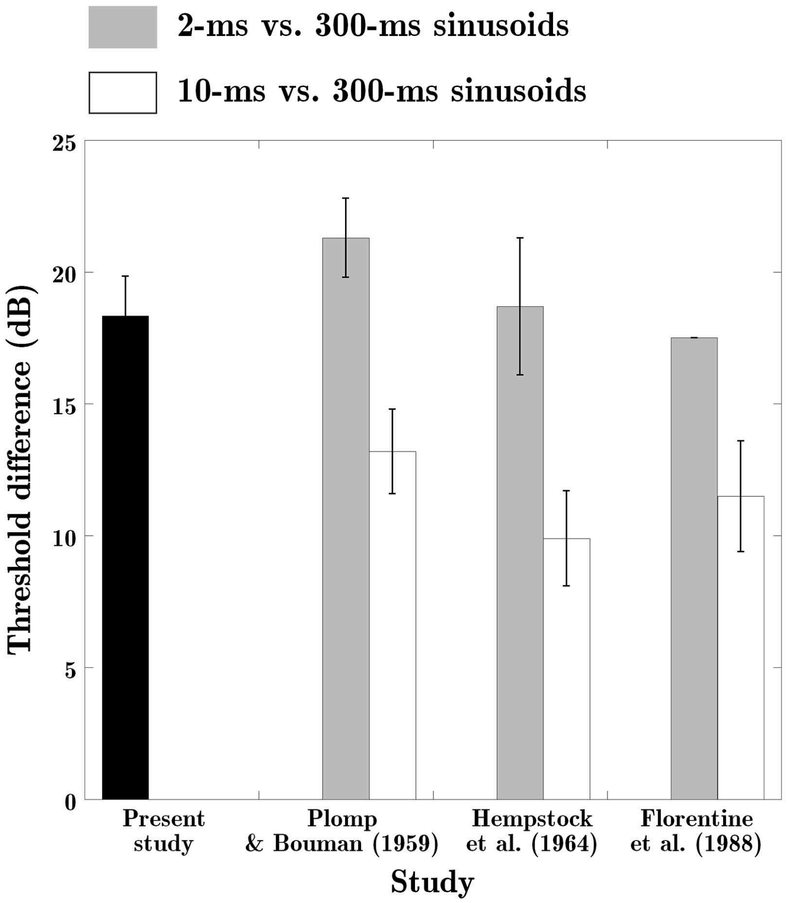


Figure 7.2: Mean threshold differences (in dB) between 2-ms and 300-ms sinusoids (gray bars) and between 10-ms and 300-ms (white bars) in three studies on temporal integration. The black bar on the left shows the mean threshold difference observed in the present study between 9.6-ms Gaussian-shaped and 300-ms steady-state sinusoids. Error bars show ±1 standard deviation.

Summary

In Experiment 1, absolute thresholds were measured for 11 carrier frequencies of the Gaussian target. These frequencies correspond to the different ΔFs involved in Experiment 2 and 6. The obtained thresholds allowed determining the amount of masking at each ΔF in these experiments. Additionally, the thresholds obtained at 4 kHz allowed determining the individual SL of the Gaussian masker used in Experiment 2, 3, 5, and 6.

To examine how absolute thresholds for very short Gaussians compare to those for long, steady-state signals, absolute thresholds for 300-ms steady-state sinusoids with same frequencies were measured on a subset of listeners. We found that the detection of 9.6-ms Gaussian targets is poorer than that of 300-ms sinusoids, independently of frequency. The difference in threshold was similar to that between 2-ms (rectangular- or Hamming-shaped) sinusoids and 300-ms sinusoids reported in the literature. This suggest that our data for brief Gaussian stimuli should be compared with literature data for stimuli with comparable ERDs rather than for stimuli with comparable effective durations.

Chapter 8

Frequency masking with Gaussian stimuli

Contents

This chapter presents Experiment 2–4, in which frequency masking was measured with Gaussians as both masker and target. Signals were presented simultaneously and the frequency separation (ΔF) between masker and target was varied. Because masking patterns provide an estimation of the "effective" spread of frequency masking induced by the masker, the masker had fixed frequency (F_M) and level (L_M), and the amount of masking was measured for various target frequencies (F_T).

In Experiment 2, F_M was fixed to 4 kHz and L_M was fixed to 60 dB SL. Masked thresholds were measured for 11 ΔFs. In Experiment 3, we examined the dependence of the masking patterns' shape on masker level. Masked thresholds were measured for eight ΔFs and three masker levels ($L_M = 30$, 45, and 60 dB SL). In Experiment 4, we examined how masking patterns measured at 4 kHz compare to those measured in the low-frequency portion of the audible spectrum. Masked thresholds were measured for eight ΔFs with $F_M = 0.75$ kHz and $L_M = 60$ dB SL.

Note that the goal of these experiments was not the improved understanding of the mechanisms underlying auditory masking. Rather, these experiments were intended for verifying that the masking patterns for Gaussian stimuli are consistent with those previously reported in the literature for stimuli with comparable frequencies and levels but with various durations and bandwidths (see Sec. 3.2).

8.1 Effect of frequency separation between masker and target (Experiment 2)

8.1.1 Methods

8.1.1.1 Stimuli

Both masker and target were Gaussian-shaped sinusoids (defined in Eq.(5.1)) with a duration of 9.6 ms ($ERB_{GW} = \Gamma = 600$ Hz; $ERD_{GW} = 1.7$ ms). The carrier frequency of the masker was 4 kHz. Its level was 60 dB SL (72–75 dB SPL depending on the listener). Masker and target were presented simultaneously ($\Delta T = 0$). Masking patterns were measured for 11 values of ΔF, defined in the ERB scale (see Eq. (2.11)): 0, ±1, ±2, ±3, ±4, +5, and +6 ERB units (the corresponding F_Ts are listed in Sec. 7.1). This range of ΔF values was chosen to avoid having a too large difference in the number of excited critical bands. With ERB_{F_T} being the ERB of the CB centered at F_T, the ratio ERB_{GW}/ERB_{F_T} was limited to values between 0.5 and 2.0.

To prevent cochlear combination products ("combination tones", CTs, see Sec. 2.1.4) from being detected and thus from producing irregularities in the masking patterns, a continuous background noise was added. In particular, CTs are considered as critical when F_T is above F_M ($\Delta F > 0$). A low-pass filtered (-96 dB/octave) white noise was used, whose cut-off frequency and level were determined so as to mask the most prominent CT: the cubic difference tone (CDT). Regarding the target and the masker as two primary components, when $F_T > F_M$, the CDT frequency (F_{CDT}) is $2F_M$ - F_T. The difference tone as well as higher order CTs (whose frequencies lie below F_{CDT}) were presumably easily masked by the noise (Goldstein, 1967). According to the power spectrum model (see Eq. (2.6) and Patterson and Moore 1986), which assumes that the detection of a sinusoidal component in a narrowband noise depends on a critical target-to-noise power ratio within a CB, the cut-off frequency of the noise was chosen as the upper edge of the ERB centered at F_{CDT}. For example, if F_T was 4480 Hz, F_{CDT} was 3520 Hz. The ERB centered at 3520 Hz has the upper edge at 3722 Hz. Therefore, in this condition, the cut-off frequency of the continuous noise was 3722 Hz. The SPL of the masker was used as a reference to evaluate the most "critical" level (*i.e.*, the highest level) of the CDT. The latter was estimated at about 25 dB below L_M (Goldstein, 1967). Hence, the noise level was adjusted so as to totally mask a Gaussian with a carrier frequency equal to F_{CDT} and a level 25 dB below L_M. For all listeners, this was achieved using an overall noise level of 50 dB SPL.

8.1.1.2 Listeners

The six listeners from Experiment 1 participated in Experiment 2. However, in this and the subsequent experiments, we used an audiometric room different from that used in Experiment 1. Therefore, we first verified that the absolute thresholds for the Gaussian targets measured at 2521, 3181, 3568, 4000, 4480, 5015, 6274, and 7835 Hz were not affected by the room change. Only listener L2 obtained slightly different thresholds in the second room, as illustrated in Figure 8.1. For the other five listeners, the difference in threshold between the two rooms was less than

2 dB at all target frequencies.

Five of the six listeners (L1–L5) performed the conditions $\Delta F = 0$ and $+1$ ERB with *and* without noise, with three threshold estimates per condition.

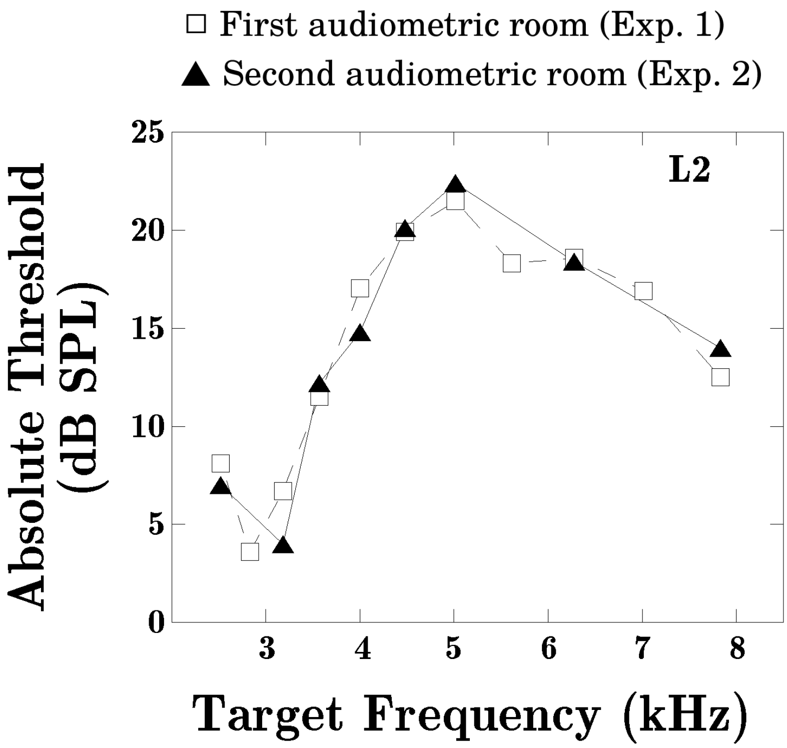


Figure 8.1: Absolute thresholds (dB SPL) as a function of target frequency (kHz) obtained with Gaussian targets in Experiment 1 (□) and after the audiometric room change (▲) for listener L2.

8.1.1.3 Thresholds determination

To verify that the presence of the background noise did not affect the detection of masker or target, the "$\Delta F = 0$" condition was measured with *and* without noise. In this condition, no CT was generated, since masker and target had the same frequency. Hence, the thresholds measured with and without noise should be roughly the same. Furthermore, to assess whether listeners actually used the presence of CTs as a cue to detect the target, the "$\Delta F = +1$ ERB" condition was also measured with *and* without noise. At $\Delta F = +1$ ERB, $F_T = 4480$ Hz $= 1.12\ F_M$. Based on Goldstein's findings (1967), this frequency ratio between the two primaries leads to the highest CT levels. Hence, in this condition, the listeners who have detected CTs should have obtained lower thresholds without noise than when CTs were masked by this noise.

Masked thresholds were measured using the 3-IFC adaptive procedure described in Section 6.2. Each session contained conditions measured with ($\Delta F = 0$, $+1$, $+2$, $+3$, $+4$, $+5$ and $+6$ ERB units) or without ($\Delta F =$ -4, -3, -2, -1, 0, and $+1$ ERB units) background noise. Within a session, the target frequencies were chosen at random.

8.1.2 Results

8.1.2.1 Background noise

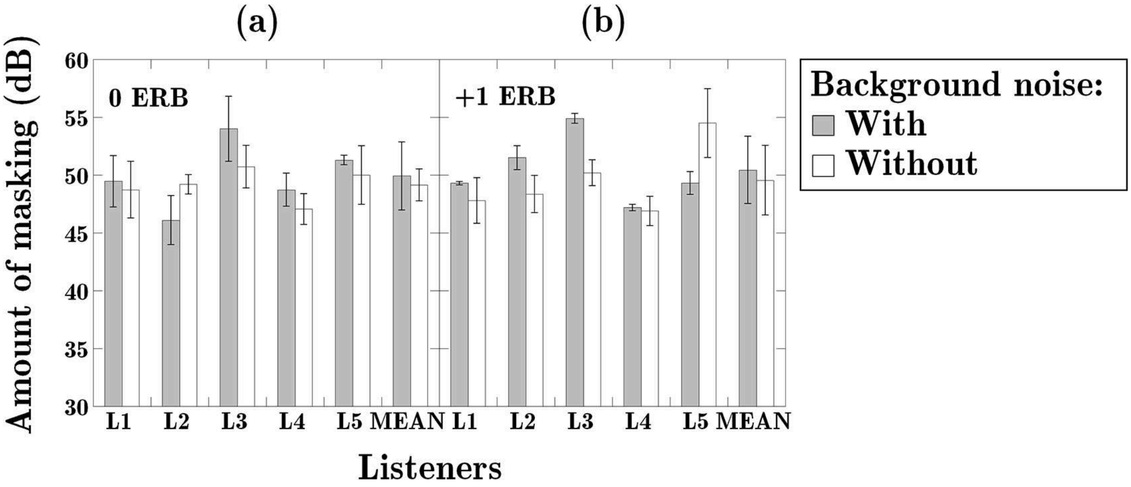


Figure 8.2: Individual and mean amounts of masking (in dB) obtained with (gray bars) and without background noise (white bars) at (**a**) $\Delta F = 0$ ERB and (**b**) $+1$ ERB. Bars show ±1 standard deviation.

Figure 8.2 shows individual and mean amounts of masking (in dB) obtained with (gray bars) and without background noise (white bars) at $\Delta F = 0$ ERB (Fig. 8.2a) and $+1$ ERB (Fig. 8.2b). First, the data obtained at 0 ERB generally showed no effect of background noise (with the exception of listeners L2 and L3 who exhibited differences < 3 dB). Repeated measures ANOVAs performed on the amounts of masking of all five listeners at $\Delta F = 0$ indicated no significant main effect of noise ($F_{(1,4)} = 0.54$; $p = 0.50$). This confirms that the presence of the noise did not affect the detection of masker or target.

Second, at $\Delta F = +1$ ERB, listeners L1 and L4 obtained identical thresholds with and without noise. Listeners L2 and L3 obtained higher thresholds with noise (differences < 5 dB). Surprisingly, listener L5 conversely obtained lower thresholds with noise (difference $= 5$ dB). Repeated measures ANOVAs performed on the amounts of masking of all five listeners at $\Delta F = +1$ indicated no significant main effect of noise ($F_{(1,4)} = 0.27$; $p = 0.63$). Removing listener L5 from the statistical analysis did not change the issue ($F_{(1,3)} = 6.3$; $p = 0.09$). Therefore, two listeners (L2 and L3) may have used CTs as a cue to detect the target when the low-pass noise was absent. Consequently, in the subsequent experiments, the noise was systematically used when F_T was above F_M.

8.1.2.2 Masking patterns

Figure 8.3 presents individual and mean amounts of masking (in dB) as a function of ΔF (in ERB units). First, a dip (listeners L1, L3 and L4) or a plateau (listener L5) was observed instead of a peak at $\Delta F = 0$. This is discussed in Section 8.1.3.1.

Second, for all listeners and $|\Delta F| \geq 2$ ERB units, the amount of masking decreased as $|\Delta F|$ increased. This decrease was more abrupt for $F_T < F_M$ than

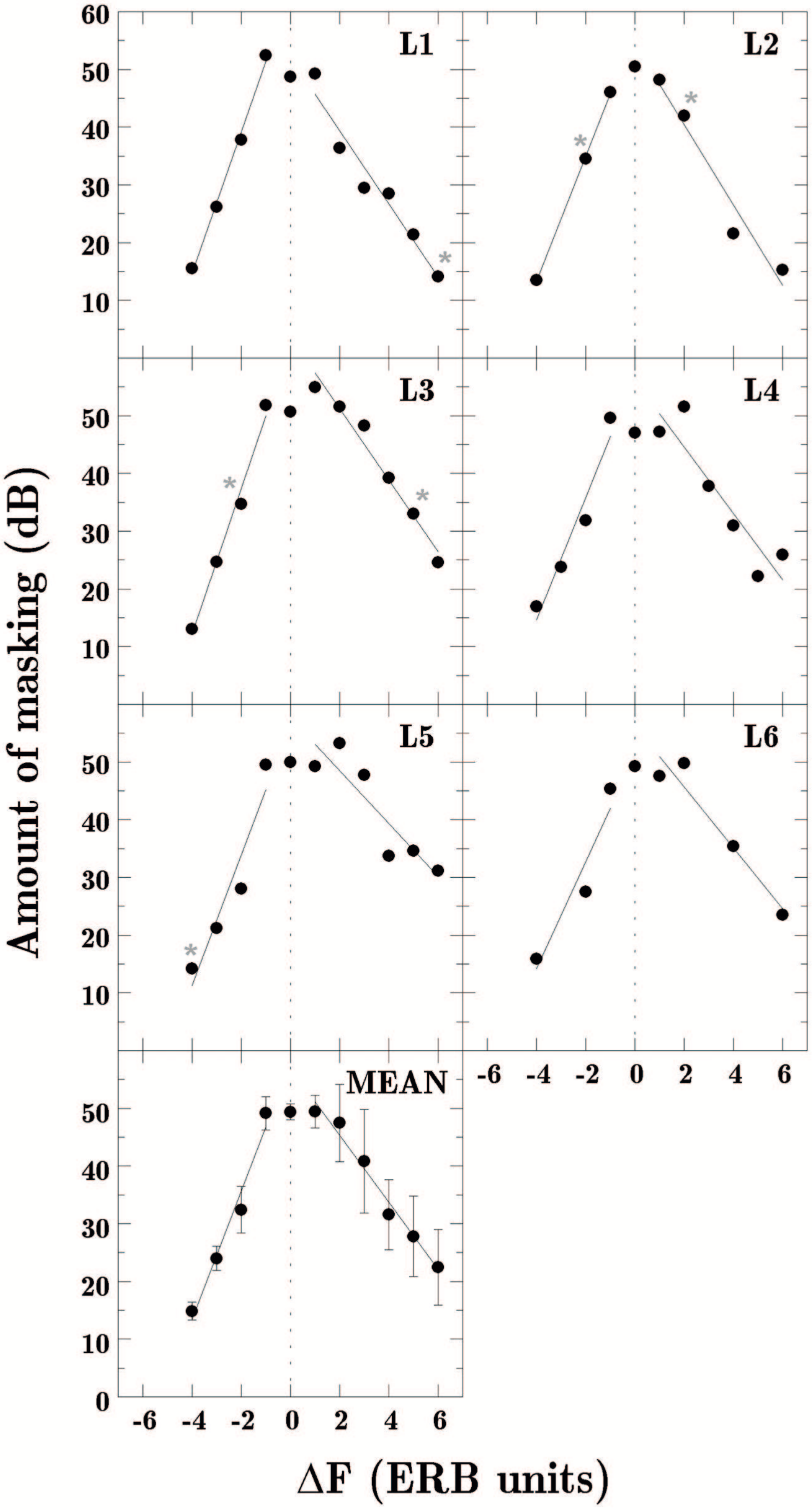


Figure 8.3: Individual amounts of masking (in dB) as a function of ΔF (in ERB units). Data were fitted with linear regression lines on each side of the masking patterns (excluding the point at $\Delta F = 0$). The fit parameters are given in Table 8.1. Stars indicate values with a standard deviation greater than 3 dB. The bottom panel shows the mean data with ±1 standard deviation bars.

Listener	Linear fit	Side of the masking pattern	
		$F_T < F_M$	$F_T > F_M$
L1	Slope	+63.19	−52.95
	r^2	0.99	0.95
L2	Slope	+54.45	−51.78
	r^2	1.00	0.95
L3	Slope	+66.41	−39.66
	r^2	0.99	0.97
L4	Slope	+59.02	−35.28
	r^2	0.94	0.84
L5	Slope	+64.91	−25.49
	r^2	0.91	0.81
L6	Slope	+53.01	−31.40
	r^2	0.91	0.93
MEAN	Slope	+60.12	−38.54
	r^2	0.97	0.98

Table 8.1: Slope (in dB/octave) and r^2 values of the linear regression lines plotted in Figure 8.3.

for $F_T > F_M$. To further examine this asymmetry, regression lines were computed in dB per octave for each side of the masking patterns and listener (straight lines in Fig. 8.3). Because the data obtained at $\Delta F = 0$ represent a special condition of masking (see Sec. 8.1.3.1), they were excluded from the fit. Table 8.1 lists the individual and mean values of slope and r^2. Except for listener L2 who showed no asymmetry, the slopes for $F_T < F_M$ were, on average, 1.6 times those for $F_T > F_M$. The masking patterns' asymmetry is best illustrated in Figure 8.4, in which the amount of masking averaged across four listeners (L1, L3, L4, and L5 who were tested for $\Delta F = \pm1$, ±2, ±3, and ±4 ERB units) is plotted as a function of ΔF for each side of the masker frequency. The figure clearly shows that the data for $F_T < F_M$ (black triangles) and $F_T > F_M$ (gray circles) do not fall on parallel lines. A straight-line fit of these data provided slopes of +63.38 and -38.35 dB/octave for $F_T < F_M$ and $F_T > F_M$, respectively ($r^2 \geq 0.97$). Overall, this steeper masking decay for $F_T < F_M$ is consistent with that reported in previous frequency masking studies (see Sec. 3.2).

The ΔF values chosen for the experiment (guided by pretests) did not allow the listeners to reach 0 dB of masking. Indeed, if the low-frequency and high-frequency slopes in Figure 8.3 are extrapolated, 0 dB of masking would have been reached, on average, at ΔFs = -5 and +10 ERB units, respectively. Testing such large ΔF values would have caused the ratio ERB_{GW}/ERB_{F_T} to exceed the 0.5–2.0 constraint stated above.

Finally, it should be noted that across-listener variability was larger with $F_T > F_M$ than with $F_T < F_M$ (see error bars in Fig. 8.3).

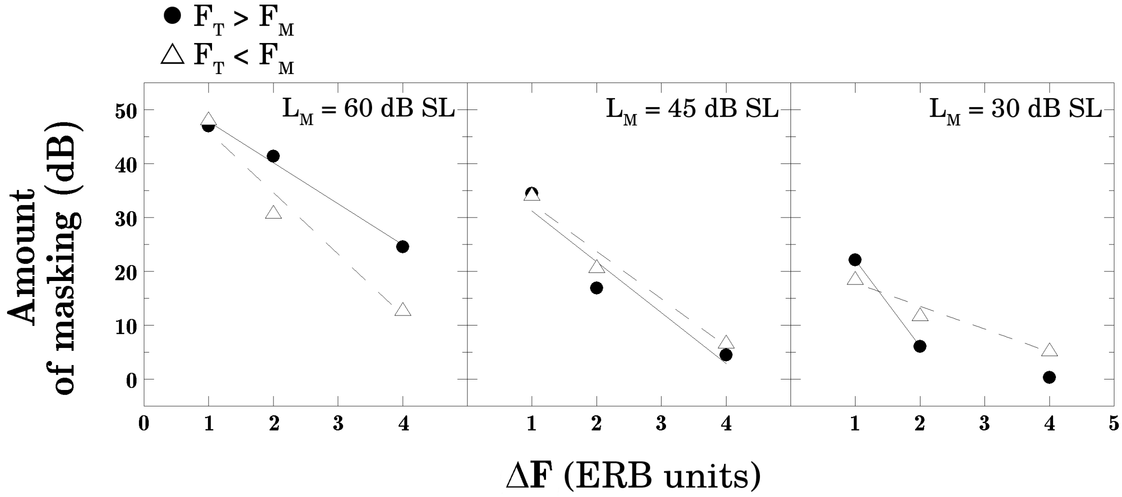


Figure 8.4: The amount of masking averaged across four listeners (L1, L3, L4, and L5 who were tested for $\Delta F = \pm 1$, ± 2, ± 3, and ± 4 ERB units) is plotted as a function of ΔF (in ERB units) for $F_T < F_M$ (▲) and $F_T > F_M$ (●). Data were fitted with linear regression lines whose parameters are specified in the text.

8.1.3 Discussion

8.1.3.1 The "$\Delta F = 0$" condition

In studies on simultaneous masking, greatest masking effects occur when ΔF equals 0, leading to a peak in the masking patterns at this frequency separation (see Sec. 3.2). In the present study, five of the six listeners conversely exhibited a dip or a plateau at $\Delta F = 0$. It has to be considered that this represents a special condition, where masker and target were exactly the same stimuli presented at the same time. Thus, the listeners could only use the intensity increase in the interval containing the target as a cue. In other words, the listener performed an intensity discrimination task in this condition.

Green (1969), who used identical 10-ms, rectangular-shaped sinusoids as masker and target, also reported a dip at $\Delta F = 0$. He showed that the target level at threshold was identical to that predicted by the just-noticeable difference (jnd) in intensity between the interval containing the masker alone and that containing both masker and target. Let I_M and I_T denote the masker and target intensities, respectively, and I_{M+T} the intensity of "masker plus target"[1]. The relative difference ("$\Delta I/I$") between the two intensities is $(I_{M+T} - I_M)/I_M$. Table 8.2 lists individual and mean $\Delta I/I$ values at target threshold in Experiment 2 for $\Delta F = 0$ (expressed as $10\log(\Delta I/I)$, in dB). Note that because the stimulus duration is implicated in the present discussion, the signal levels used to compute the $\Delta I/I$ values were taken in dB SPL referenced at the stationary carriers (see Sec. 6.4). Hence, the masker, or "pedestal[2]" levels, were 82–85 dB SPL depending on the listener. The $\Delta I/I$ values range from $+1$ to -3 dB (mean $=$ -1.8 dB, standard deviation $=$

[1] Note that $I_{M+T} \neq I_M + I_T$. Rather, I_{M+T} depends on the exact phase relation between masker and target. For a review, see Grantham and Yost (1982).

[2] Given that the pedestal is the baseline intensity (I) from which an intensity increment (ΔI) is to be detected, it is equivalent to the masker intensity (I_M) in the present study.

0.7 dB). The value reported by Green (1969) was -6.8 dB, which is much lower. Two studies can be considered that measured intensity discrimination thresholds for short Hamming-shaped sinusoids. The first is that of Florentine (1986), who used 2-ms ($\approx ERD_{GW}$) sinusoids at 1 kHz (pedestal level = 85 dB SPL). She obtained an average jnd in intensity of -2.33 dB, compatible with the data reported in Table 8.2. The second is that of Carlyon and Moore (1984), who measured an average $\Delta I/I$ of 0 dB with 6-ms sinusoids at 4 kHz (steady-state portion = 1 ms; pedestal level = 55 dB SPL), which is higher than our mean value. This can be reasonably explained by the lower pedestal level used by Carlyon and Moore. The cited authors indeed showed that intensity discrimination improves (*i.e.*, $\Delta I/I$ decreases) as the pedestal level increases. The main difference between the four studies is the temporal windowing of the stimuli. Green used a rectangular window producing a large amount of spectral spread. In the other three studies, this spread was minimized by the use of Hamming (Carlyon and Moore, 1984; Florentine, 1986) or Gaussian (present study) windows. More off-frequency band information was therefore available to the listeners in Green (1969), which could explain his lower $\Delta I/I$. Accordingly, Carlyon and Moore (1984) showed that the use of a band-stop noise centered at the target frequency to prevent off-frequency listening caused an increase in the discrimination threshold (*i.e.*, a higher $\Delta I/I$).

In summary, the ΔT and $\Delta F = 0$ condition effectively represented an intensity discrimination task. The presence of a peak in the masking pattern for one of the six listeners could be due to the fact that this listener has the highest intensity jnd. Finally, the amount of intensity increase in the target interval depends on the exact phase relationship between masker and target. If we had used, for example, a 90-degree phase shift (as done by Moore et al., 1998), we would have therefore obtained higher thresholds, and the dip in the masking patterns would probably not have been observed.

Listener	$10 \log \left[\frac{(I_{M+T} - I_M)}{I_M} \right]$ (dB)
L1	-2.07
L2	-1.07
L3	-1.16
L4	-3.00
L5	-1.35
L6	-1.75
MEAN	-1.82

Table 8.2: Individual and mean $\Delta I/I$ values, expressed as $10 \log[(I_{M+T} - I_M)/I_M]$ (in dB), at target threshold in Experiment 2 for $\Delta F = 0$. The signal levels at threshold were taken in dB SPL referenced at the stationary carriers.

8.1.3.2 Masking patterns

In Figure 3.4, masking patterns for maskers with comparable frequencies and levels but with various bandwidths were compared. This comparison revealed that the spectral width of masking patterns (qualified by the quality factor at the -3-dB bandwidth, see Sec. 3.2.3) and, in other words, the spread of frequency masking, increases with the broadening of masker bandwidth. In the present study, we used Gaussians as masker and target to reduce spectral and temporal spreads in the signal-induced excitation. Therefore, to examine how masking patterns for Gaussian maskers compare to those for maskers with various spectral and temporal characteristics, the mean results from Experiment 2 (**d**, bold line) are plotted together in Figure 8.5 with some literature data. Precisely, we considered (**a**) a continuous noise that excites several critical bands (Bilger and Hirsh, 1956), (**b**) a 10-ms sinusoid temporally shaped with a rectangular window inducing considerable spectral broadening (Green, 1969), (**c**) a 50-ms sinusoid temporally shaped with a Hamming window restricting this spectral broadening (Bacon and Viemeister, 1985), and (**e**) a continuous sinusoid whose spectral energy is optimally concentrated at the masker frequency (Ehmer, 1959b). All studies used maskers with comparable levels (70–80 dB SPL) and frequencies (4 kHz, except Bilger and Hirsh 1956 who used a broadband noise centered at 1210 Hz). The corresponding spreads of the stimuli across the TF plane as well as the values of Q_{3dB} estimated for each of the corresponding masking patterns are specified in the insert.

It can be seen that broadband maskers (open squares) and rectangular-shaped sinusoids (open circles) produce the broadest masking patterns (Q_{3dB} = 7–8). Narrowband maskers (filled symbols) conversely result in narrower patterns (Q_{3dB} = 11–15). Figure 8.5 further demonstrates that the amount of spectral spread caused by shortening the duration of the signal can be limited by the use of appropriate window shapes (Bacon and Viemeister, 1985; Hartmann and Wolf, 2009). Indeed, the 10-ms rectangular-shaped sinusoid produced a pattern as broad as that for a continuous broadband noise, while the 9.6-ms Gaussian-shaped sinusoid produced a much narrower pattern, though slightly broader than that for the 50-ms Hamming-shaped sinusoid (Bacon and Viemeister, 1985). Given that the cited authors used stimuli with similar spectra but different durations (50-ms masker *vs.* 20-ms target), the listeners could have better distinguished the target from the masker, resulting in lower amounts of masking (Weber and Moore, 1981; Neff, 1985).

The difference in frequency spread of masking between the 50-ms Hamming-shaped and the 9.6-ms Gaussian-shaped sinusoids can be due to the temporal development of auditory contours (Elliott, 1967). The experiments carried out by Elliott (1967) showed that stimulation by a narrowband stimulus (*i.e.*, whose bandwidth is narrower than or equal to the width of the CB centered at the stimulus frequency) produces a broadband excitation pattern at stimulation onset, which narrows *and* decreases in amplitude over time. Stimulation by a broadband stimulus also produces a broadband excitation that decreases in amplitude over time, but whose spread is unchanged. Elliott therefore suggested that short (duration $\leq$ 20 ms), narrowband maskers should produce more masking than long (duration $>$ 200 ms) and narrowband maskers for short targets presented at the masker onset. Accordingly, Bacon et al. (2002) measured psychophysical tuning curves (PTCs, see Sec. 2.3.2) in various temporal conditions (*i.e.*, with a short target presented at or after the onset of a long masker) and showed that PTCs are broader when measured at masker onset than when measured with a time delay between masker and target onsets. Though often debated, the Elliott's hypothesis was addressed in various studies on frequency selectivity, masking, or overshoot

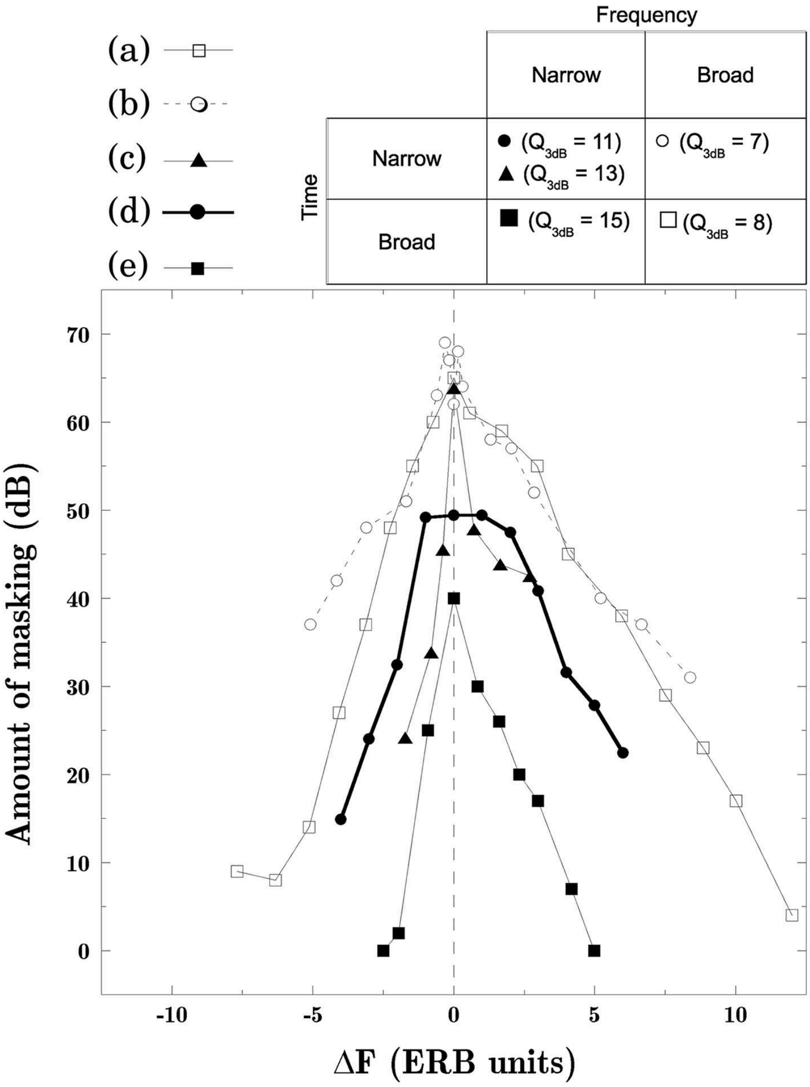


Figure 8.5: Amount of masking (in dB) as a function of ΔF (in ERB units). The mean results for the present 9.6-ms Gaussian-shaped masker (**d**, bold line) are compared with (**a**) a continuous, 420-Hz-wide band of noise centered at 1210 Hz (Bilger and Hirsh, 1956), (**b**) a 10-ms rectangular-shaped sinusoid (Green, 1969), (**c**) a 50-ms (incl. 10-ms Hamming rise/fall times) sinusoid (Bacon and Viemeister, 1985), and (**e**) a continuous sinusoid (Ehmer, 1959b). The corresponding spreads of the stimuli across the TF plane as well as the values of Q_{3dB} estimated for each masking pattern are specified in the insert.

(Green, 1969; Zwicker and Fastl, 1972; Bacon and Viemeister, 1985; Strickland, 2001; Bacon et al., 2002). Overall, because (1) the Gaussian-shaped masker in the present study was much shorter than the Hamming-shaped masker and (2) the short target was temporally centered within the longer masker in Bacon and Viemeister (1985), the Gaussian masker may not have had sufficient time to develop its narrow frequency contours, thus producing a broader masking pattern.

Otherwise, all types of maskers in Figure 8.5 produce the greatest masking at $\Delta F = 0$ (except in the present study and in Green, 1969, see Sec. 8.1.3.1 above), and asymmetrical patterns with steeper slopes for F_T below F_M than for F_T above.

This asymmetry has been explained by suppressive masking, which dominates for F_Ts above F_M for masker levels greater than 50 dB SPL, and by the increase of the auditory filter bandwidth with increasing center frequency (see Sec. 3.2.6). One listener from Experiment 2 obtained a symmetrical pattern. He might have needed a higher masker level to produce the expected asymmetry (Egan and Hake, 1950; Ehmer, 1959b; Vogten, 1978b).

Larger across-listener variability was observed in Experiment 2 for $F_T > F_M$ than for $F_T < F_M$. A similar difference was previously reported by Egan and Hake (1950) with a narrowband noise masker. This could be due to individual differences in the degree of suppression. Accordingly, Rodriguez et al. (2010) made psychophysical estimates of suppression in 22 normal-hearing listeners for a target signal ("suppressee") to $F_T = 4000$ Hz and two masker ("suppressor") frequencies ($F_M = 2141$ or 4281 Hz). Their data revealed a greater inter-listener variability for $F_T > F_M$ than for $F_T < F_M$, consistent with our data. Individual differences in the perception of CTs were also proposed as a possible explanation (Moore et al., 1998), but this explanation is unlikely here, because the detection of CTs was avoided.

Summary

In Experiment 2, masked thresholds were measured as a function of ΔF for simultaneous presentation ($\Delta T = 0$) of the Gaussian masker and target. The masker had a fixed frequency ($F_M = 4$ kHz) and level ($L_M = 60$ dB SL). This allowed us to examine the shape of frequency masking patterns for a Gaussian masker. The patterns showed a dip when the target frequency equaled the masker frequency. Because the task at this frequency likely reflects intensity discrimination, the presence and size of the dip will depend on the phase relationship between masker and target. The amount of masking was found to decrease more abruptly when ΔF increased for target frequencies below the masker frequency than for target frequencies above. This asymmetry is similar to that previously reported in studies using maskers with various spectro-temporal characteristics. The width of the masking pattern for the 9.6-ms Gaussian-shaped masker was (1) much narrower than that for broadband noise maskers and 10-ms rectangular-shaped maskers producing considerable spectral spread, (2) broader than that for continuous sinusoidal maskers, and (3) comparable to that for sinusoidal maskers temporally controlled with windows limiting spectral broadening.

8.2 Effect of masker level (Experiment 3)

8.2.1 Methods

8.2.1.1 Stimuli

In this experiment, masking patterns were measured for eight ΔFs (0, ±1, ±2, ±4 and +6 ERB units) and three levels of the Gaussian masker, namely

L_M = 30, 45, and 60 dB SL[3] (the corresponding SPLs spanned 40–45, 55–60 and 70–75 dB depending on the listener). The masker frequency (F_M) was fixed to 4 kHz.

When F_T was above F_M ($\Delta F > 0$), we used a background noise identical to that of Experiment 2 to mask potential CTs. The background noise level was adjusted to each of the tested masker levels using the same procedure as above, *i.e.*, to totally mask a Gaussian with a carrier frequency equal to F_{CDT} and a level 25 dB below L_M (Goldstein, 1967). For all listeners, the overall noise levels were 20, 35, and 50 dB SPL for L_M = 30, 45, and 60 dB SL, respectively.

8.2.1.2 Listeners

Four of the six listeners (L1–L4) participated in Experiment 3. Because Experiment 3 was conducted about 8 months after Experiment 1 and 2, we first verified that the absolute thresholds for the Gaussian targets measured at 2521, 3181, 3568, 4000, 4480, 5015, 6274, and 7835 Hz were the same. This was the case for listeners L2–L4 (differences < 3 dB for all target frequencies). Only listener L1 obtained different thresholds, as illustrated in Figure 8.6.

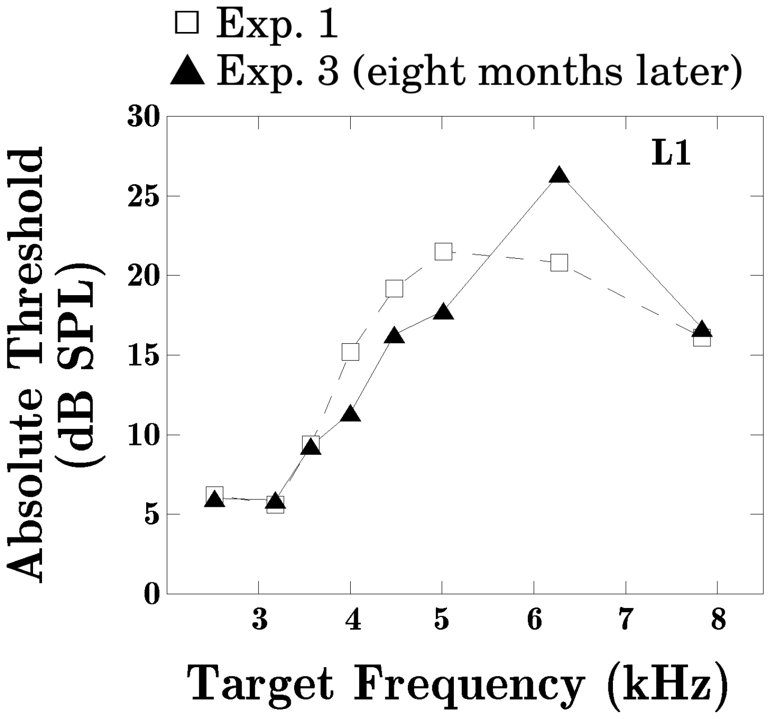


Figure 8.6: Absolute thresholds (dB SPL) as a function of target frequency (kHz) obtained with Gaussian targets in Experiment 1 (□) and in Experiment 3 performed eight months later (▲) for listener L1.

[3] The 60-dB SL condition was measured in Experiment 2. However, because Experiment 3 was conducted about eight months after Experiment 2, the 60-dB SL condition was remeasured.

8.2.1.3 Thresholds determination

Each session contained conditions measured with (ΔF = +1, +2, +4, and +6 ERBs) and without background noise (ΔF = -4, -2, -1, and 0 ERBs). Within a session, the target frequencies and masker levels were chosen at random.

8.2.2 Results

The condition with $L_M = 60$ dB SL in the present experiment (Exp. 3) replicated the conditions from Experiment 2 on four listeners. In Figure 8.7, the data obtained with $L_M = 60$ dB SL in Experiment 3 (filled circles) are compared to those obtained in Experiment 2 (open triangles). Listener L1 showed less masking (-14 dB) in Experiment 3 than in Experiment 2 at almost all ΔFs > 0. Remind that this listener obtained lower absolute thresholds in Experiment 3 than in Experiment 1 (conducted eight months prior to Exp. 3, see Fig. 8.6). In particular, the threshold in quiet for the 4-kHz Gaussian target changed from 15 (Exp. 1) to 11 dB SPL (Exp. 3). This resulted in a lower SPL for the 60-dB-SL masker in Experiment 3, which may have resulted in the lower amounts of masking in Experiment 3. Listeners L2–L4 obtained comparable thresholds in both experiments (differences < 3 dB).

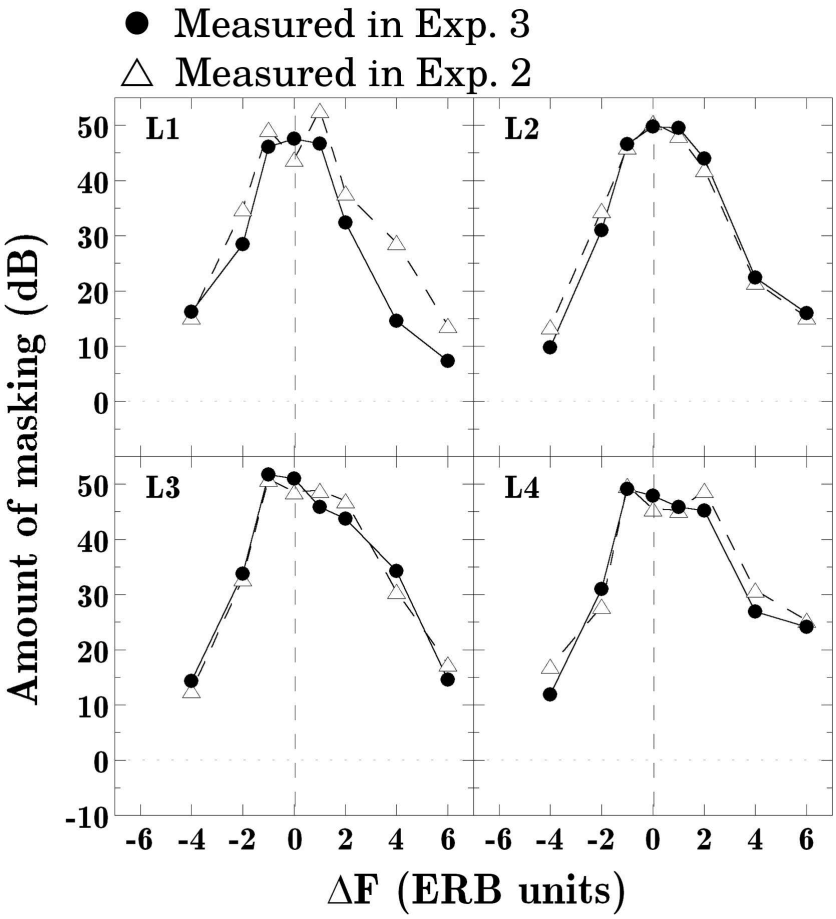


Figure 8.7: Individual amounts of masking (in dB) as a function of ΔF (in ERB units) obtained at $L_M = 60$ dB SL in Experiment 3 (●) and Experiment 2 (△) for listeners L1–L4.

Figure 8.8 presents individual and mean amounts of masking (in dB) as a function of ΔF (in ERB units), for each L_M obtained in Experiment 3. For all listeners, the dip/plateau observed at $\Delta F = 0$ for $L_M = 60$ dB SL (see also Fig. 8.3) disappeared at the lower L_Ms and gave rise to a peak. Otherwise, the results from all listeners reveal a clear decrease in the amount of masking as L_M decreased from 60 to 30 dB SL. For all frequency separations, the largest amount of masking was obtained with $L_M = 60$ dB. Masking dropped by 10–25 dB as L_M decreased to 45 dB, then by 10–15 dB as L_M decreased to 30 dB. For $|\Delta F| > 2$ ERB units, masking was generally less than 5 dB at $L_M = 30$ dB SL.

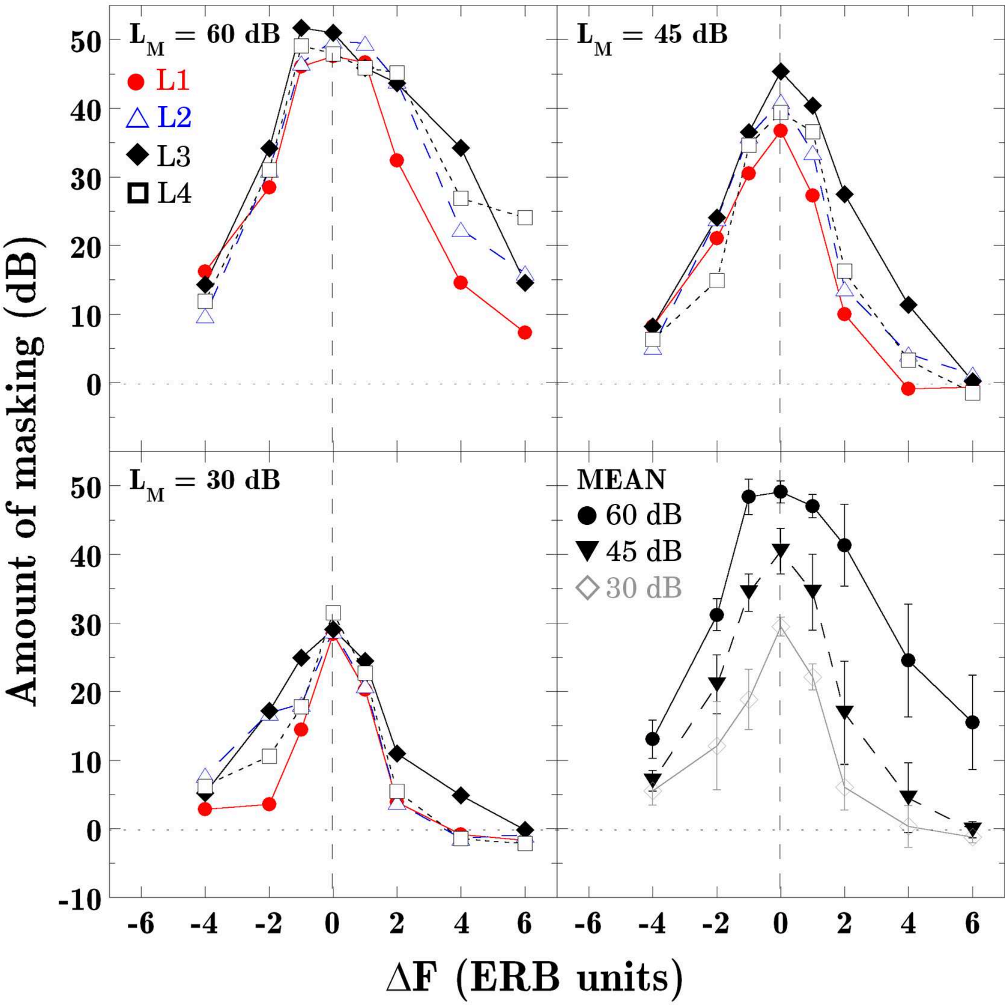


Figure 8.8: Individual amounts of masking (in dB) as a function of ΔF (in ERB units) obtained at $L_M = 60$, 45, and 30 dB SL. Each panel is for a masker level. The bottom right panel shows the mean data with ±1 standard deviation bars with L_M as the parameter.

To investigate the effect of L_M on the steepness of masking decay on each side of the masking patterns, Figure 8.9 presents the mean amount of masking (in dB) obtained at each L_M as a function of ΔF (in ERB units), for each side of the masking pattern. In each panel, the data were fitted with linear regression lines whose parameters are listed in Table 8.3. For $L_M = 60$ dB, the slope of the masking decay for $F_T < F_M$ (dashed line) is about twice as steep as for $F_T > F_M$ (straight line), which reflects the upward spread of masking. For $L_M = 45$ dB, comparable slopes are obtained on both sides of the masking patterns (parallel lines in Fig. 8.9). For $L_M = 30$ dB, the slope for $F_T > F_M$ is about four times that for $F_T < F_M$, *i.e.*, the asymmetry observed in masking patterns for $L_M = 60$ dB is reversed. This steeper masking decay for $F_T > F_M$ than for $F_T < F_M$ is consistent with previous frequency masking studies (see Sec. 3.2.5), and is sometimes referred to as the “downward spread of masking". Overall, these data reveal that a decrease in L_M results in a shallower slope for $F_T < F_M$ and a steeper slope for $F_T > F_M$.

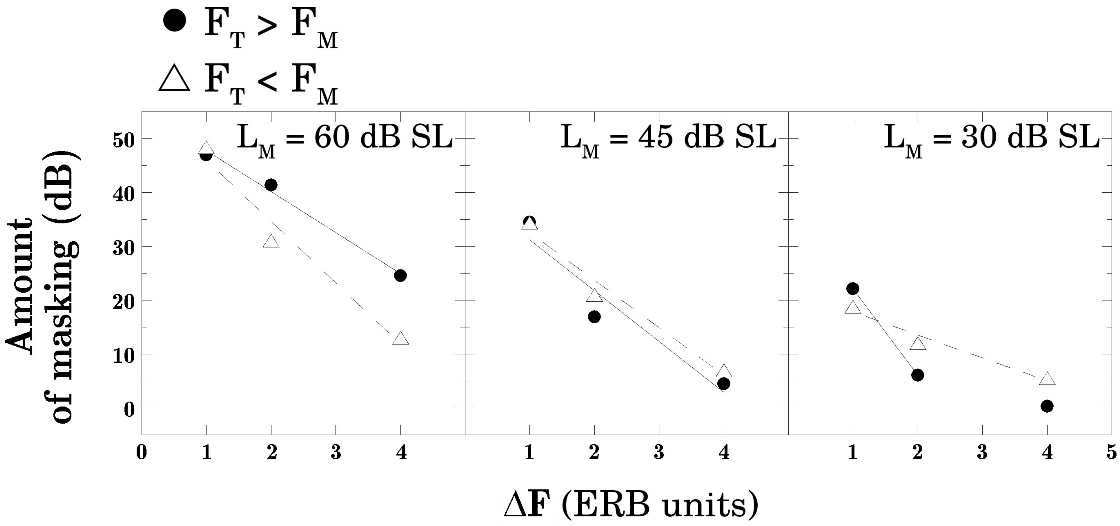


Figure 8.9: The mean amount of masking (in dB) obtained at each L_M is plotted as a function of ΔF (in ERB units) for $F_T < F_M$ (△, dashed lines) and for $F_T > F_M$ (•, straight lines). For each side of the masking patterns, data were fitted with linear regression lines whose parameters are listed in Table 8.3. Because of floor effects, the data point with $\Delta F = +4$ ERBs for $L_M = 30$ dB SL was excluded from the fit.

Masker level	Linear fit	Side of the masking pattern	
(dB SL)		$F_T < F_M$	$F_T > F_M$
60	Slope	−11.38	−7.62
	r^2	0.98	0.99
45	Slope	−8.84	−9.45
	r^2	0.98	0.96
30	Slope	−4.27	−16.02
	r^2	0.98	1.00

Table 8.3: Slope (in dB/ERB) and r^2 values of the linear regression lines plotted in Figure 8.9.

To asses whether the masking patterns become broader of narrower with decreasing L_M, we computed the quality factors at the 3-dB bandwidth (conditions

$\Delta F = 0$ were excluded from the fits, see Sec. 3.2.3), Q_{3dB}, for all listeners and L_Ms. Individual and mean values of Q_{3dB} are listed in Table 8.4. Except for listener L1 who showed a constant Q_{3dB} in all conditions, Q_{3dB} remained approximately the same as L_M decreased from 60 to 45 dB, but roughly decreased (*i.e.*, patterns broadened) as L_M further decreased to 30 dB.

	L_M (dB SL)		
Listener	60	45	30
L1	12.4	11.7	11.6
L2	13.0	14.0	7.0
L3	11.9	12.4	9.1
L4	10.1	14.0	7.3
MEAN	12.0	13.4	7.7

Table 8.4: Individual and mean values of Q_{3dB} for three L_Ms estimated from the data in Figure 8.8.

Finally, it should be noted that for $L_M = 60$ and 45 dB SL, across-listener variability was larger with $F_T > F_M$ than with $F_T < F_M$ (see error bars in Fig. 8.8). Conversely, for $L_M = 30$ dB SL, larger variability was obtained for $F_T < F_M$ than for $F_T > F_M$.

8.2.3 Discussion

For $L_M = 60$ dB SL, five of the six listeners from Experiment 2 and three of the four listeners from Experiment 3 exhibited a dip/plateau instead of a peak at $\Delta F = 0$ (see Fig. 8.3). The presence of this dip was attributed to the fact that listeners performed an intensity discrimination task in this special condition, where masker and target were strictly identical stimuli presented at the same time. In Experiment 3, for $L_M = 45$ and 30 dB SL, all listeners showed a peak at $\Delta F = 0$. This presence of peaks for the lower values of L_M is likely due to a deterioration in intensity discrimination with decreasing pedestal (*i.e.*, L_M) level. Table 8.5 lists the individual and mean $\Delta I/I$ values (expressed as $10\log[(I_{M+T} - I_M)/I_M]$, in dB) at target threshold in Experiment 3 for $\Delta F = 0$ and $L_M = 30$, 45, and 60 dB SL. Because the stimulus duration is implicated in the present discussion, the signal levels used to compute the $\Delta I/I$ values were taken in dB SPL referenced at the stationary carriers (see Sec. 6.4). For all listeners, $\Delta I/I$ decreases (*i.e.*, intensity discrimination performances improve) as L_M increases. Accordingly, previous studies with short (duration < 30 ms) Hamming-shaped (Carlyon and Moore, 1984; Florentine, 1986) and Gaussian-shaped (Nizami et al., 2001) sinusoids showed that intensity discrimination improves as pedestal level increases (pedestal levels spanned 25–85 dB SPL across the cited studies).

The reversal of the masking patterns' asymmetry observed when L_M decreased from 60 to 30 dB SL (see Figs. 8.8 and 8.9) is closely consistent with almost all previous studies on frequency masking which investigated the effect of masker

Listener	L_M dB SL	L_M dB SPL	$10\log\left[\frac{(I_{M+T}-I_M)}{I_M}\right]$ dB
L1	30	50	4.00
	45	65	−0.14
	60	80	−2.48
L2	30	54	4.00
	45	69	2.20
	60	84	−1.07
L3	30	49	4.31
	45	64	5.04
	60	79	−0.61
L4	30	52	6.06
	45	67	1.46
	60	82	−2.31
MEAN	30	51	4.58
	45	66	2.14
	60	81	−1.61

Table 8.5: Individual and mean $\Delta I/I$ values, expressed as $10\log[(I_{M+T}-I_M)/I_M]$ (in dB), at target threshold in Experiment 3 for $\Delta F = 0$ and $L_M = 30$, 45, and 60 dB SL. The signal levels at threshold were taken in dB SPL referenced at the stationary carriers.

level (Egan and Hake, 1950; Bilger and Hirsh, 1956; Ehmer, 1959b; Vogten, 1978b,a; Zwicker and Jaroszewski, 1982; Lutfi and Patterson, 1984; Moore et al., 1998). The cited studies indeed reported "upward" asymmetrical patterns at high levels ($L_M >$ 50 dB SPL), symmetrical patterns at moderate levels (45–50 dB SPL), and "downward" asymmetrical patterns at low levels ($<$ 40 dB SPL). The high-level asymmetry (or "upward spread of masking") is commonly attributed to the increase in the auditory filters bandwidth with increasing center frequency, and to the contribution of suppression effects (see Sec. 3.2.6). The low-level asymmetry can be attributed to the level-dependent changes in the auditory filters' shape (Lutfi and Patterson, 1984), and to the contribution of suppression effects (Vogten, 1978b,a). Although suppression effects were found to be dominant for F_Ts above F_M and for levels greater than 50 dB SPL (Delgutte, 1990; Yasin and Plack, 2005; Rodriguez et al., 2010), at low levels, the response of auditory nerve fibers was found to be suppressive only for suppressor frequencies above the suppressee frequency (see Fig. 2.10 and Vogten, 1978a, Fig. 6(b)), *i.e.*, for F_Ts below F_M.

Furthermore, the flattening of masking patterns observed at the lowest masker level (see Tab. 8.4) is compatible with previous studies. For comparison, we computed the values of Q_{3dB} for the masking patterns in Figures 3.2 (data from Egan and Hake, 1950, L_M = 20–80 dB SPL) and 3.6 (data from Zwicker and Jaroszewski, 1982, L_M = 20–60 dB SPL). In Egan and Hake (1950), the spectral width of the masking patterns remained approximately constant (Q_{3dB}

= 6–9) for L_M = 40–80 dB, then drastically broadened (Q_{3dB} = 1–4) for levels below 30 dB. Similar observations were made from Zwicker and Jaroszewski (1982): Q_{3dB} was about 8–11 for levels greater than or equal to 30 dB, then decreased to about 3 for L_M = 20 dB. This broadening of the patterns at low levels can result from the shallow masking decay for $F_T < F_M$ (see Tab. 8.3) combined with the fact that the amount of masking for $\Delta F = 0$ decreases as L_M decreases, *i.e.*, the masking decay "starts" from a lower level.

Finally, for L_M = 60 and 45 dB SL, larger across-listener variability was observed with $F_T > F_M$ than with $F_T < F_M$. As for Experiment 2, this is probably due to individual differences in the degree of suppression (see Sec. 8.1.3.2). Also, because suppression effects are possibly involved in the downward asymmetry with L_M = 30 dB SL, individual differences in the degree of suppression probably accounted for the large variability observed for $F_T < F_M$ and L_M = 30 dB SL.

Summary

In Experiment 3, masking patterns were measured for eight ΔFs and three masker levels (L_M = 60, 45 and 30 dB SL) in four listeners. The masker had a fixed frequency (F_M = 4 kHz).
Except for one listener, the patterns for L_M = 60 dB were comparable to those obtained in Experiment 2. For lower values of L_M, the patterns showed a peak at $\Delta F = 0$. This is probably due to the degradation of intensity discrimination with decreasing level. For target frequencies above the masker frequency, the decrease in L_M resulted in a steeper masking decay. Conversely, for target frequencies below the masker frequency, the decrease in L_M resulted in a shallower masking decay. Consequently, patterns were symmetric for L_M = 45 dB, and asymmetric for L_M = 30 dB with steeper slopes with $F_T > F_M$ than with $F_T < F_M$. Overall, the "upward spread of masking" observed at high levels was found to reverse ("downward spread of masking") at low levels. This is in close agreement with previous data for various types of maskers. The results can be attributed to (1) the increase of the auditory filters bandwidth with increasing center frequency, especially at high levels, (2) the level-dependent changes in the auditory filters' shape, and (3) the contribution of suppression effects.

8.3 Effect of masker frequency (Experiment 4)

To assess whether masking patterns for Gaussian stimuli measured in the low-frequency portion of the audible spectrum (see Fig. 3.1) have different shapes compared to those measured at 4 kHz, in Experiment 4, the carrier frequency of the Gaussian masker was fixed to 0.75 kHz. To minimize the number of activated TF observation windows of the auditory system (*i.e.*, to keep the ratio ERB_{GW}/ERB_{F_M} as closest as possible to one), the stimulus parameters had to be modified based on the results from van Schijndel et al. (1999) (see Chap. 5).

8.3.1 Methods

8.3.1.1 Stimuli

Both masker and target were Gaussian-shaped sinusoids (see Eq. (5.1)) with a duration of 51 ms ($ERB_{GW} = \Gamma = 112.5$ Hz, $ERD_{GW} = 8.9$ ms). The carrier frequency of the masker was 0.75 kHz. Its level was 60 dB SL. Masker and target were presented simultaneously ($\Delta T = 0$). Masking patterns were measured for eight values of ΔF: 0, ±1, ±2, ±4, and +6 ERB units (the corresponding F_Ts were 408, 560, 650, 750, 861, 985, 1276, and 1638 Hz). As for Experiment 2, this range of ΔF values was chosen to avoid having a too large difference in the number of excited critical bands: with ERB_{F_T} being the ERB of the CB centered at F_T, the ratio ERB_{GW}/ERB_{F_T} was limited to values between 0.5 and 2.0. One listener (L1) was tested with two additional ΔFs, namely +8 and +10 ERB units ($F_T = 2081$ and 2642 Hz).

When F_T was above F_M ($\Delta F > 0$), we used a continuous background noise to mask potential CTs. The cut-off frequency of the noise was fixed to 555 Hz. This frequency corresponds to the upper edge of the ERB centered at $F_{CDT} = 515$ Hz, *i.e.*, the CDT expected for $F_T = 985$ Hz ($\Delta F = +2$ ERB units). In this condition, the frequency ratio between the two primaries ($F_T/F_M = 1.3$) leads to the most critical CT level. For the condition $\Delta F = +1$ ERB unit ($F_T = 861$ Hz), F_{CDT} was 639 Hz, thus too close to the masker frequency to be resolved by the ear. Nonetheless, the noise was used in this condition to mask higher-order CTs (Goldstein, 1967). The noise level was adjusted so as to totally mask a Gaussian with a carrier frequency equal to 515 Hz and a level 25 dB below L_M. This was achieved using an overall noise level of 45 dB SPL for two listeners (L1 and L3) and 50 dB SPL for one listener (L4).

8.3.1.2 Listeners

Because of time constraints, only three of the six listeners (L1, L3 and L4) could participate in Experiment 4.

8.3.1.3 Thresholds determination

To determine the amount of masking at each ΔF, absolute thresholds for the eight Gaussian targets were measured in a preliminary test. Additionally, the thresholds obtained at 0.75 kHz allowed determining the individual SL of the Gaussian masker (64–69 dB SPL).

To verify that the background noise did not affect the detection of masker or target, the condition $\Delta F = 0$ was measured with *and* without noise on all listeners. The data showed no differences (< 2 dB). Furthermore, to assess whether the presence of CTs was used as a cue to detect the target, the conditions $\Delta F = +1$ and +2 ERB units were measured with *and* without noise on one listener (L1). At $\Delta F = +1$ ERB unit, this listener obtained identical thresholds (differences < 2 dB) with and without noise, indicating that he probably did not detect CTs. Conversely, at $\Delta F = +2$ ERB units, this listener obtained higher thresholds with noise (differences > 5 dB), indicating that he may have used CTs as a cue to detect the target when the noise was absent.

For masking measurements, each session contained conditions measured with ($\Delta F = 0$, +1, +2, +4, +6, +8, and +10 ERB units) and without background noise ($\Delta F =$ -4, -2, -1, 0, +1, and +2 ERB units). Within a session, the target frequencies were chosen at random.

8.3.2 Results and discussion

Figure 8.10 presents individual and mean amounts of masking (in dB) as a function of ΔF (in ERB units). First, all listeners exhibited a peak at $\Delta F = +1$ instead of 0. Such a shift of the maximum masking frequency towards F_Ts above F_M was previously observed in simultaneous (Vogten, 1978b) and forward (Munson and Gardner, 1950; Fastl, 1979; Kidd Jr. and Feth, 1981; Lutfi, 1988) masking studies with high-level maskers ($L_M > 50$ dB SPL). This shift is commonly attributed to the displacement towards the base (*i.e.*, towards high frequencies, see Sec. 2.1.2) of the "peak" of the traveling wave pattern with increasing stimulus intensity (for a review, see, *e.g.* McFadden, 1986). Although it is a plausible explanation for the data in Figure 8.10, our results do not allow any conclusion.

Second, for all listeners and $|\Delta F| > 1$ ERB unit, the amount of masking decreased as $|\Delta F|$ increased. Only listener L4 exhibited an atypical re-increase of masking (by about 5 dB) as ΔF increased from +4 to +6 ERB units. Interestingly, this listener had exhibited the same re-increase in Experiment 2 (see Fig. 8.3) as ΔF increased from +4 to +6 ERB units. The reason of this unusual re-increase of masking in this listener remains unclear.

Large across-listener variability (11–13 dB) was obtained at $\Delta F = +4$ and +6 ERB units. This variability seems to arise from the large amounts of masking for listener L1, who obtained up to 20 dB more masking than listeners L3 and L4. Excluding listener L1 from the calculation of the mean data almost halved across-listener variability at $\Delta F = +4$, and the variability almost disappeared (< 1 dB) at $\Delta F = +6$ ERB units. To observe the masking decay over a broader frequency range for listener L1, two additional ΔF values (+8 and +10 ERB units) were tested for that listener.

This greater masking for listener L1 than for other listeners at $\Delta F = +4$ and +6 ERB units resulted in an asymmetrical pattern. Regression lines were computed for each side of the masking patterns and listener (straight lines in Fig. 8.10). To take the variability of each data point into account in the estimation of slopes, data were fitted using a weighted-least-squares method. The weight of each data point corresponded to the reciprocal of the variance of the measurement. The fit parameters are listed in Table 8.6.

For listener L1, the slope for $F_T < F_M$ was about twice that for $F_T > F_M$. Conversely, listeners L3 and L4 obtained quasi symmetrical patterns. Previous masking data measured at low frequencies (*i.e.*, with $F_M \leq 1.0$ kHz) with sinusoidal maskers of various durations and $L_M = 60$–70 dB SPL mostly reported symmetrical patterns (*e.g.*, Ehmer, 1959b,a; Green, 1969; Zwicker and Jaroszewski, 1982). Two explanations have been proposed to account for this symmetry. The first is the absence of suppression in the low-frequency region of the audible spectrum. Accordingly, physiological data showed that suppression is weak for auditory nerve fibers with CFs near and below 1 kHz (Fahey and Allen, 1985; Delgutte, 1990).

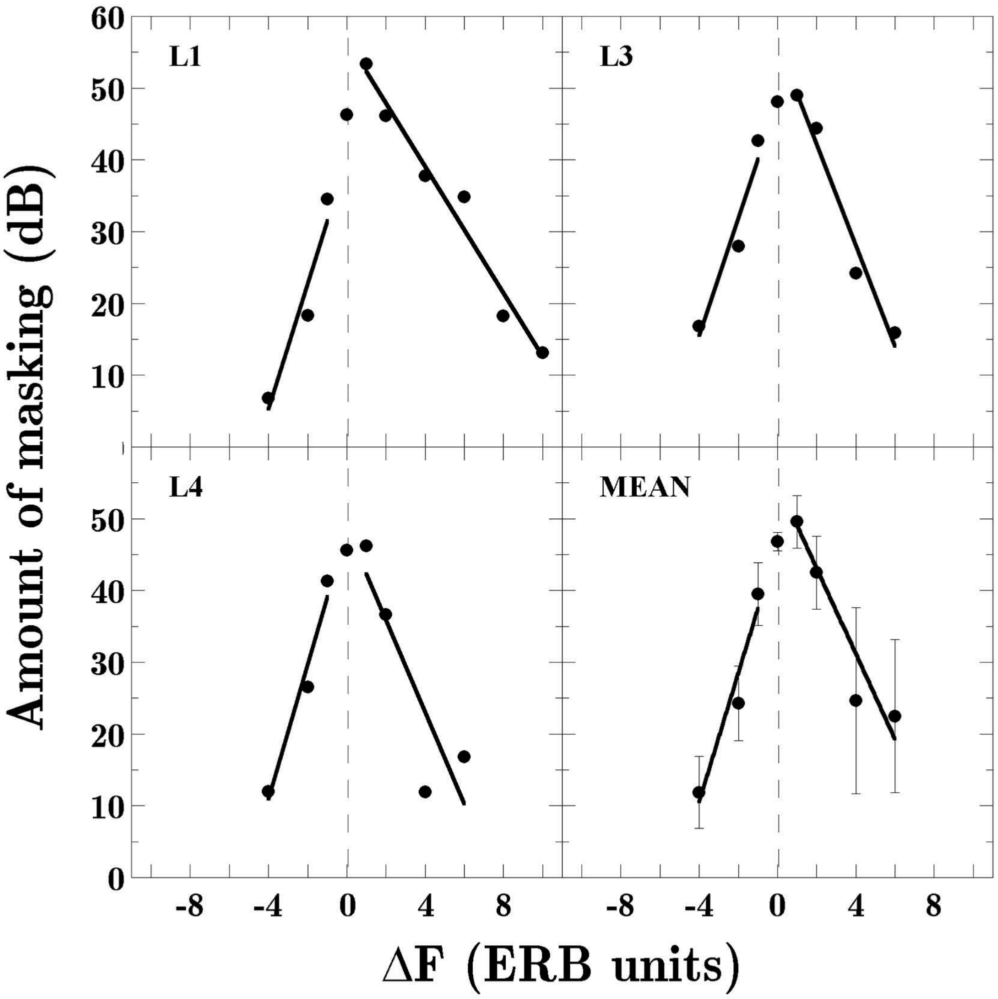


Figure 8.10: Individual amounts of masking (in dB) as a function of ΔF (in ERB units). Data were fitted with linear regression lines on each side of masking patterns (excluding the point at $\Delta F = 0$) using a weighted-least-squares method. The weight of each data point corresponded to the reciprocal of the variance of the measurement. The fit parameters are listed in Table 8.6. The bottom right panel shows the mean data with ±1 standard deviation bars.

The second explanation is the activation of the middle-ear reflex (see Secs. 2.1.1 and 3.2.6). Pang and Guinan (1997) and Liberman and Guinan (1998) indeed showed that the activation of the reflex can decrease the amount of suppressive masking (by as much as 40 dB) of high-frequency targets ($F_T > 1.5$ kHz) in the presence of low-frequency maskers ($F_M < 1$ kHz). The middle-ear reflex is known to be activated around 90 dB SPL for long-lasting sinusoids in normal-hearing humans. The SPLs (referenced at the stationary carrier, see Sec. 6.4) of the 0.75-kHz Gaussian maskers in Experiment 4 were about 73, 74 and 78 dB for listeners L1, L3 and L4, respectively. Although these SPLs are much lower than the reported thresholds for triggering the middle-ear reflex, to our knowledge, no data are available on the reflex activation for impulsive sounds such as the brief Gaussian stimuli in the present study. Therefore, we do not exclude that listeners L3 and L4, who obtained symmetrical patterns, were subject to the reflex activation.

Listener	Linear fit	Side of the masking pattern	
		$F_T < F_M$	$F_T > F_M$
L1	Slope	+8.74	−4.41
	r^2	0.92	0.97
L3	Slope	+8.20	−7.04
	r^2	0.93	0.97
L4	Slope	+9.41	−6.41
	r^2	0.96	0.77

Table 8.6: Slope (in dB/ERB) and r^2 values of the linear regression lines plotted in Figure 8.10.

The asymmetrical pattern for listener L1 could be attributed to individual differences in the degree of suppression (Egan and Hake, 1950; Rodriguez et al., 2010), and/or to the inactivation of the middle-ear reflex in that listener.

To assess whether the masking pattern for $F_M = 0.75$ kHz was broader or sharper than that for $F_M = 4.0$ kHz, we estimated the quality factor at the 3-dB bandwidth, Q_{3dB}, from the mean data of Figure 8.10. The estimation provided a value of $Q_{3dB} = 8.6$ for $F_M = 0.75$ kHz, against $Q_{3dB} = 11$ for $F_M = 4.0$ kHz (see Fig. 8.5). This indicates that the pattern broadened as F_M was decreased from 4.0 to 0.75 kHz, consistent with previous data (see Sec. 3.2). Nevertheless, the broadening of the pattern observed here is less pronounced than that reported in the literature (Ehmer, 1959b,a; Green, 1969; Zwicker and Jaroszewski, 1982; Moore et al., 1998). For example, in Ehmer (1959a), Q_{3dB} dropped from 14 to 6 as the frequency of a pure-tone masker was decreased from 4.0 to 0.5 kHz, and from 21 to 8 for a narrowband-noise masker with same frequencies (see Fig. 3.5). The broadening of masking patterns at low frequencies is commonly attributed to the facts that (1) low-frequency maskers excite a wider portion of the BM than high-frequency ones, and (2) because of the very narrow CB width at low frequencies, low-frequency maskers excite several adjacent CBs around F_M (see Sec. 3.2.6). In the present study, we attempted to restrict the energy spread of the masker (ERB_{GW}) to one CB (*i.e.*, by keeping $ERB_{GW}/ERB_{F_M} \approx 1$). This likely explains the fact that our pattern only slightly broadened when decreasing F_M.

If the 4.0- and 0.75-kHz Gaussian maskers actually produced similar energy spreads in their respective frequency regions, then their masking patterns should have roughly similar shapes when plotted on an ERB scale (Moore et al., 1998; Zwicker and Feldtkeller, 1999). This is verified in Figure 8.11, where the mean results from Experiment 4 (filled circles) are plotted with those from Experiment 2 (empty circles) as a function of ΔF (in ERB units). For almost all ΔFs, lower amounts of masking were obtained with the 0.75- than with the 4.0-kHz masker (comparable amounts of masking were solely obtained at $\Delta F = +1$ and $+6$ ERB units). For ΔFs below 0 (*i.e.*, $F_T \leq 0.65$ kHz), this could be attributed to the elevated threshold in quiet in the lower frequency portion of the audible spectrum (Zwicker and Jaroszewski, 1982). For ΔFs above +1 (*i.e.*, $F_T \geq 0.98$ kHz), this could be attributed to the absence of suppression and/or to the activation of the middle-ear reflex. Overall, both masking patterns in Figure 8.11 have similar shapes and comparable spectral widths ($Q_{3dB} = 11$ *vs.* 8.6 with $F_M = 4.0$ and 0.75 kHz, respectively). This is compatible with the constant-Q frequency analysis by the human auditory system (see Sec. 2.3).

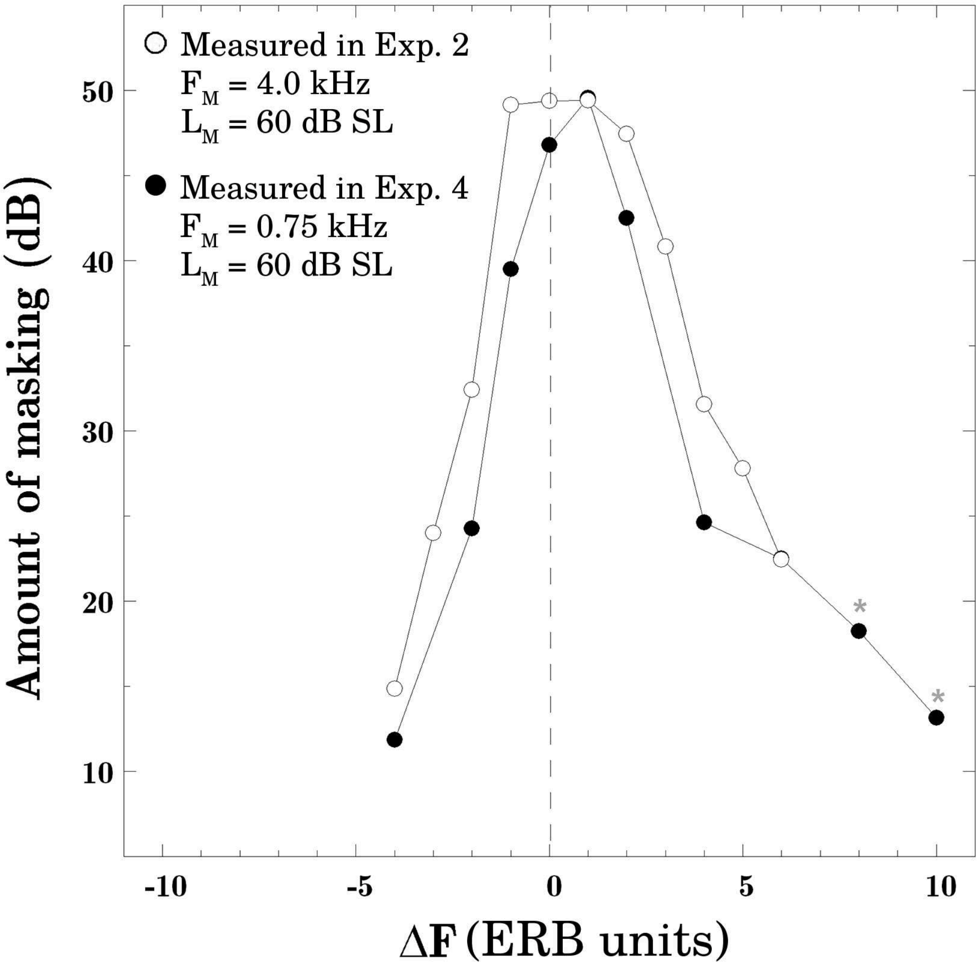


Figure 8.11: The amount of masking (in dB) is plotted as a function of ΔF (in ERB units). The mean results from Experiment 4 (•) are plotted together with the mean results from Experiment 2 (○). Stars indicate values measured for listener L1 only. For clarity, the standard deviation bars were omitted.

Summary

In Experiment 4, masked thresholds were measured for eight ΔFs with $F_M = 0.75$ kHz and $L_M = 60$ dB SL in three listeners. To minimize the number of activated TF windows of the auditory system, the parameter Γ of the Gaussian window was fixed to 112.5. Thus, masker and target had an ERB of 112.5 Hz and an ERD of 8.9 ms (effective duration = 51 ms).

For all listeners, the patterns showed a peak at $\Delta F = +1$ instead of 0. Although our data did not allow any conclusion on this effect, this might be related to the displacement towards the base of the top of the cochlear vibration pattern resulting from masker and target excitations on the BM. Two listeners obtained symmetrical patterns, consistent with previous data for short and long sinusoidal maskers with comparable level and frequency. One listener obtained an asymmetrical pattern with a steeper masking decay for $F_T < F_M$ than for $F_T > F_M$. This asymmetry could be attributed to individual differences in the degree of suppression and/or to the inactivation of the middle-ear reflex in that listener. The masking pattern for $F_M = 0.75$ kHz was only slightly broader than that for $F_M = 4.0$ kHz. This could be attributed to the fact that low-frequency maskers excite a wider portion of the BM than high-frequency ones. Both patterns had similar shapes and comparable spectral widths when plotted on an ERB scale. This is compatible with the constant-Q frequency analysis by the auditory system.

The goal of the frequency masking experiments presented in this chapter was not the improved understanding of the mechanisms underlying auditory masking. Rather, the results from these experiments allowed us to show that masking patterns for Gaussian maskers are consistent with those previously reported in the literature for sinusoidal maskers with comparable frequencies and levels and temporally controlled with windows limiting spectral broadening.

Chapter 9

Temporal masking with Gaussian stimuli (Experiment 5)

Contents

In Experiment 5, we investigated temporal masking for Gaussian stimuli. Masker and target had the same frequency ($\Delta F = 0$), and the amount of masking was measured for various temporal separations (ΔT) between masker and target. Because a pilot test indicated very little backward masking for our 9.6-ms Gaussian masker[1], in this and subsequent experiments, we focused on forward masking, *i.e.*, the masker always temporally *preceded* the target.

9.1 Methods

9.1.1 Stimuli

Stimuli were Gaussian-shaped sinusoids (defined in Eq. (5.1)) with a duration of 9.6 ms ($ERB_{GW} = \Gamma = 600$ Hz, $ERD_{GW} = 1.7$ ms). Masker and target had the same carrier frequency ($F_T = F_M = 4$ kHz). L_M was fixed to 60 dB SL (72–75 dB SPL depending on the listener). The temporal separation between masker and target (ΔT), defined as the “peak-to-peak" distance, was 0, 5, 10, 20, or 30 ms.

[1] Accordingly, in a collaborative study using identical stimuli, we found that a backward masker at $\Delta T = 8$ ms (counting from target peak to masker peak) required, on average, a level of 81 dB SPL to produce about 8 dB of masking (see Laback et al., 2008). In comparison, the forward masker at $\Delta T =$ -8 ms in the same study required an average level of 66 dB SPL to produce about the same amount of masking. This demonstrates that backward masking is much weaker than forward masking, consistent with previous studies (see Sec. 3.3).

9.1.2 Thresholds determination

The six listeners (L1–L6) participated in the present experiment. Masked thresholds were measured using the 3-IFC adaptive procedure described in Section 6.2. Within a session, the ΔT values were chosen at random.

9.2 Results

Figure 9.1 presents individual and mean amounts of masking as a function of ΔT on a linear (Fig. 9.1a) and a logarithmic scale (Fig. 9.1b). On average, masking decreased from 50 dB for $\Delta T = 0$ to about 6 dB for $\Delta T = 30$ ms. When plotted on a logarithmic scale, the data for $\Delta T > 0$ are well fitted with straight lines. A straightforward description of these data is provided by

$$AM = a \log(\Delta T) + b \tag{9.1}$$

where AM is the amount of masking, a is the slope of the forward masking decay, and b is the offset of the forward masking decay. Table 9.1 lists the values of a and b determined by applying a weighted-least-squares fit of Equation (9.1) to the data for $\Delta T > 0$ in Figure 9.1b. To take the variability of each data point into account in the estimation of parameters a and b, the weight of each data point corresponded to the reciprocal of the variance of the measurement. The slope estimates (a) largely depend on the listener, ranging from -36 to -14 dB per log(ΔT). This variability seems to arise from the variability of threshold estimates for the condition $\Delta T =$ 5 ms (see error bars in the bottom panels of Fig. 9.1), discussed below. Excluding the condition $\Delta T = 5$ ms from the calculation of the slopes almost halves across-listener variability in the slope estimates.

	Parameter values		
Listener	a	b	r^2
L1	−14.39	28.46	0.98
L2	−23.38	39.93	1.00
L3	−25.61	44.28	0.99
L4	−29.37	48.60	0.95
L5	−36.00	56.01	0.96
L6	−17.76	28.26	0.97
MEAN	−23.18	39.12	0.97

Table 9.1: Values of parameters a (in dB/log(ΔT)) and b (in dB) determined by fitting Equation (9.1) to the data for $\Delta T > 0$ in Figure 9.1b using a weighted-least-squares criterion. The last column indicates the r^2 values.

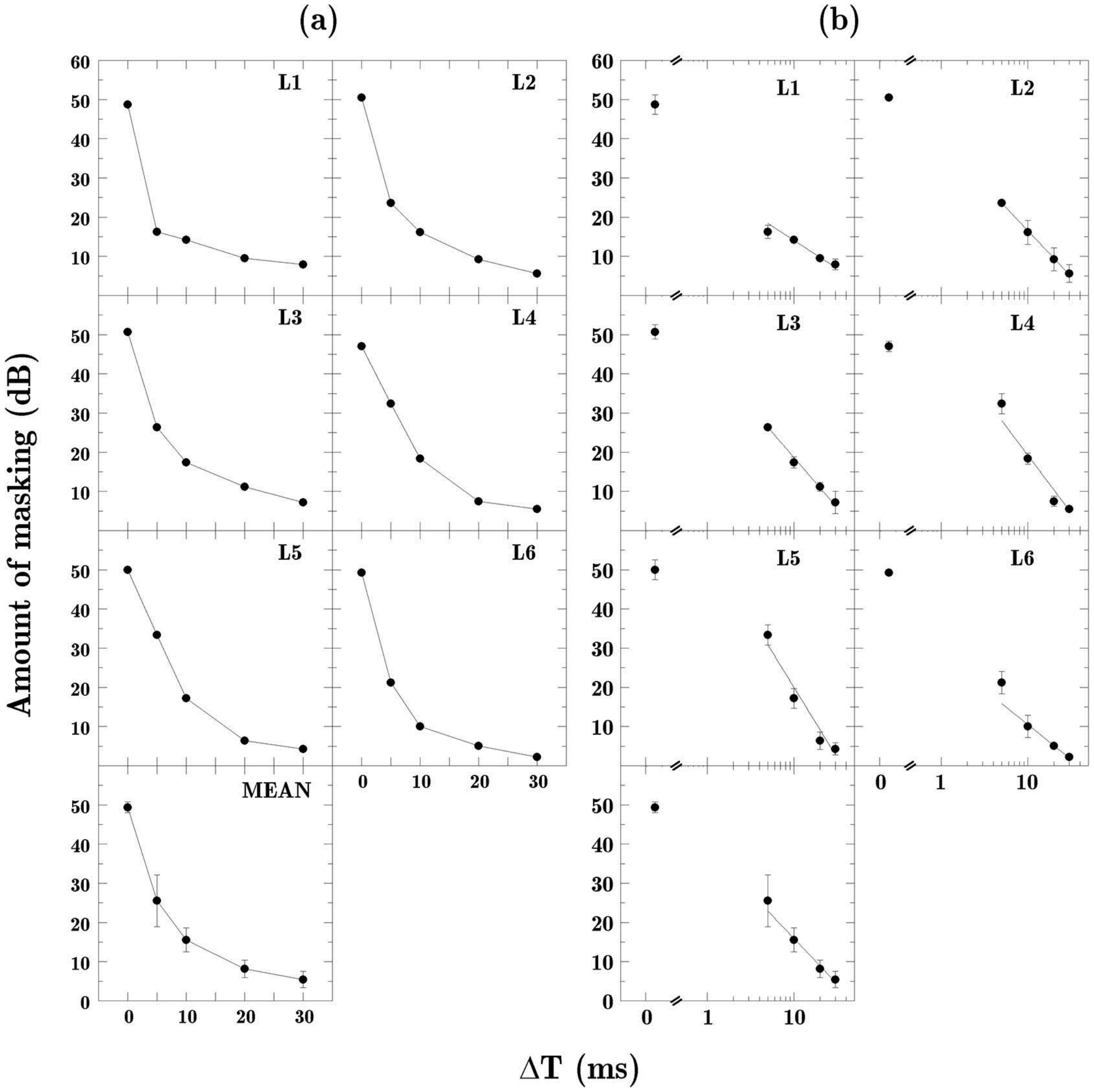


Figure 9.1: Individual amounts of masking (in dB) as a function of ΔT (in ms) (a) on a linear scale and (b) on a logarithmic scale with straight-line fits to the data for $\Delta T > 0$ (see Eq. (9.1)). The fit parameters are listed in Table 9.1. Error bars in the individual panels indicate ±1 standard deviation of each measurement. The bottom panels show the mean data with ±1 standard deviation bars.

9.3 Discussion

In Figure 9.1b, the decay of forward masking was a linear function of log(ΔT), a result consistent with almost all previous forward masking studies using various types of maskers (see Sec. 3.3.2 and, *e.g.*, Duifhuis 1973; Penner 1974; Fastl 1979; Jesteadt et al. 1982; Zwicker 1984; Carlyon 1988). To examine how the temporal spread of masking for our narrowband *and* short Gaussian masker compares to those for maskers with various spectral and temporal characteristics, Figure 9.2 compares the mean results from Experiment 5 (**b**, bold line) with studies using (**a**) a 2.5-ms Hamming-shaped sinusoid (Duifhuis, 1973), (**c**) a 10-ms uniformly-masking noise (Zwicker, 1984), (**d**) a 300-ms sinusoid (Fastl, 1979), and (**e**) a 500-ms uniformly-masking noise (Fastl, 1976). All cited studies used maskers with comparable frequencies ($F_M = 4$ kHz) and levels ($L_M = 70$–80 dB SPL, except Duifhuis, 1973, who used a 40-dB SL masker) and short (duration = 1–20 ms) sinusoidal targets. The corresponding spread of the stimuli across the TF plane and the recovery times, defined as the offset-onset interval at which 5 dB of masking is reached, are specified in the insert for each masking curve.

First, for masker offset-to-target onset intervals of 0–20 ms, it can be seen that long maskers (squares) produce the greatest masking, thus the longest recovery times ($\geq$ 65–80 ms). The 2.5-ms masker (triangles) conversely produces the shortest recovery time ($<$ 10 ms). This is consistent with studies of forward masking studies investigating the effect of masker duration (*e.g.*, Penner, 1974; Kidd Jr. and Feth, 1982; Zwicker, 1984). Second, for a given masker duration (either temporally narrow or broad), it can be seen that increasing the masker bandwidth results in an increase of masking (Duifhuis, 1973; Widin and Viemeister, 1979b; Moore, 1981). Overall, the spread of temporal masking produced by the 9.6-ms Gaussian masker (recovery time = 20 ms) is about twice that produced by the 2.5-ms Hamming-shaped masker ($<$ 10 ms), but smaller than that produced by the 10-ms noise masker (40 ms).

The physiological mechanisms underlying forward masking were described in Section 3.3.4. The present results are compatible with an explanation based on the exponential decay of masker-induced activity over time in the cochlea and in the auditory nerve (short-term adaptation), and by more central effects of persistence of the neural activity evoked by the masker. With small values of ΔT (remember that in the present study ΔT was defined as peak-to-peak and varied from 0 to 30 ms) inducing a temporal overlap of the cochlear responses to masker and target, masking can be attributed to the mechanisms of simultaneous masking (Duifhuis, 1973; Carlyon, 1988; Nizami and Schneider, 2000). The cited authors suggested that the transition from simultaneous to forward masking should be marked by a radical decrease in the amount of masking. They proposed that the “critical" value of ΔT at which the transition occurs is frequency-dependent. Namely, the “ringing” of the auditory filters may result in the overlap of the BM responses to masker and target even if there is no physical overlap of the signals. The ringing time at 4 kHz being about 2.2 ms (Carlyon, 1988; Nizami and Schneider, 2000), the temporal overlap may have contributed to the present results for the smallest ΔT values. Accordingly, we measured forward masking for ΔTs of 1 and 2 ms on two listeners (L1 and L2) in pilot tests. Thresholds were identical to those for $\Delta T = 0$. According to the hypothesis from Duifhuis (1973) and Carlyon (1988), this suggests

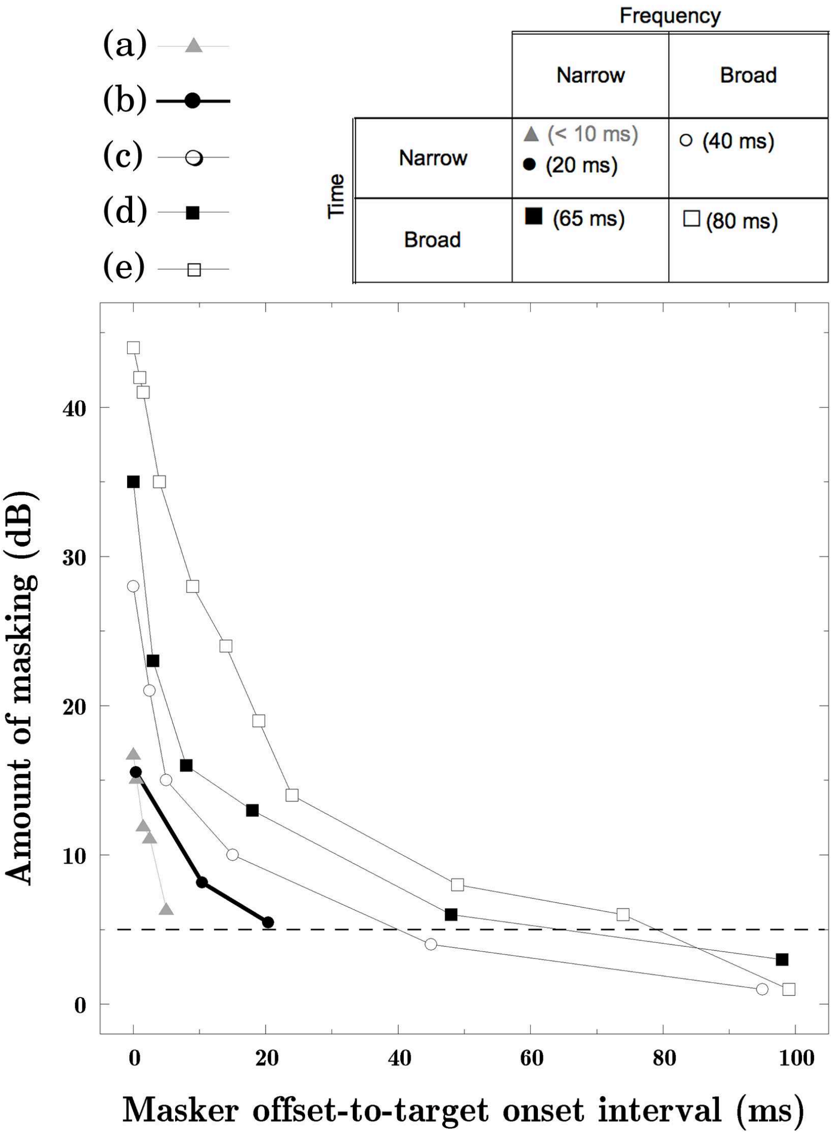


Figure 9.2: Amount of masking (in dB) as a function of masker offset-to-target onset interval (in ms). The mean results from Experiment 5 (**b**, bold line) are compared with (**a**) a 2.5-ms Hamming-shaped sinusoid (Duifhuis, 1973), (**c**) a 10-ms (incl. 1-ms Gaussian rise/fall times) uniformly-masking noise (Zwicker, 1984), (**d**) a 300-ms (incl. 1-ms Gaussian rise/fall times) sinusoid (Fastl, 1979), and (**e**) a 500-ms (incl. 0.5-ms Gaussian rise/fall times) uniformly-masking noise (Fastl, 1976). The corresponding spread of the stimuli across the TF plane and the recovery times, defined as the offset-onset interval at which 5 dB of masking is reached, are specified in the insert. For clarity, the simultaneous conditions (*i.e.*, offset-onset intervals < 0 ms) measured in (a), (b) and (c) were not plotted.

that for $\Delta T = 1$ and 2 ms, the mechanisms underlying simultaneous masking were much more effective than those underlying forward masking.

Interestingly, large across-listener variability was observed at $\Delta T = 5$ ms. Because the effective duration of the signals was 9.6 ms and that ΔT was defined as the peak-to-peak separation, in that condition, there was a 4.6-ms overlap of the two signals. This variability may be attributed to individual differences in temporal resolution. Indeed, studies on temporal resolution provided estimates in the range 2–8 ms (see Sec. 2.2.1). Thus, the listeners who obtained the lowest thresholds would be those with sufficiently good temporal resolution to detect the temporal gap between the two peaks. Another explanation for this variability might be individual differences in cochlear filter ringing (Nizami and Schneider, 2000). If the ringing actually contributed to the results for $\Delta T = 5$ ms, then the listeners who obtained the highest thresholds would be those with the longest ringing times.

Summary

In Experiment 5, masked thresholds were measured as a function of the temporal separation (ΔT) between the Gaussian masker and target, which had the same frequency ($F_M = F_T = 4$ kHz). The masker had a fixed level of 60 dB SL.

The decay of forward masking was a linear function of $\log(\Delta T)$, a result consistent with previous data for maskers with various spectro-temporal characteristics. Forward masking was limited to ΔTs $\leq$ 30 ms, whereas the literature reported masking for ΔTs up to 100–150 ms with long (duration $\geq$ 200 ms) maskers. This narrow temporal spread of masking can be attributed to the short duration of the Gaussian masker. The present results are consistent with an explanation based on the temporal decay of masker-induced excitation at both peripheral and more central levels. For small ΔT values, masking may also be due to the overlap of the BM responses to masker and target.

The goal of Experiment 5 was not the improved understanding of the mechanisms underlying forward masking. Rather, the result from this experiment allowed us to show that the temporal spread of forward masking for a 9.6-ms Gaussian masker is consistent with those previously reported for maskers with comparable durations, bandwidths, frequencies and levels.

Chapter 10

Time-frequency masking (Experiment 6)

Contents

In Experiment 2, auditory masking was measured as a function of the frequency separation (ΔF) between the Gaussian masker and target, which were presented simultaneously ($\Delta T = 0$). In Experiment 5, masking was measured as a function of the temporal separation (ΔT) between masker and target, which had the same frequency ($\Delta F = 0$). In Experiment 6 presented here, both ΔT and ΔF were varied.

10.1 Methods

10.1.1 Stimuli

Both masker and target were Gaussian-shaped sinusoids (defined in Eq. (5.1)) with a duration of 9.6 ms ($ERB_{GW} = \Gamma = 600$ Hz, $ERD_{GW} = 1.7$ ms). The masker had fixed frequency ($F_M = 4$ kHz) and level ($L_M = 60$ dB SL). The initial panel of TF conditions included five ΔTs (0, 5, 10, 20, and 30 ms) and eight ΔFs (0, ±1, ±2, ±4, and +6 ERB units). Masked thresholds were measured for 32 out of the 40 possible $\Delta T \times \Delta F$ combinations, based on a pilot test.

Although the effect of CTs is usually ignored in forward masking studies, we used a background noise identical to that of Experiment 2 to mask potential CTs (when

F_T was above F_M) because of the small ΔT values. To verify that the background noise did not affect masker or target detection, the condition $\Delta F = 0$ was measured with *and* without noise for all five ΔTs. The difference in threshold was less than 3 dB for all listeners and ΔTs.

10.1.2 Thresholds determination

Because of time constraints, only four of the six listeners (L1–L4) could participate in Experiment 6. The whole set of conditions was split into two groups: frequency separations measured with ($\Delta F = 0$, +1, +2, +4, and +6 ERB units) and without background noise ($\Delta F =$ -4, -2, -1, and 0 ERB units). Then, experimental blocks were formed that contained the ΔT conditions for each ΔF. The order of blocks and groups was randomized across sessions. Within a session, the target frequency was fixed and ΔT was chosen randomly.

10.2 Results

Figure 10.1 presents the results as a function of frequency (ΔF) with time (ΔT) as the parameter. First, the dip/plateau observed at $\Delta F = 0$ for $\Delta T = 0$ (see also Fig. 8.3) almost disappeared when ΔT increased. For ΔTs > 0, two of the four listeners (L1 and L3) exhibited a peak at $\Delta F = +1$ instead of 0. Second, the results from all listeners revealed a drastical decrease in the amount of masking as ΔT increased from 0 to 30 ms. For all frequency separations, the largest amount of masking was obtained in the simultaneous condition ($\Delta T = 0$). This masking dropped by 20–40 dB as ΔT increased to 5 ms for small frequency separations (*i.e.*, $|\Delta F| \leq 2$ ERB units). For ΔTs > 10 ms, masking was generally less than 10 dB for all frequency separations.

To examine the asymmetry of the patterns in Figure 10.1, we computed the linear regression lines for each side and for $\Delta T = 0$, 5, and 10 ms (conditions $\Delta F = 0$ were excluded from the fits). Also, by calculating the intersection of the two regression lines and dividing the frequency (in Hz) of the intersection point by the bandwidth 3 dB down from the amount of masking at the intersection point, we obtained the estimates of Q_{3dB} for each pattern. Individual and mean slopes (in dB/ERB) and Q_{3dB} are listed in Table 10.1. For $\Delta T = 0$ and for all listeners, the patterns are asymmetric. For ΔTs > 0, listener L2 showed no asymmetry, and listener L4 showed asymmetry only for $\Delta T = 10$ ms. Listeners L1 and L3 showed asymmetry for all ΔTs. The mean data indicate an asymmetry for $\Delta T = 0$ and 10 ms, and symmetry for $\Delta T = 5$ ms.

The values of Q_{3dB} in the last column of Table 10.1 allow assessing whether the masking patterns in Figure 10.1 become broader or narrower with increasing ΔT. For all listeners, Q_{3dB} roughly decreased (and hence, masking patterns roughly broadened) as ΔT increased from 0 to 5 ms. When ΔT further increased to 10 ms, the broadening of the patterns was much less pronounced. This might be due to floor effects, *i.e.*, to the fact that for $\Delta T = 10$ ms and $|\Delta F| \geq +1$ ERB unit, masking was generally less than 10 dB. Finally, for $\Delta T = 5$ ms, there was large across-listener variability for small frequency separations (*i.e.*, $-1 \leq \Delta F \leq +2$ ERB units).

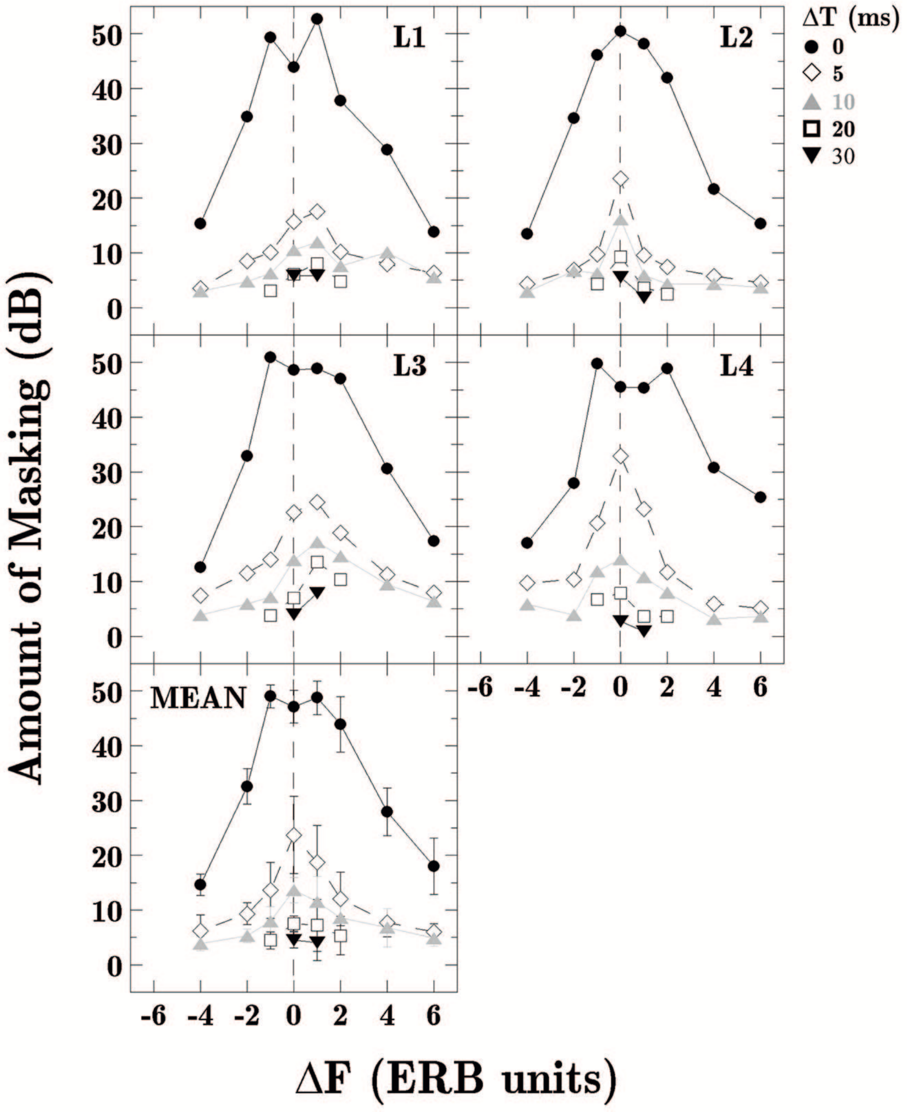


Figure 10.1: Individual amounts of masking (in dB) as a function of ΔF (in ERB units) obtained for five ΔT values. The bottom panel shows the mean data with ±1 standard deviation bars.

In Figure 10.2, the same data are presented as a function of ΔT on a logarithmic scale with ΔF as the parameter. For all listeners and ΔTs, the amount of masking decreased as $|\Delta F|$ increased. Surprisingly, only listener L2 showed this decrease when ΔF increased from 0 to +1 ERB unit. For each ΔF, the decay of forward masking was a linear function of log(ΔT). For $\Delta F = 0$, ±1 and +2 ERB units, a weighted-least-squares fit of Equation (9.1) was applied to the data for $\Delta T > 0$ (the fit was not applied to other ΔFs because only two points were measured). As for Experiment 5, the weights corresponded to the reciprocal of the variance of the measurements as to take into account the variability of each data point. The values of parameters a and b obtained for each ΔF are listed in Table 10.2. For all listeners, the slope of this decay (a) decreased as $|\Delta F|$ increased. On average, the slope decreased from about -21 to -14 dB/log(ΔT) as ΔF “increased” from 0 to -1 ERB units, and from about -21 to -11 dB/log(ΔT) as ΔF increased from

Listener	ΔT	Side of the masking pattern		Q_{3dB}
	(ms)	$F_T < F_M$	$F_T > F_M$	
L1	0	+11.04	−4.75	9.52
	5	+2.24	−1.96	3.03
	10	+1.50	$-0.93^\star$	1.57
L2	0	+10.82	−6.96	12.30
	5	+1.73	−0.94	1.63
	10	+1.24	−1.60	2.10
L3	0	+12.40	−6.64	12.51
	5	+2.18	−3.28	3.96
	10	+1.09	−2.18	2.31
L4	0	+10.15	−4.75	9.30
	5	$+3.19^\star$	−3.28	4.74
	10	+8.00	−2.51	5.34
MEAN	0	+11.12	−6.39	11.75
	5	+2.34	−2.36	3.45
	10	+2.60	−1.24	2.29

Table 10.1: Slopes (in dB/ERB) and values of Q_{3dB} for three ΔTs as estimated from the data in Figure 10.1. Stars indicate linear regressions with r^2 values < 0.75.

0 to +2 ERB units. Accordingly, Soderquist et al. (1981), who measured forward masking for various ΔFs with a 250-ms sinusoidal masker at 1 kHz, obtained an average slope of 17 dB/log(ΔT) at $\Delta F = +0.4$ ERB units. This slope decreased to about 5 dB/log(ΔT) as ΔF increased to +3.7 ERB units.

Kidd Jr. and Feth (1981), who measured forward masking for various ΔFs on two listeners with a 300-ms sinusoidal masker at 1 kHz, found that for each listener, the ratio b/a was approximately constant regardless of ΔF. Given that the ΔT-axis intercept is $10^{-b/a}$, their finding would imply that for each listener in Figure 10.2, all straight lines reach 0 dB of masking around a common ΔT value, referred to as ΔT_{0dB}. From the values of a and b indicated in Table 10.2, average values of ΔT_{0dB} were computed for each listener across all ΔFs. These values are 58.2 ms for listener L1 (standard deviation = 14.6 ms), 46.7 ms for L2 (6.7 ms), 63.0 ms for L3 (31.4 ms), and 32.8 ms for L4 (4.0 ms). Apart from listener L3 for whom a great variability was observed, our results are consistent with those from Kidd Jr. and Feth (1981).

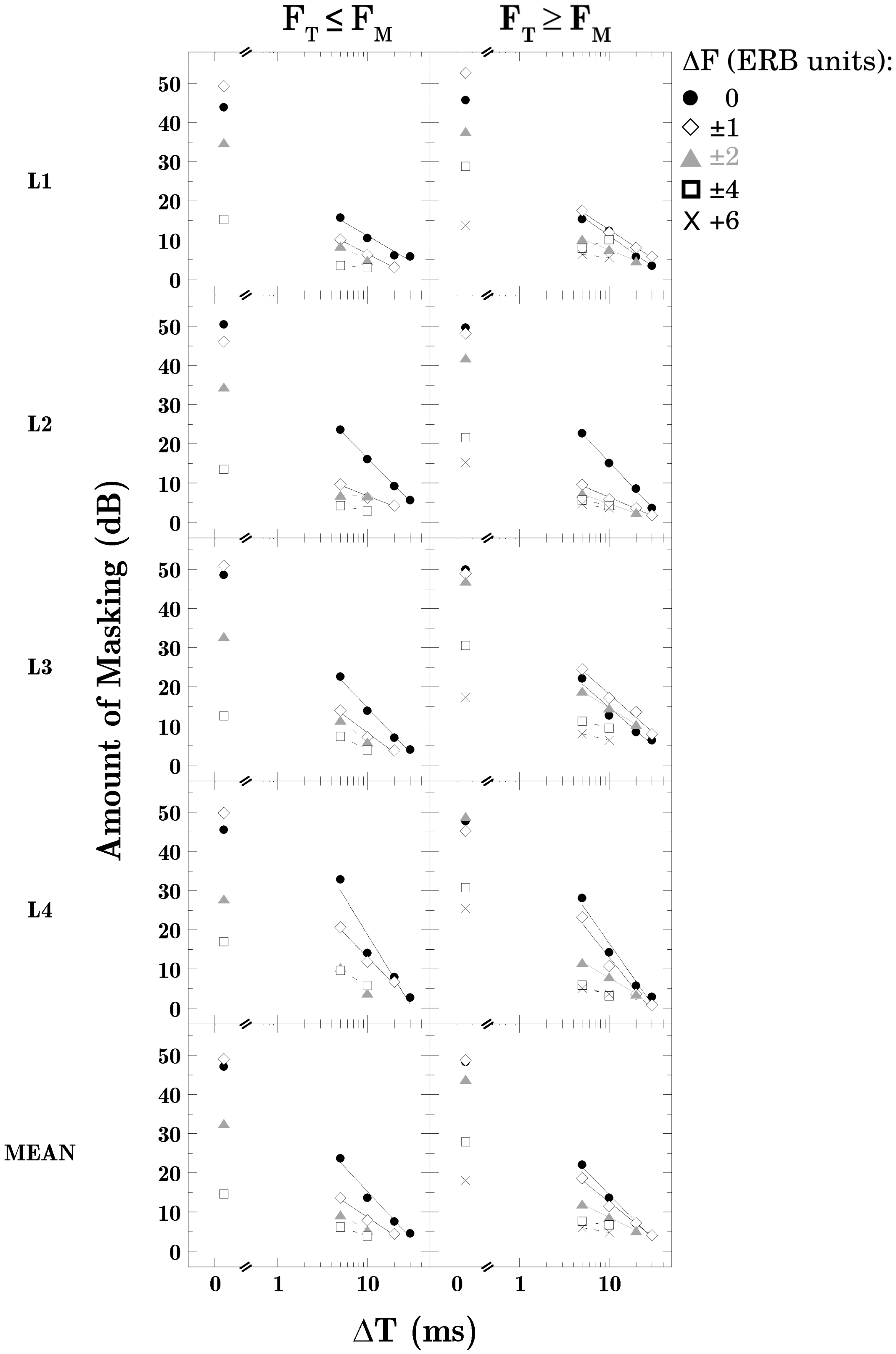


Figure 10.2: Individual amounts of masking (in dB) as a function of ΔT (in ms) on a logarithmic scale. The parameter is ΔF. Panels in the left column show data for $F_T \leq F_M$. Panels in the right column show data for $F_T \geq F_M$. Mean data are shown in the bottom panels. For clarity, standard deviation bars are not represented. For $\Delta F = 0$, ±1 and +2 ERB units, a weighted-least-squares logarithmic regression was applied to the data for $\Delta T > 0$ (see Tab. 10.2).

	ΔF	$F_T \leq F_M$		$F_T \geq F_M$	
Listener	(ERB units)	a	b	a	b
L1	0	−12.82	23.54	−15.77	26.77
	1	−11.50	17.94	−15.17	27.84
	2			−9.05	16.58
L2	0	−23.38	39.93	−23.66	39.38
	1	−9.32	16.15	−10.00	16.54
	2			−8.57	13.39
L3	0	−25.46	40.20	−19.17	34.24
	1	−17.63	26.10	−20.21	38.22
	2			−14.12	28.73
L4	0	−34.68	53.90	−31.00	47.24
	1	−25.31	38.00	−29.88	42.48
	2			−13.90	21.69
MEAN	0	−20.46	34.45	−20.86	34.63
	1	−13.66	22.15	−17.53	29.89
	2			−11.19	19.86

Table 10.2: Values of parameters a (in dB/log(ΔT)) and b (in dB) determined by fitting Equation (9.1) to the data for $\Delta F = 0, \pm 1, +2$ ERB units and for $\Delta T > 0$ in Figure 10.2 using a weighted-least-squares criterion. In all cases, r^2 values were greater than 0.94.

10.3 Discussion

As stated above (see Secs. 3.3.2 and 4.2), there are a few TF masking data for narrowband stimuli in the literature. Some recent studies (Lopez-Poveda et al., 2003; Yasin and Plack, 2005) measured temporal masking functions for various ΔFs. These studies involved sinusoidal maskers at 4 kHz with steady-state durations of 100 and 200 ms, respectively. The masker level required to mask a fixed-level target was measured (as for the measurement of PTCs, see Sec. 2.3.2). For that reason, their results could hardly be compared to ours.

In other studies (Fastl, 1979; Kidd Jr. and Feth, 1981; Soderquist et al., 1981), frequency masking patterns were measured for various ΔTs. These studies involved sinusoidal maskers ($F_M = 1$ or 4 kHz) with comparable levels (60–77 dB SPL) but with much longer durations (250–300 ms) than in the present study. Their results indicated that (1) the amount of frequency masking rapidly decreases as ΔT increases, (2) forward masking patterns are broader than simultaneous masking patterns, and (3) the masking patterns' asymmetry remains for ΔTs > 0. These three findings are all in agreement with our data.

Forward masking patterns are thought to reflect the temporal decay of masker-induced excitation (otherwise called “residual masking" by Kidd Jr. and Feth, 1981) on the BM and the auditory nerve. Based on this assumption, the broadening of the masking patterns with increasing ΔT is expected since (1) the decay of forward masking becomes slower as ΔF increases (see Tab. 10.2), and (2) for $\Delta F = 0$, the amount of masking for $\Delta T > 0$ is smaller than for $\Delta T = 0$. In other terms, the flattening of the masking patterns with increasing ΔT could result from the steepness of the forward masking decay, which itself depends on ΔF.

In Section 9.3, it has been suggested that the temporal overlap of the cochlear responses to masker and target with small values of ΔT may have increased masking. Similarly, this phenomenon may have played a role in the results from Experiment 6 with small ΔFs (Duifhuis, 1973; Soderquist et al., 1981). The large across-listener variability observed for $\Delta T = 5$ ms with small values of ΔF indeed strengthens the temporal overlap hypothesis. In these conditions, the masker and the target had close carrier frequencies. Thus, the listeners probably focused on temporal cues to detect the target. This variability may therefore be attributed to individual differences in temporal resolution and/or cochlear filter ringing (see Sec. 9.3).

The occurrence of peaks at $\Delta F = +1$ is consistent with previous forward masking studies (Munson and Gardner, 1950; Fastl, 1979; Kidd Jr. and Feth, 1981; Lutfi, 1988). The shift of the maximum masking frequency towards F_Ts above F_M is commonly attributed to the displacement towards the base of the “peak” of the traveling wave pattern with increasing stimulus level (McFadden, 1986). This is a plausible interpretation for the present data obtained with small ΔT values inducing a temporal overlap of the BM responses to masker and target. However, the origin of the peaks observed at $\Delta F = +1$ in listeners L1 and L3 for ΔTs up to 30 ms remains unexplained.

In Figure 10.2 and Table 10.2, we showed that (1) varying ΔF does not affect the linear decay characteristic of forward masking as a function of $\log(\Delta T)$, (2) the slope of this decay decreases as $|\Delta F|$ increases, and (3) for a given listener, 0 dB of masking would be reached, if tested, at approximately the same value of ΔT for all ΔFs. These results are closely consistent with those from Kidd Jr. and Feth (1981), and Soderquist et al. (1981).

Interestingly, an analogy can be made between the forward masking curves for various ΔFs (see Fig. 10.2) and the level dependency of forward masking, well reported in the literature (see Fig. 3.10 and *e.g.*, Widin and Viemeister 1979a; Jesteadt et al. 1982; Moore and Glasberg 1983a). This suggests that the amount of masking for any combination of ΔF and ΔT could be predicted from forward masking data measured at $\Delta F = 0$, simply by assuming the presence of a lower-level forward masker. This issue is addressed below.

10.4 Prediction of time-frequency masking patterns

In Chapter 4, the currently implemented masking models in perceptual audio codecs (such as MPEG 1 Layer III) were described. Emphasis was placed on the fact that most of audio coding schemes use simultaneous masking functions only to estimate the masking thresholds of TF components. A few models were developed to exploit both temporal and frequency masking in such coders (Lincoln, 1998; Vafin et al., 2000; Huang and Chiueh, 2002; Najaf-Zadeh et al., 2003; He and Scordilis, 2008). To achieve combined TF masking models, most of the cited studies simply assumed a linear combination of simultaneous and forward masking functions. However, given the highly non linear behavior of cochlear mechanics (see Sec. 2.1.4 and, *e.g.*, Ruggero et al. 1997; Lopez-Poveda et al. 2003), such a simple combination of temporal and frequency masking functions is unlikely to provide an accurate representation of TF masking. Moreover, these studies mostly based their models on psychoacoustical data for stimuli that are not maximally concentrated in the TF domain (see Sec. 4.1.3).

To prove that the spread of TF masking cannot be deduced from masking data measured separately in the time and frequency domains, we tested two simple prediction schemes assuming a linear combination of temporal and frequency masking. Specifically, we assessed whether the results from Experiment 6 (TF masking) could be predicted by the results from Experiment 2 (frequency masking) and 5 (temporal masking). The goal of the two approaches presented below is to question a representation of TF masking used in audio codecs. These approaches *do not intend* to model the non linear processing of the auditory system. The general idea of the prediction is that the spread of TF masking caused by a masker can be described by the frequency masking pattern combined with the decay of forward masking from each point of the masking pattern. In the following, let $AM(\Delta T, \Delta F)$ denote the amount of masking produced by the masker on a target separated from the masker by ΔT and ΔF in the TF plane ($\Delta T > 0, \Delta F <$ or > 0).

10.4.1 Prediction according to simple superposition of frequency and temporal masking functions

We first considered a simple superposition of the frequency and temporal masking functions to predict the TF masking data. Lincoln (1998); Vafin et al. (2000) and

He and Scordilis (2008) used a similar approach (see Sec. 4.1.1). This prediction scheme, referred to as "Prediction A" below, is given by

$$AM(\Delta T, \Delta F) = AM(0, \Delta F) - (AM(0, 0) - AM(\Delta T, 0)) \tag{10.1}$$

where $AM(0, \Delta F)$ represents the "initial" spread of masking produced by the masker at the target frequency (derived from Fig. 8.3) from which is subtracted the temporal decay of forward masking over time ΔT (derived from Fig. 9.1a). Individual and mean masking patterns predicted with Prediction A for ΔT values of 5, 10, and 20 ms are depicted in Figures 10.3–10.5, respectively (solid lines). It is clear from the figures that Prediction A overestimates the amount of masking for small frequency separations ($|\Delta F| \leq 2$ ERB units) and underestimates masking for larger ΔFs. One obvious reason for the inefficiency of Prediction A is the fact that it does not take into account the ΔF dependency of the forward masking decay.

10.4.2 Prediction according to superposition of frequency masking function and level-dependent temporal masking function

An approach which takes into account the ΔF-dependency of forward masking has been employed by Huang and Chiueh (2002, see Sec. 4.1.1). In this approach (referred below to as "Prediction B"), each point of the frequency masking pattern with $F_M \neq F_T$ is considered as a hypothetical forward masker with $F_M = F_T$ but with a lower level. Since the ΔF-dependency of forward masking presents and analogy with the level dependency of forward masking, the forward masking functions for lower-level maskers are estimated by assuming that (1) the amount of masking for $\Delta T = 0$ and $\Delta F = 0$ decreases linearly with masker level (*i.e.*, the growth of masking is linear), and (2) the forward masking decay for $\Delta F = 0$ is linear on a logarithmic time scale and decays to 0-dB masking at the same point for all masker levels. The latter assumption seems to be fulfilled given the data presented in Figure 10.2. The former assumption might be fulfilled if one assumes that the condition $\Delta T = 0$ and $\Delta F = 0$ corresponds to an intensity discrimination threshold (see Sec. 8.1.3.1) which decreases linearly with masker level and, in other words, by assuming that the Weber's law holds for brief Gaussian-shaped sinusoids (Carlyon and Moore, 1984; Florentine, 1986). Prediction B has thus the form

$$AM(\Delta T, \Delta F) = AM(0, \Delta F) - a' \log(\Delta T) \tag{10.2}$$

with $a' = AM(0, \Delta F)/\log(\Delta T_{0dB})$, ΔT_{0dB} being the ΔT-axis intercept, or 0-dB masking point, at which the forward masking functions converge. Given the parameters a and b determined by fitting Equation (9.1) to the temporal masking data presented in Figure 9.1b, ΔT_{0dB} is equal to $10^{-b/a}$ (see Tab. 9.1). As for Prediction A, $AM(0, \Delta F)$ is determined from the frequency masking data. Individual and mean masking patterns predicted with Prediction B for ΔT values of 5, 10, and 20 ms are depicted in Figures 10.3–10.5, respectively (dashed lines). It can be seen that the shape of the masking patterns with Prediction B is closer to the data than that with Prediction A. Nevertheless, it is clear that Prediction B overestimates the amount of masking in almost all conditions, the overestimation being particularly large for small frequency separations ($|\Delta F| \leq$ 2 ERB units). This may be due to an incorrect estimation of the level-dependent forward masking function. The estimation of a' was indeed based on the results for $\Delta F = 0$, and assumed a linear growth of masking for $\Delta F \neq 0$. However, previous forward-masking studies reported a non linear growth of masking in some listeners for $\Delta F \neq 0$ (Munson and Gardner, 1950; Widin and Viemeister, 1979a). Another explanation might be that intensity discrimination thresholds for short Gaussian-shaped sinusoids do not vary linearly with masker level. Accordingly, Nizami et al. (2001) measured intensity discrimination thresholds as a function of level for 2-kHz Gaussian-shaped sinusoids with an ERD of 2.51 ms (for comparison, the ERD of a 2-kHz signal defined by Eq. (5.1) with $\alpha = 0.15$ would be 3.33 ms). The resulting function showed a mid-level hump at about 40 dB SL, which suggests that the growth of masking for $\Delta T = 0$ and $\Delta F = 0$ is non linear. This is further supported by our data obtained with $\Delta T = 0$ and $\Delta F = 0$ in Experiment 3 (see Sec. 8.2). The mean $\Delta I/I$ value at target threshold in Experiment 3 for $\Delta F = 0$ and $L_M = 45$ dB SL indeed showed large across-listener variability, thus a probable non linear growth of masking in some listeners.

An additional, more fundamental problem with this approach — which concerns both Predictions A and B — is the fact that the simultaneous masking pattern does not reflect the actual excitation pattern on the BM, because of suppression effects which affect both the masker- and target-induced activities (Delgutte, 1990; Yasin and Plack, 2005; Rodriguez et al., 2010). Finally, for $|\Delta F\text{s}| > 2$ ERB units and $\Delta T\text{s} > 0$, predictions and data should be compared with caution because of floor effects.

Overall, both prediction schemes failed in predicting our TF masking data. This supports the idea that collecting actual data on the spread of TF masking for a single atom is important. It further suggests that the cited audio codecs exploiting TF masking provide rather erroneous predictions of masking, and that using the TF masking function shown in Figure 11.1 might improve the efficiency of such coders.

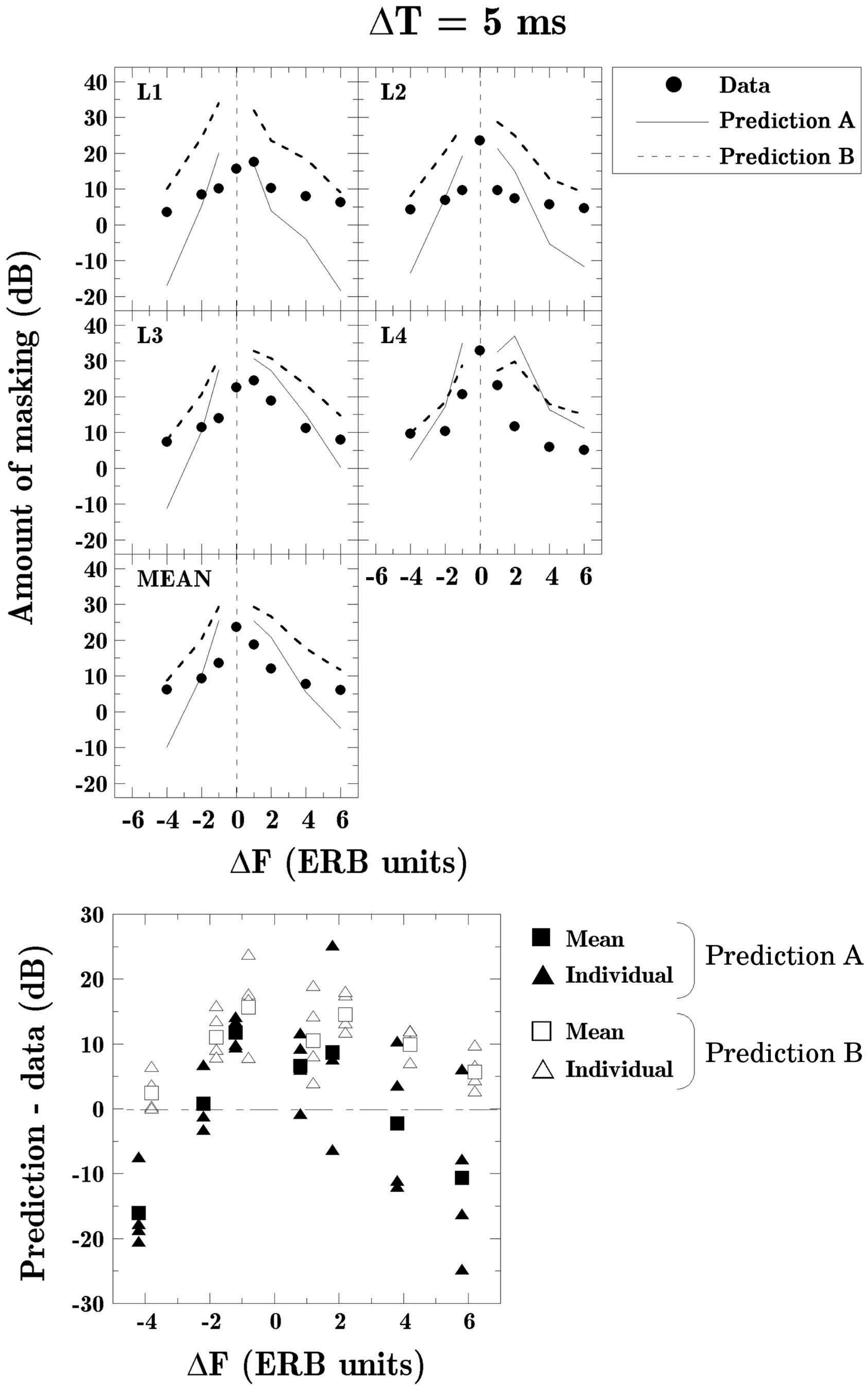


Figure 10.3: Individual and mean forward masking patterns for $\Delta T = 5$ ms predicted with Prediction A (solid lines) and Prediction B (dashed lines). The observed data are shown with filled circles. The bottom picture shows the signed prediction error of the models.

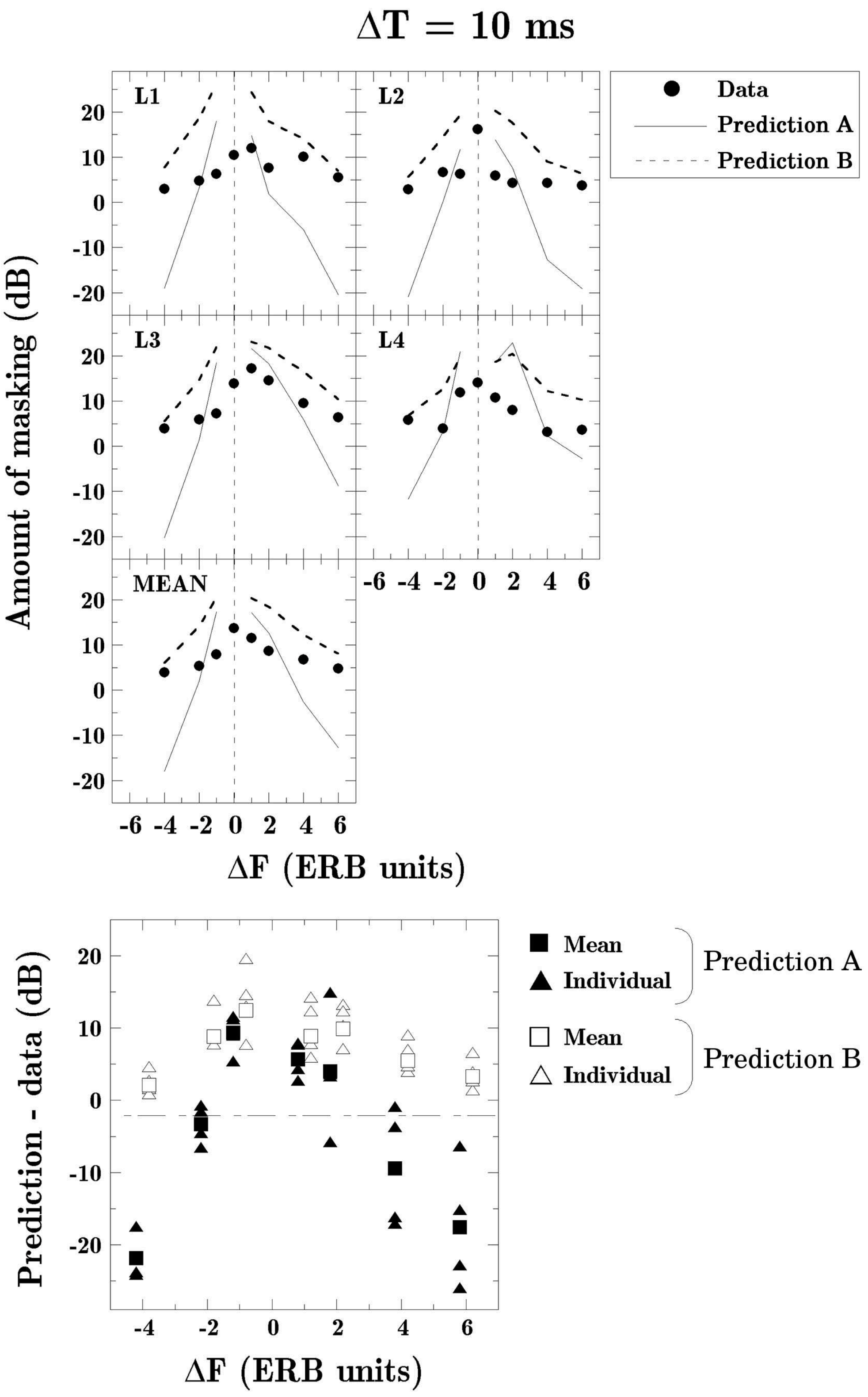


Figure 10.4: Same as Figure 10.3 but for $\Delta T = 10$ ms.

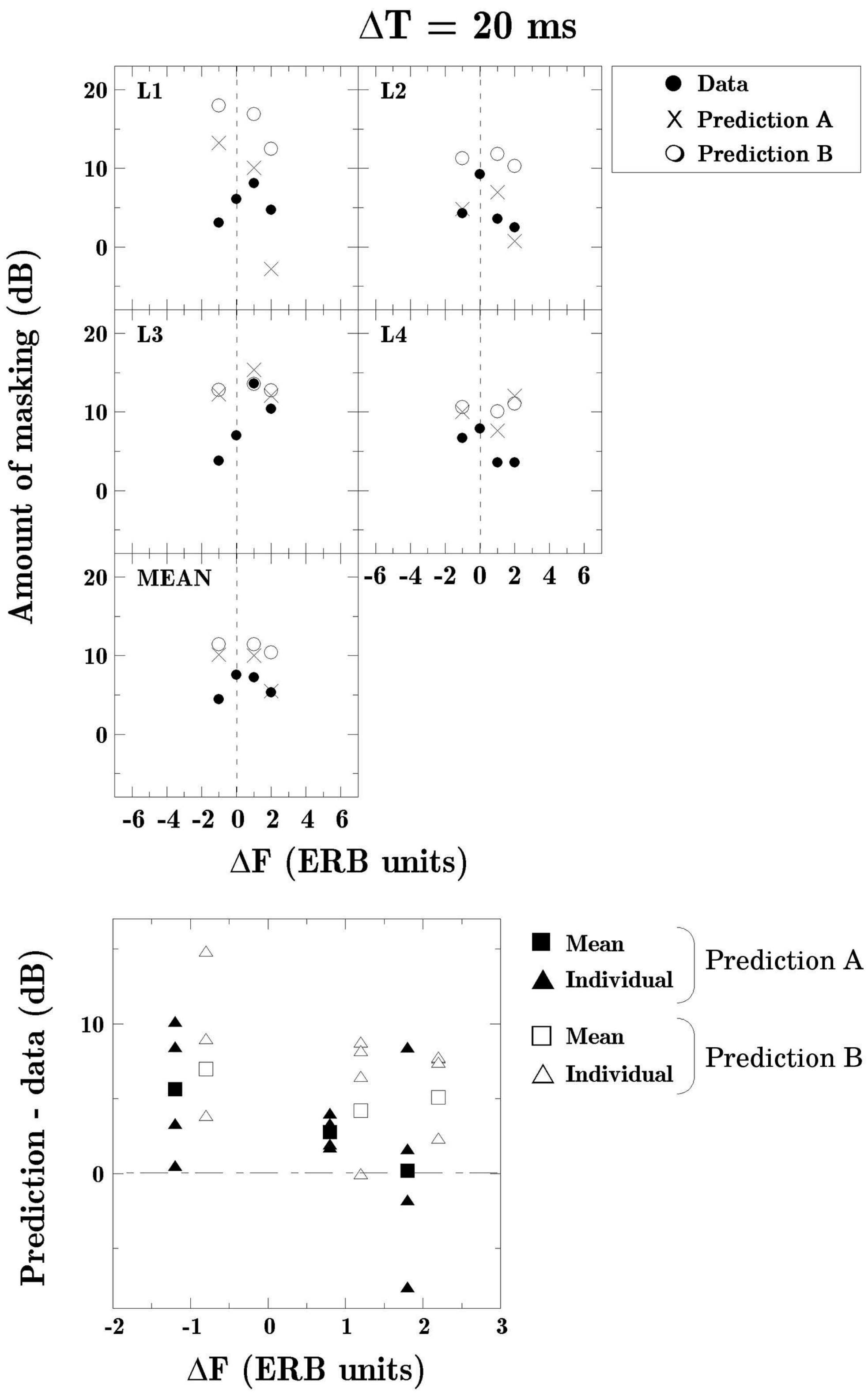


Figure 10.5: Same as Figure 10.3 but for $\Delta T = 20$ ms. Filled circles show actual data. Crosses show predictions with Prediction A. Open circles show values predicted with Prediction B.

Summary

Experiment 6 combined frequency (ΔF) and temporal (ΔT) separations between masker and target. The masker had a fixed frequency ($F_M = 4$ kHz) and level ($L_M = 60$ dB SL).

Examined as a function of frequency, the results indicated that (1) the amount of frequency masking abruptly decreases as ΔT increases, (2) forward masking patterns broaden with increasing ΔT, and (3) the masking patterns' asymmetry remains for ΔTs > 0. Examined as a function of time, the results indicated that (1) varying ΔF does not affect the linear decay characteristic of forward masking as a function of $\log(\Delta T)$, (2) the slope of this decay decreases as $|\Delta F|$ increases, and (3) forward masking decays to 0 dB at approximately the same ΔT value (which is listener-dependent) for all ΔFs. These results are in close agreement with the preceding studies that have varied both ΔF and ΔT separations with long-duration sinusoidal maskers. The results can be attributed to the temporal decay of masker-induced excitation pattern on the BM and in more central auditory processing stages. The temporal overlap of the BM responses to masker and target may have also contributed to the results for small ΔT and ΔF values. We attempted to predict the TF masking data by assuming a linear combination of temporal and frequency masking results. Such a simplistic approach has been used in some perceptual audio coding algorithms to account for TF masking. We assumed that the spread of TF masking caused by the masker could be described by the frequency masking pattern (measured in Exp. 2) combined with the decay of forward masking (measured in Exp. 5) from each point of the masking pattern. Two prediction schemes were tested that both failed in predicting the TF masking data from Experiment 6. The inefficiency of these two predictors indicates that combining temporal and frequency masking data measured separately does not provide an accurate representation of TF masking results. It is therefore important to collect actual TF masking data.

Chapter 11

Towards the modeling step: Summary of psychoacoustical data on time-frequency masking for a Gaussian masker

The three-dimension plot in Figure 11.1 summarizes the results from Experiment 6 in the TF domain. To account for the exponential decay of masker-induced activity over time in the cochlea and in the auditory nerve (see Sec. 10.3), for ΔFs between -4 and +6 ERB units, the decay of forward masking was modeled with an exponential function of the form $AM = C(\Delta F)e^{-\Delta T/\lambda(\Delta F)}$. Parameter C specifies the "initial spread of masking" (*i.e.*, at $\Delta T = 0$) produced by the masker at the target frequency, and λ is a time constant that characterizes the temporal decay of forward masking. The values of parameters C and λ estimated for each ΔF are listed in Table 11.1. To provide a "complete" representation of TF masking (*i.e.*, that reaches 0 dB of masking), the data for ΔFs below -4 and above +6 ERB units were extrapolated based on a two-dimensional cubic spline fit along the TF plane.

Note that the masking function in Figure 11.1 resembles that previously proposed by Huang and Chiueh (2002) (see Eq. (4.4)) to account for forward masking in perceptual audio codecs. Their model included a frequency-dependent time constant (parameter $\tau(z)$ in Eq. (4.4)). However, the values of the model parameters in Huang and Chiueh were chosen based on psychoacoustical data measured separately in the time and frequency domains (see Sec. 4.1.3). Conversely, the present TF masking function is issued from actual TF masking data. A model of TF masking is designed below based on the values of parameters C and λ. This model might be used to improve the efficiency of perceptual audio coding algorithms such as those presented in Chapter 4.

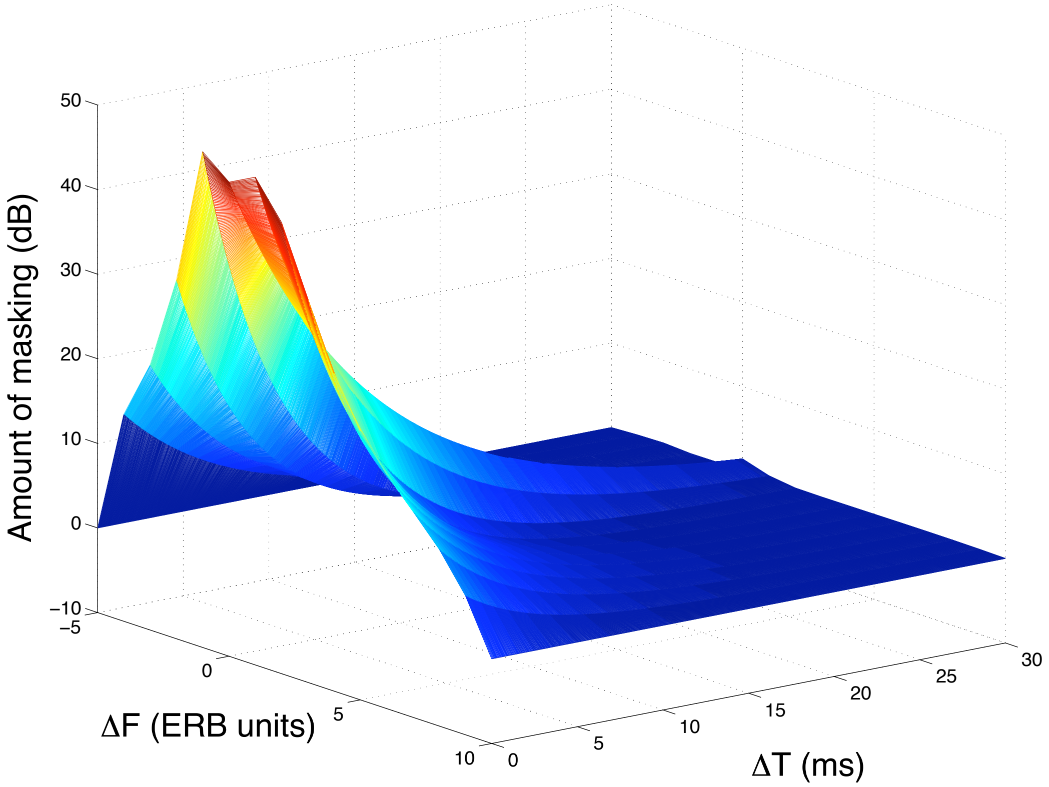


Figure 11.1: Mean amount of masking (in dB) produced by the Gaussian masker as a function of ΔT (in ms) and ΔF (in ERB units). The time axis was sampled at 44.1 kHz.

	Exponential fit		
ΔF (ERB units)	C (dB)	λ	r^2
-4	14.45	6.69	0.99
-2	32.37	4.46	0.99
-1	48.60	4.49	0.99
0	46.03	8.67	0.99
+1	47.65	6.56	0.98
+2	43.30	4.77	0.97
+4	27.54	4.93	0.95
+6	17.60	5.77	0.96

Table 11.1: Values of parameters C (in dB) and λ determined by fitting the equation $AM = Ce^{-\Delta T/\lambda}$ to the results from Experiment 6 for ΔFs between -4 and +6 ERB units. The last column indicates values of r^2.

Part III

MODELING OF EXPERIMENTAL DATA

Contents of the Third Part

Chapter 12

Design of a time-frequency masking model

Contents

In Section 4.1.1, the TF masking models that are currently implemented in perceptual audio codecs were presented. These models were shown to have several limitations to accurately predict the masking effects between TF atoms (see Secs. 4.1.3 and 10.4). On the purpose of overcoming the limitations of these models, in Experiment 6, we gathered masking data for stimuli with maximal concentration in the TF plane in various time and frequency configurations (see Chap. 10). Based on these masking data representing the basic spread of TF masking evoked by an atom (see Figs. 4.4b and 11.1), the idea was to develop a new TF masking model. To achieve this goal, two different approaches are possible, namely (1) to design an explicit, mathematical TF masking function from the experimental data or (2) to discretize the TF masking pattern so as to get a TF masking *kernel*. We attempted to model the experimental data using both approaches. We insist on the fact that the models presented below solely aim to quantitatively describe the masking effects in the TF domain and *do not intend* to model nor account for the non linear auditory process.

12.1 Design of an explicit time-frequency masking function

12.1.1 Exponential function

To develop a TF masking model based on the data gathered in Experiment 6, a straightforward approach was to use the exponential function proposed in Chapter 11

$$AM(\Delta T, \Delta F) \;=\; C(\Delta F)\, e^{-\Delta T/\lambda(\Delta F)} \tag{12.1}$$

with

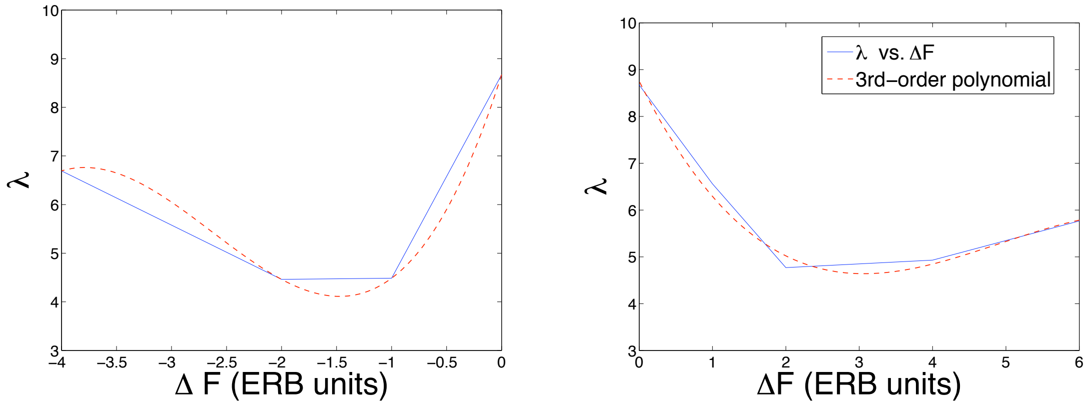
$$C(\Delta F) = \begin{cases} +11.1\,\Delta F + 58.0 & \text{if} \quad \Delta F < 0 \\ -6.4\,\Delta F + 55.4 & \text{if} \quad \Delta F \geq 0 \end{cases} \tag{12.2}$$

and

$$\lambda(\Delta F) = \begin{cases} +0.43\,\Delta F^3 + 3.4\,\Delta F^2 + 7.1\,\Delta F + 8.7 & \text{if} \quad \Delta F < 0 \\ -0.05\,\Delta F^3 + 0.75\,\Delta F^2 - 3.2\,\Delta F + 8.8 & \text{if} \quad \Delta F \geq 0 \end{cases} \tag{12.3}$$

$AM(\Delta T, \Delta F)$ is the amount of masking (in dB) produced by the masker on a target separated from the masker by ΔT and ΔF in the TF plane. The ΔF-dependencies of parameters C and λ were determined based on the values of C and λ estimated for ΔFs between -4 and +6 ERB units in Experiment 6 (see Tab. 11.1). Because parameter C predicts the amount of masking for $\Delta T = 0$, we used as $C(\Delta F)$ the linear regressions computed for each side of the masking pattern with $\Delta T = 0$ in Experiment 6 (see Tab. 10.1).

The ΔF-dependency of the time constant λ was approximated based on a third-order polynomial fit of the data, as shown in Figure 12.1. Because $\lambda(\Delta F)$ could not be fitted with a single function over the whole range of ΔFs, we provide two functions (see Eq. (12.3)).



(a) (b)

Figure 12.1: Third-order polynomial fit (dashed lines) of λ *vs.* ΔF (straight lines) for (a) $\Delta F < 0$ ($r^2 = 1.0$) and (b) $\Delta F \geq 0$ ($r^2 = 0.99$). Polynomial coefficients are specified in Equation (12.3).

Then, we assessed whether the results from Experiment 6 could be predicted by the exponential model above. Specifically, we attempted to predict the mean forward masking patterns in Figure 10.1 for $\Delta T = 0$, 5, 10 and 20 ms. The results (**a**) and the signed prediction error (**b**) are plotted in Figure 12.2. Regardless of ΔT, when $\Delta F = 0$, the model overestimates masking (up to 9 dB for $\Delta T = 0$). This is a consequence of the fact that the prediction at $\Delta T = 0$ and $\Delta F = 0$ corresponds to the offset of simultaneous masking in Equation (12.2) and does not account or the dips observed in the simultaneous masking patterns (see Fig. 8.3). Otherwise, the model is satisfactory ($|\text{error}| \leq 5$ dB). Nevertheless, at $\Delta T = 20$ ms, predictions and data should be compared with caution because of floor effects.

Overall, the exponential model of TF masking in Equation (12.1) provided a good fit to the experimental data. Therefore, this model might be used to improve the

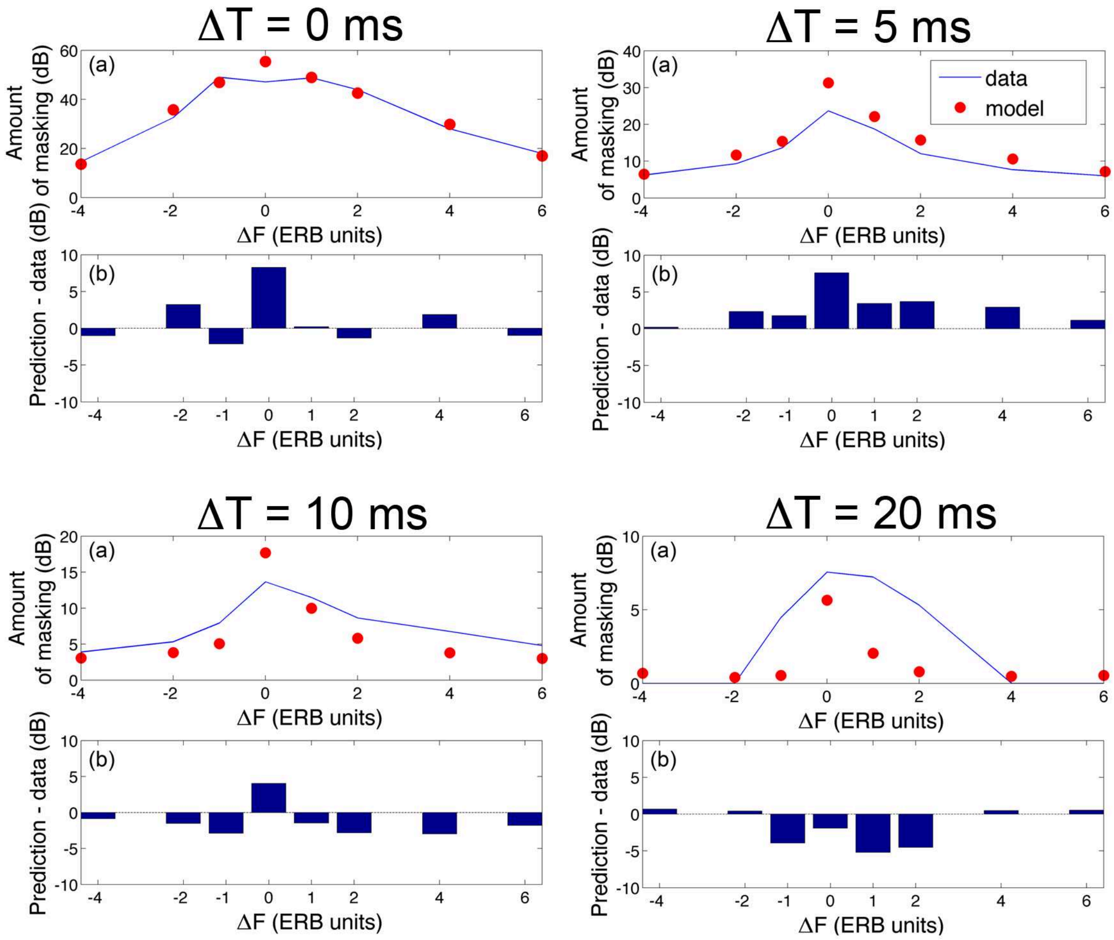


Figure 12.2: (**a**) Predictions (●) of the mean data from Experiment 6 (straight lines) by the exponential model of TF masking in Equation (12.1), for $\Delta T = 0$, 5, 10 and 20 ms. (**b**) Signed prediction error of the model.

efficiency of perceptual audio coding algorithms such as those described in Chapter 4. However, because the ΔT and ΔF dependencies of the amount of masking lie on a set of five equations, the model's implementation in a sound signal processing algorithm might be tricky. We therefore attempted to design a second model with a simpler formulation for $AM(\Delta T, \Delta F)$.

12.1.2 Conic function

To provide a simpler formulation for the TF masking model, we made a second modeling attempt by considering the polar coordinates associated with the Cartesian plane $(\Delta T, \Delta F)$. The mapping from Cartesian to polar coordinates is

$$\begin{aligned} r &= \sqrt{\Delta T^2 + \Delta F^2} \\ \Theta &= \arctan\left(\frac{\Delta F}{\Delta T}\right) \end{aligned}$$

Making an analogy with the currently implemented spreading function of masking in audio codecs (see Eq. (4.1)), the basic idea here was to develop a two-dimensional function $AM(r,\Theta)$ capable of predicting the amount of masking for any direction in the TF plane. We therefore attempted to match the TF masking data to a conic function with a maximum at $(\Delta T = 0, \Delta F = 0)$ and an exponential decay of masking such that

$$AM(r,\Theta) \; = \; C_0 \, e^{-\alpha_\Theta r} \tag{12.4}$$

where C_0 is a constant characterizing the amount of masking at (0,0) (*i.e.*, the top of the cone) and α_Θ is a damping factor characterizing the decay of masking in the TF domain. Figure 12.3 illustrates the concept behind the conic function. Considering all directions Θ ($\Theta \in [-\frac{\pi}{2}; \frac{\pi}{2}]$) around the point (0,0) and estimating the exponential masking decay in each direction, one should get the desired cone. Note that the estimation of such a conic function requires that the greatest amount of masking be obtained at (0,0), which was not the case in practice. To fulfill this requirement, and thus properly estimate the conic function, the mean amount of masking obtained at $\Delta T = 0$ and $\Delta F = 0$ in Experiment 6 (≈ 47 dB) was set to the value predicted by the intersection of the slopes of the simultaneous masking pattern (≈ 60 dB). Accordingly, the value of parameter C_0 in Equation (12.4) was set to 60. Then, the damping function α_Θ remained to be estimated. This was achieved by varying Θ from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ in 0.1-radian steps, and estimating α for each Θ. In a preliminary step, the initial set of experimental data (blue circles in Fig. 12.3) was interpolated based on a two-dimensional cubic spline method in order to get more "masking samples" in the TF domain. Figure 12.4 shows the exponential fit of AM *vs.* r for Θ values of $\pm\frac{\pi}{2}$, $\pm\frac{\pi}{4}$, and 0. It can be seen that the exponential fit was successful in all directions ($r^2 \geq 0.97$). The dependence of α upon Θ is illustrated in Figure 12.5. The function α_Θ (filled circles) can be well approximated by a 8th degree polynomial (dashed line) whose expression is specified in the figure. Given the function α_Θ, the conic function in (12.4) could be built. This function is plotted in Figure 12.6 in the Cartesian plane $(\Delta T, \Delta F)$.

As for the exponential model, we finally verified whether the results from Experiment 6 could be predicted by the conic function. The results (**a**) and the signed prediction error (**b**) are plotted in Figure 12.7. For $\Delta T = 0$, the shape of the masking pattern predicted by the conic function is close to the data, but the amounts of masking for all ΔFs $\neq 0$ are underestimated. This might be due to the very small deviation between α *vs.* Θ and the 8th degree polynomial. Even if residuals were smaller than 0.1, an underestimated value of α can greatly affect the exponential decay of masking. Also, note that the two-dimensional cubic spline interpolation applied to the data prior to the determination of α_Θ may have biased the estimation. We attempted to apply the exponential fit to the actual data only, but this provided poor results, especially at large $\Delta T/\Delta F$ configurations for which a few number of data points were measured. Consequently, we conserved the interpolation.

The overestimation at $\Delta F = 0$ is most likely due to the constraint on the maximum of the function at (0,0). For $\Delta T = 5$ and 10 ms, the model over predicts the amount of masking in almost all conditions. This can be thought of as a straight consequence of the bad predictions at $\Delta T = 0$ which involve the overestimation of masking at (0,0), combined with a possibly underestimated damping factor α_Θ. For $\Delta T = 20$ ms, the model predicts little or no masking for all ΔFs but, again, predictions and data should be compared with caution in this condition because of

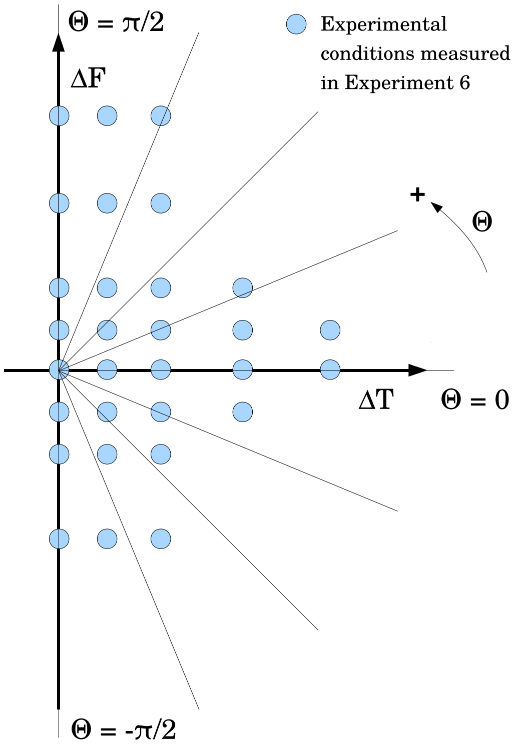


Figure 12.3: Schematic representation of the polar grid associated with the Cartesian plane $(\Delta T, \Delta F)$ and defined by the rotation of angle Θ ($\Theta \in [-\frac{\pi}{2}; \frac{\pi}{2}]$) around the point (0,0). $\Theta = \{-\frac{\pi}{2}; \frac{\pi}{2}\}$ corresponds to frequency masking conditions. $\Theta = 0$ corresponds to forward masking conditions. The blue circles represent experimental conditions measured in Experiment 6.

floor effects.

Overall, the present conic function was not able to accurately predict the TF masking data from Experiment 6. Although it has the advantage over the exponential function above to provide a simpler masking function in the TF domain, the conic function *in its present form* is not eligible for implementation in a sound signal processing algorithm.

12.2 Design of a time-frequency masking kernel

Finally, a more pragmatic approach to model the experimental data consisted in discretizing the TF masking pattern in Figure 11.1 according to the TF analysis-synthesis scheme chosen for signal decomposition. The so-called discrete pattern could then be used as a masking *kernel* in the TF plane. This is the modeling approach we actually opted for implementation in a sound signal processing tool. It is treated in Chapter 13.

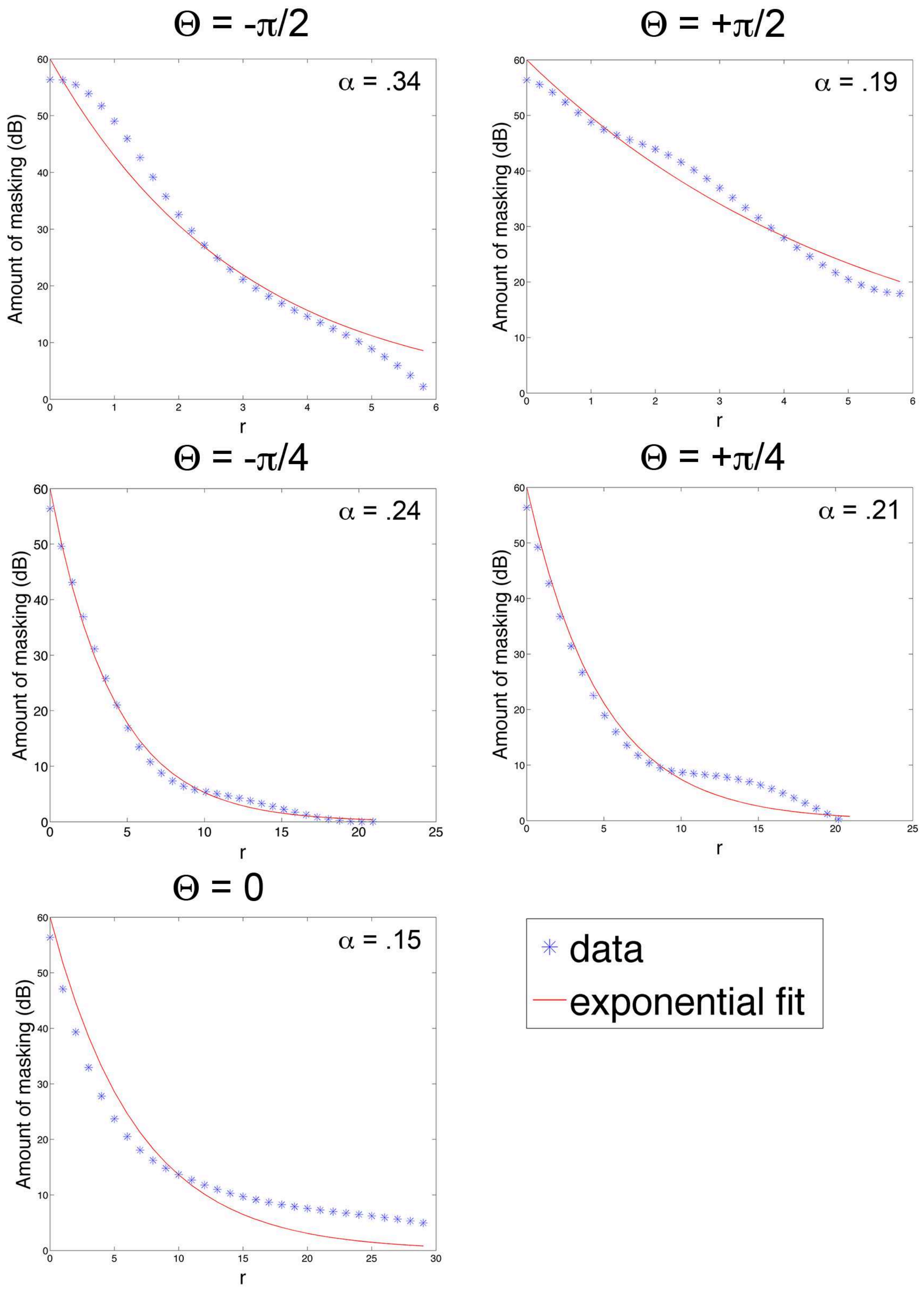


Figure 12.4: Exponential fit of $AM\, vs.\, r$ by the function $60\, e^{-\alpha_\Theta r}$ at five Θ values.

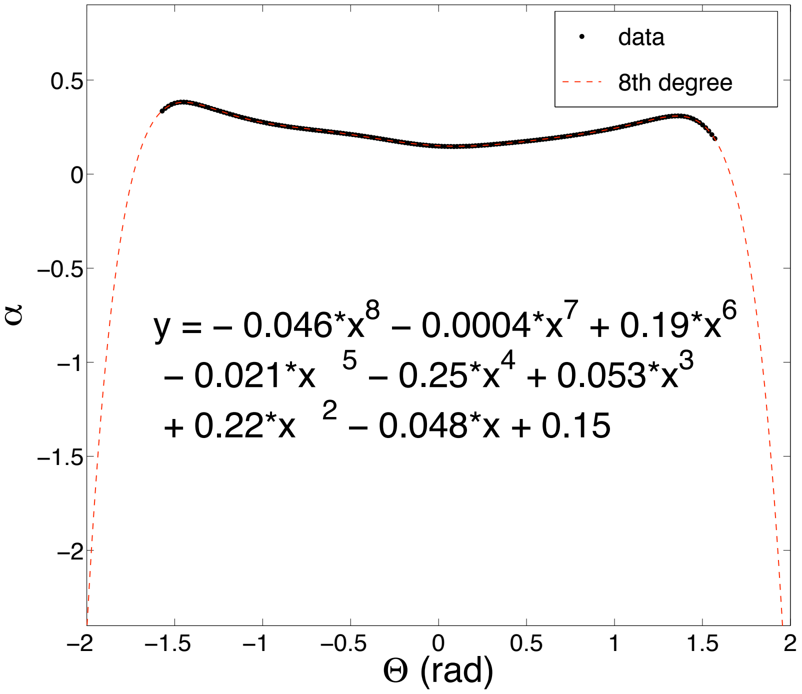


Figure 12.5: 8th degree polynomial fit (dashed line) of α *vs.* Θ (●, in radians). The polynomial coefficients are specified in the figure. Residuals were < 0.1.

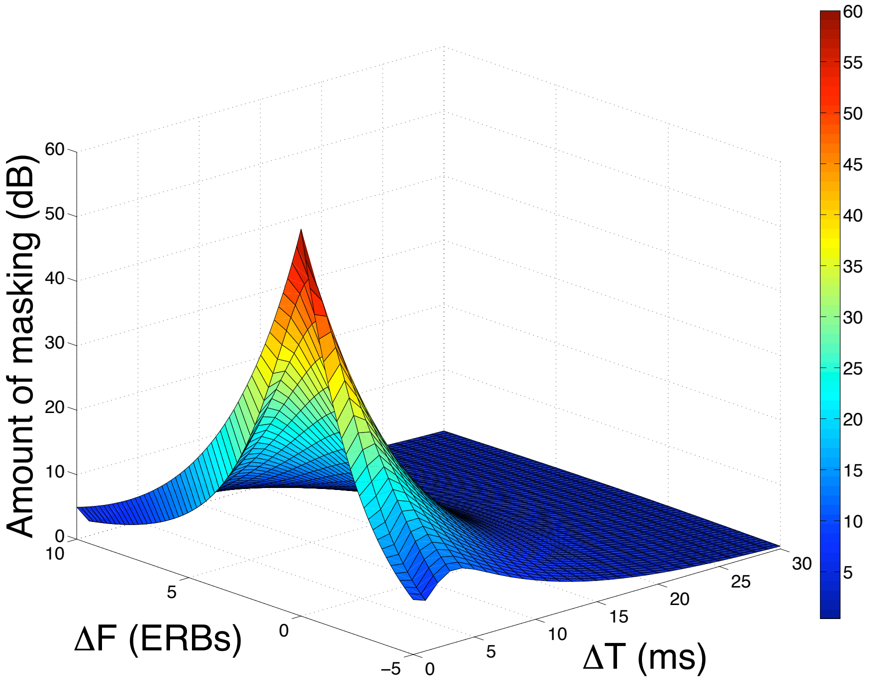


Figure 12.6: Representation of the conic function of masking in (12.4) in the Cartesian plane $(\Delta T, \Delta F)$. $C_0 = 60$ and α_Θ is the 8th degree polynomial in Fig. 12.5.

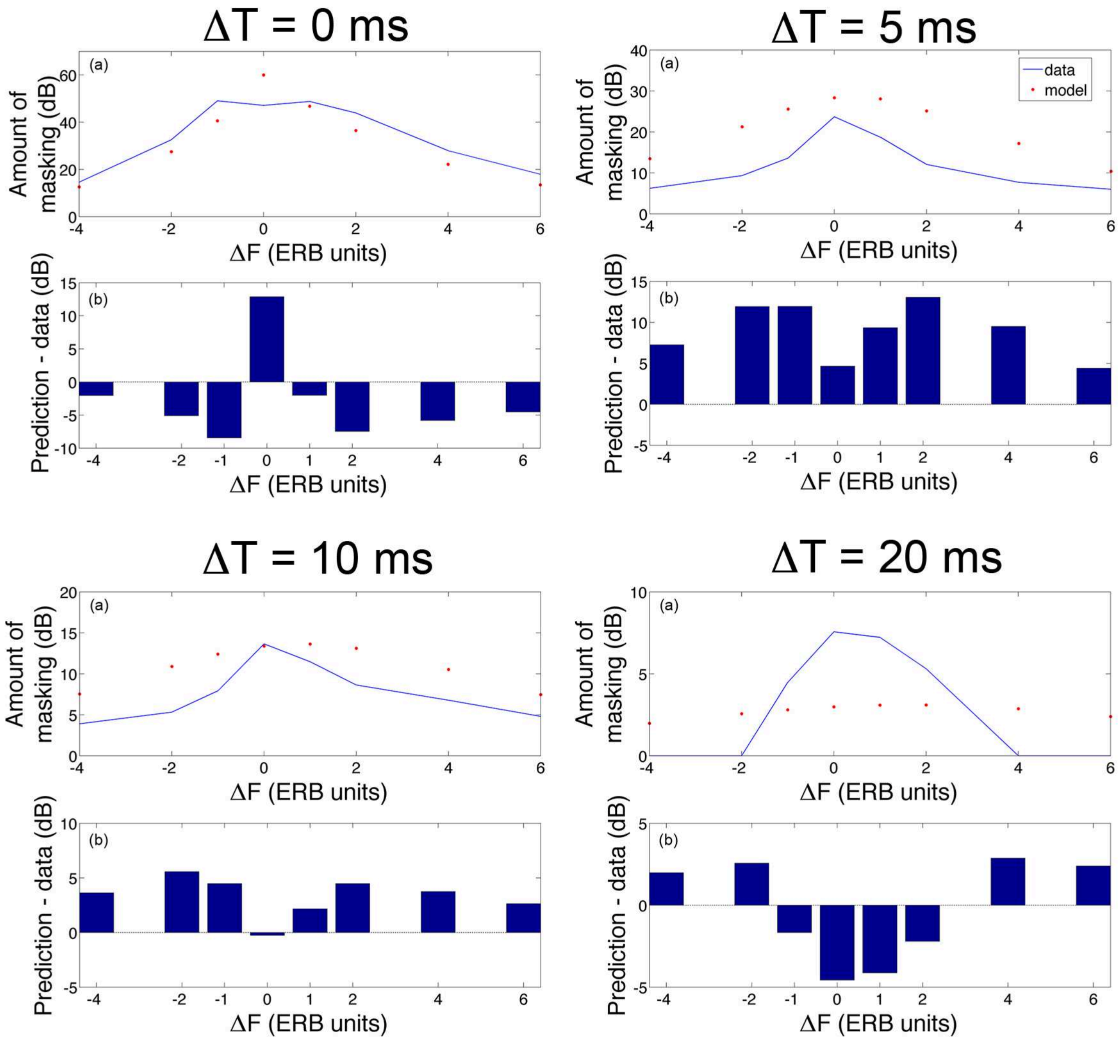


Figure 12.7: (**a**) Predictions (●) of the mean data from Experiment 6 (straight lines) by the conic model of TF masking in Equation (12.1), for $\Delta T = 0$, 5, 10 and 20 ms. (**b**) Signed prediction error of the model.

Summary

Two modeling attempts were made to develop an analytic TF masking function accounting for the TF masking data gathered in Experiment 6.

First, an exponential function of the form $AM(\Delta T, \Delta F) = C(\Delta F)\, e^{-\Delta T/\lambda(\Delta F)}$ was provided. The function $\lambda(\Delta F)$ was approximated by two 3rd order polynomials, while $C(\Delta F)$ was approximated by two linear regressions. This model could globally predict ($|\text{error}| < 5$ dB) the experimental data for ΔFs $\neq 0$ and ΔTs up to 10 ms. It is therefore eligible for implementation in a sound signal processing tool.

Second, in an attempt to provide a simpler masking function (*i.e.*, to simplify the ΔF dependencies of C and λ), we designed a conic function with an exponential radial decay of masking of the form $AM(r, \Theta) = C_0\, e^{-\alpha_\Theta r}$ where (r, Θ) are the polar coordinates associated with the Cartesian plane $(\Delta T, \Delta F)$, C_0 is a constant, and α_Θ is a damping factor. The latter was approximated by a 8th order polynomial. However, this model was not able to predict the experimental data, and is not eligible for implementation in a signal processing algorithm.

Finally, we opted for a more pragmatic approach to model the experimental data, namely to discretize the TF masking pattern according to the TF analysis-synthesis scheme so as to use it as a masking *kernel* in the TF domain. This is treated below.

Chapter 13

Implementation of the model in a sound signal processing tool

Contents

This chapter presents preliminary investigations for implementing the TF masking data in a sound signal processing algorithm. Precisely, two algorithms are presented, which aimed at matching the signal representation to human auditory perception by removing the perceptually irrelevant components in the signal decomposition. Because of time constraints, the two implementation attempts described below could not be formally evaluated (*i.e.*, by conducting listening tests with real-world sounds). Therefore, we merely evaluate the performance of each algorithm based on preliminary results obtained with simple sounds (such as the Gaussian stimuli used in psychoacoustical experiments) and informal listening by the author.

Prior to the description of the two proposed algorithms, the TF analysis-synthesis scheme chosen for signal decomposition and reconstruction is detailed.

13.1 Choice of the signal representation: Wavelet transform

In Section 2.3, we showed that the frequency analysis performed by the human auditory system can be approximated by a bank of bandpass filters with a constant relative bandwidth. Each of these filters is associated with a frequency channel named *auditory filter*, or *critical band* (CB), and is characterized by its equivalent rectangular bandwidth (ERB). Based on this concept, the ERB scale was defined (see Eq. (2.11)) which allows to plot signals or psychoacoustical data on a frequency scale related to human auditory perception.

In Section 8.3, the results from Experiment 4 revealed that the masking patterns for Gaussian maskers with $F_M = 4$ kHz ($ERB_{GW} = 600$ Hz) and $F_M = 0.75$ kHz ($ERB_{GW} = 112.5$ Hz) have roughly similar shapes and bandwidths when plotted on an ERB scale (see Fig. 8.11). This is compatible with the constant-Q frequency analysis of the auditory system.

In Chapter 1, we presented two analysis methods for representing signals in the TF domain, namely the Gabor and wavelet transforms. While the Gabor transform can be interpreted as a bank of bandpass filters with fixed bandwidth, the wavelet transform can be thought of as constant-Q analysis (*i.e.*, as a bank of bandpass filters with constant relative bandwidth), analog to the frequency analysis of the human auditory system. For that reason, most of audio applications use the wavelet transform as the TF analysis-synthesis scheme. The wavelet transform was also elected to the TF analysis-synthesis scheme implemented in the signal processing algorithms presented below. This choice represents one of the possible applications of TF masking data to sound signal processing. In future works, we do not exclude to generalize the present work to other TF representations such as of the Gabor transform.

13.1.1 Discretization of the continuous wavelet transform

The general properties of the continuous wavelet transform (CWT) were introduced in Section 1.3. We specify below the properties of the basis functions used for signal decomposition in the present study.

The CWT results from the decomposition of a signal into a family of functions, which are scaled versions of a prototype function $g(t)$ according to Equation (1.14). The scale factor a defines the time and frequency resolution in the TF domain such that dilated functions ($a > 1$) "analyze" low frequencies, while compressed functions ($a < 1$) "analyze" high frequencies. In the present study, we only use scale factors $a \geq 1$ (this is discussed below). In practice, the scale factor a must be discretized so as to form a discrete time-scale grid. We opted for independent time and scale parameters. We chose the following discretization $(a_j, b_k) = (a_0^j, kb_0)$ where $a_0 > 1$ and $b_0 \in \mathbb{R}$ are two constants defining the size of the sampling grid. The family of wavelets is thus given by

$$g_{j,k}(t) = \frac{1}{a_0^j} g\left(\frac{t - kb_0}{a_0^j}\right) \quad (j \in \mathbb{Z}, k \in [0, \dots, N-1]) \tag{13.1}$$

where N is the signal length, in samples. To conserve all time samples of the signal, the constant b_0 was chosen as the sampling period $T_S = 1/F_S$. The corresponding sampling grid is illustrated in Figure 13.1.

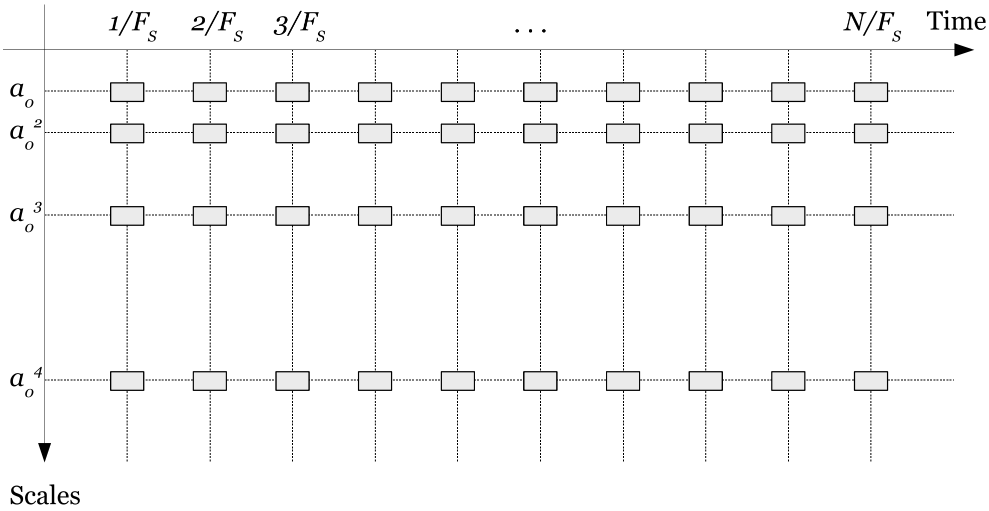


Figure 13.1: Illustration of the sampling grid defined by the wavelets family in Eq. (13.1).

Furthermore, we opted for a discretization of the scale in voices and octaves such that

$$a_0^j = 2^{\frac{m}{\mathcal{D}_v}+n} = a_{m,n}$$

where $m \in [0, \ldots, \mathcal{D}_v - 1]$, $n \in [0, \ldots, \mathcal{D}_o - 1]$ and $j \in [0, \ldots, \mathcal{D}_v\mathcal{D}_o - 1]$, $\mathcal{D}_v$ and $\mathcal{D}_o$ being the number of voices and octaves, respectively. This discretization yields an increasing scale factor step $a_0 = 2^{\frac{1}{\mathcal{D}_v}}$. Moreover, it provides two parameters, $\mathcal{D}_v$ and $\mathcal{D}_o$, for determining the total number of scales in the representation. Adding an octave amounts to adding $\mathcal{D}_v$ scales in the representation. Adding one voice amounts to adding $\mathcal{D}_o$ scales, *i.e.*, one voice in each octave. Hence, since the mother wavelet is defined for $a = 1$ and that large scales correspond to low frequencies, adding an octave amounts to adding details in the representation of low frequencies (see Fig. 13.2).

13.1.2 Choice of the mother wavelet

The family of basis functions were defined in the frequency domain. Specifically, we designed a set of wavelets with the aim to cover the whole spectrum of audible frequencies: from about 20 Hz to 20 kHz (see Fig. 3.1). Because we only used scale factors $a \geq 1$, the frequency of the mother wavelet $\hat{g}(\omega)$ defined the highest frequency in the signal representation. Overall, we were searching for a set of functions $\hat{g}_{j,k}(\omega)$ verifying (see Sec. 1.3)

(1) $\hat{g}(\omega) = 0 \,\forall\, \omega \in \mathbb{R}^-$ (*i.e.*, real-valued or complex analytic function)

(2) $\hat{g}(\omega) \neq 0 \,\forall\, \omega \in]0, \frac{F_S}{2}]$ (compact support, finite energy)

(3) $\hat{g}(0) = 0$ (function of zero mean)

Moreover, the accurate prediction of masking in the time-scale domain requires that the spectro-temporal characteristics of the wavelets match the spectro-temporal characteristics of the masker used in psychoacoustical experiments (see Sec. 4.2). Thus, a straightforward option for the mother wavelet $\hat{g}(\omega)$ was to use a Gaussian-shaped sinusoid similar to that defined in Equation (5.1) and whose Fourier transform is

$$\hat{s}(\omega) \;=\; \frac{1}{2j\sqrt{\Gamma}}\left[e^{-\pi\left(\frac{\omega-\omega_0}{\Gamma}\right)^2} - e^{-\pi\left(\frac{\omega+\omega_0}{\Gamma}\right)^2}\right]$$

This function is a bandpass filter centered at ω_0 (*i.e.*, it oscillates in time like a wave) of finite energy, but is not analytic. Therefore, we defined the following function for the mother wavelet

$$\hat{g}(\omega) \;=\; \frac{1}{2j\sqrt{\Gamma}}\, e^{-\pi\left(\frac{\omega-\omega_0}{\Gamma}\right)^2} \tag{13.2}$$

where $\Gamma = \alpha f_0 = \alpha\frac{\omega_0}{2\pi}$, α being the shape factor of the Gaussian window (see Chap. 5). To be consistent with the signals used in psychoacoustical experiments, an α value of 0.15 was used for the wavelets. Note that this α value also characterizes the quality factor of the wavelet analysis.

Because the Gaussian window has an infinite support, the function (13.2) does not fulfill the constraints (2) and (3) stated above. In practice, however, the Gaussian with $\alpha = 0.15$ has a sufficiently rapid decay and quickly tends towards 0 so as to be considered as an acceptable function for $\hat{g}(\omega)$. The steep decay of the Gaussian window can also become a drawback. Indeed, in the numerical implementation, the steep decay of $\hat{g}(\omega)$ engendered a numerical error in the reconstruction at very low frequencies (0–40 Hz, this is discussed below).

Finally, we fixed the pulsation of the mother wavelet to $\omega_0 = \frac{3\omega_S}{8}$ where $\omega_S = 2\pi F_S$. At a sampling rate of 44.1 kHz and $\alpha = 0.15$, the highest-frequency analysis filter had a center frequency $f_0 = 16.5$ kHz, and a bandpass of about 2.5 kHz. Increasing the scale factor a_j thus constituted a bank of bandpass filters with center pulsations $\omega_j = \frac{\omega_0}{a_j}$ and a constant $Q = 0.15$. Figure 13.2 shows a set of analysis windows $\hat{g}_{a_j}(\omega)$ obtained with $\mathcal{D}_o = 3$ (as represented by the three colors) and $\mathcal{D}_v = 5$. In this example, the analysis "starts" around 2 kHz (as defined by the lowest-frequency analysis filter). To analyze frequencies below 2 kHz, $\mathcal{D}_o$ must be increased.

13.1.3 Numerical implementation of wavelets

The wavelet transform is computed in the frequency domain according to Equation (1.16). Given a mother wavelet $\hat{g}(\omega) \in L^2(\mathbb{C})$, for a fixed value $a_j, j \in \mathbb{Z}$ of the scale factor, the wavelet transform of a sampled signal $s(k) \in L^2(\mathbb{R})$ is

$$S_g(a_j, b) = \mathcal{F}^{-1}\left[\hat{s}(\omega)\sqrt{a_j}\hat{g}(a_j\omega)\right] \tag{13.3}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform (see Eq. (1.3)).

Since we use analytic wavelets, the reconstruction of signal $s(k)$ from its wavelet coefficients $S_g(a_j, b)$ is achieved by (see Eq. (1.19))

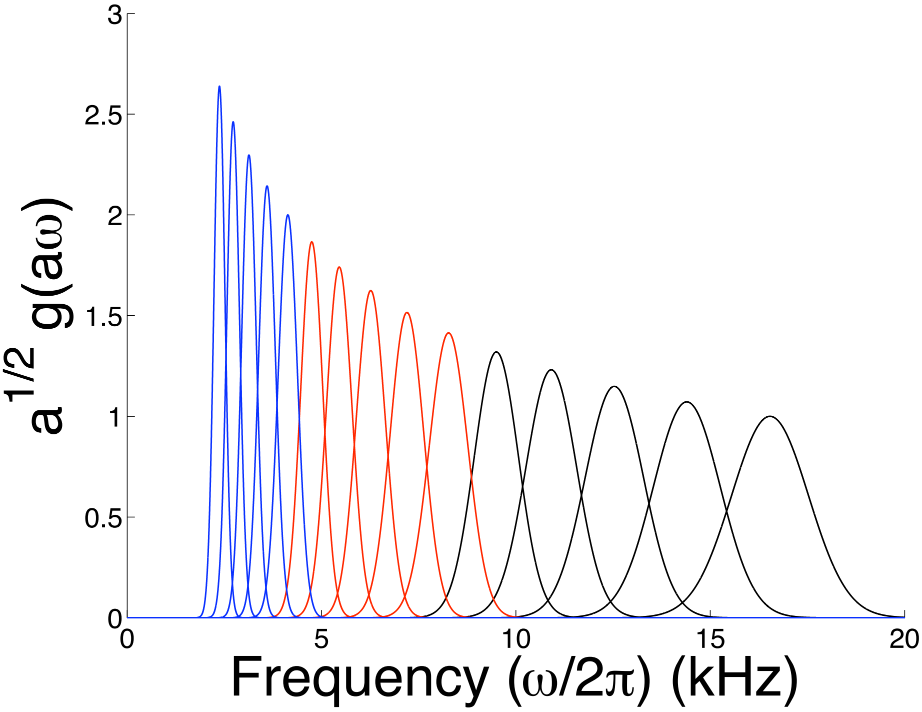


Figure 13.2: A set of 15 Gaussian analysis windows $\hat{g}_{a_j}(\omega)$ (defined in Eq. (13.2)) for three octaves ($\mathcal{D}_o = 3$) and five voices per octave ($\mathcal{D}_v = 5$) is plotted as a function of frequency. Each color represents an octave.

$$s(k) \;=\; 2\Re\left\{ c \sum_{j\,\in\,\mathbb{Z}} \sum_{b\,\in\,\mathbb{Z}} S_g(a_j, k)\, g_{a_j,b}(k) \right\} \tag{13.4}$$

where c is a normalization constant that ensures the conservation of signal energy and is equal to

$$c \;=\; \frac{1}{\sum\limits_{j\,\in\,\mathbb{Z}} |\hat{g}(a_j\omega)|^2} \tag{13.5}$$

Thus, from Equations (13.4) and (13.5) comes the synthesis formula

$$s(k) = 2\Re\left\{ \sum_{j\in\mathbb{Z}} \mathcal{F}^{-1}\left[\frac{\widehat{S_g}(a_j,\omega)\sqrt{a_j}\hat{g}(a_j\omega)}{\sum\limits_{j\in\mathbb{Z}} |\hat{g}(a_j\omega)|^2} \right] \right\} \tag{13.6}$$

Note that the use of a Gaussian window with a shape factor of 0.15 as $\hat{g}_{a_j}(\omega)$ engendered a numerical error in the reconstruction. Specifically, the term $\sum_j |\hat{g}(a_j\omega)|^2$, which represents the energy of the wavelets family, featured sample values as low as 10^{-120} for $\omega \approx 0$. This is due to the drastic decay of the Gaussian windows. Consequently, the normalization term c diverged numerically and rendered the signal reconstruction impossible. This was fixed by reconstructing the signal only for frequencies above 40 Hz or, in other words, by high-pass filtering the signal with a cut-off frequency of 40 Hz.

To illustrate the implemented analysis-synthesis scheme, Figure 13.3 shows the scalogram (top) of a wood impact sound synthesized in Aramaki and Kronland-Martinet (2006). Shortly, the synthesis model considers impact sounds as a combination of exponentially damped sinusoids, as represented by the horizontal stripes in Figure 13.3(a). The analysis was made with $\mathcal{D}_o = 9$ and $\mathcal{D}_v = 12$. The residual (output - input) is shown in Figure 13.3(b). It clearly reveals an error ($\approx 10^{-3}$) in the sound attack. This error is due to the fact that we reconstruct the signal only for frequencies above 40 Hz. Accordingly, the spectrum of the error (Fig. 13.3(c)) shows that the error is concentrated in the range 0–50 Hz. For frequencies > 500 Hz (not shown in the plot), the error is $< 10^{-15}$. Informal listening tests indicated no perceptual difference between the re-synthesized and original signals.

13.2 Matching the signal representation to human auditory perception

We propose below two algorithms for removing the perceptually irrelevant components in the wavelet transform of a sound signal while causing no audible difference to the original sound after resynthesis. The general idea of the approaches was to (1) estimate a global masking threshold in the time-scale domain based on the TF masking kernel in Figure 11.1, then (2) remove components below threshold. It has to be considered, however, that removing a component in a time-scale representation is tricky. Because of the strong correlation between components (as a consequence of the reproducing kernel $K_g(a', b', a, b)$, see Sec. 1.3.2), removing components in a wavelet transform leads to a new representation which is *not* a wavelet transform anymore (*i.e.*, it does not satisfy the reproducing kernel property). Applying the resynthesis formula on the modified representation is equivalent to projecting this representation onto the wavelets space (see Eq. (1.19)). This can in turn cause some of the previously removed components to re-appear in the wavelet transform of the reconstructed signal. We were conscious of this problem when implementing the TF masking data in the algorithms and discuss the problem outcomes below.

13.2.1 SPL normalization

The first step towards incorporating perceptual attributes in the signals representation consisted in scaling the representation in physical units related to perception. As far as auditory perception is concerned, sound signal information is usually scaled in sound pressure level (SPL). The difficulty in the SPL normalization lies in that the actual playback level remains unknown during the entire signal processing. To obtain wavelet coefficients scaled in dB SPL, we used a normalization similar to that used in perceptual audio coders (see Sec. 4.1.2). Specifically, we considered that an amplitude variation of ±1 bit in the signal is associated with a SPL of 0 dB, while a full-scale signal is associated with a SPL close to 92 dB. This

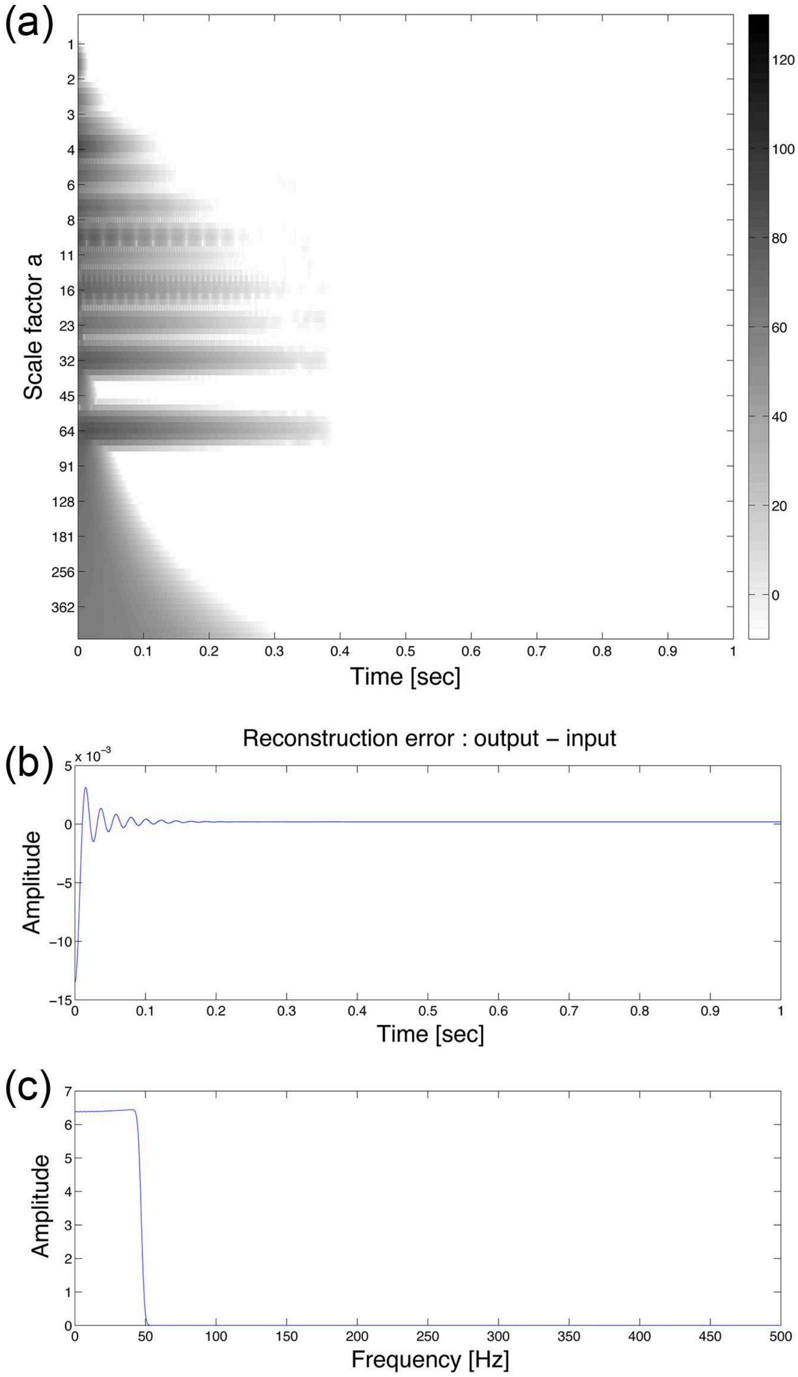


Figure 13.3: (**a**) Scalogram of a wood impact sound synthesized in Aramaki and Kronland-Martinet (2006). The analysis was made with $\mathcal{D}_o = 9$ and $\mathcal{D}_v = 12$, *i.e.*, spanned frequencies from 30 Hz to 20 kHz. The residual (output - input) is shown in (**b**). The spectrum of the error is shown in (**c**). To show up the reconstruction error at low frequencies (≤ 50 Hz), the frequency axis spans [0–500 Hz]. For frequencies > 500 Hz, the error is $< 10^{-15}$.

was achieved by normalizing the input audio samples, $s(k)$, according to

$$s(k) \;=\; s(k)\,2^{n_{\text{bits}}} \tag{13.7}$$

where n_{bits} is the number of bits per sample. In the standard audio file formats such as AIFF or WAVE, signals are sampled at 44.1 kHz using pulse code modulation with $n_{\text{bits}} = 16$. This bit resolution yields a dynamic range of $20\log(2^{16}) \approx 96$ dB.

13.2.2 The absolute threshold of hearing

The second step consisted in taking the absolute threshold of hearing into account in the time-scale representation. This was achieved by using the function $T_q(f)$ in Equation (4.5) (Terhardt, 1979), and mapping this function to the scale domain, *i.e.*, $T_q(f) \mapsto T_q(f_0/a)$. This function is used below to identify local maskers in the signal representation.

13.2.3 Discretization of the time-frequency masking kernel in the time-scale domain

The implemented time-scale analysis scheme decomposes the signals into 108 scales, split into 9 octaves ($\mathcal{D}_o = 9$) and 12 voices per octave ($\mathcal{D}_v = 12$). $\mathcal{D}_o$ was chosen so as to analyze frequencies in the range 30 Hz–20 kHz[1]. $\mathcal{D}_v$ was chosen so as to maintain a large overlap between successive wavelets, thus not to loose details in the signals. The wavelets (defined in Eq. (13.2)) had a constant relative bandwidth of 0.15, which corresponds to approximately one CB.

Accordingly, the TF masking kernel in Figure 11.1 had to be discretized in time and scales. Because we conserved all time samples of the signal, the ΔT axis was unchanged. Conversely, the ΔF axis (in ERB units) had to be matched to the scale axis (in voices and octaves). Considering that (1) the ERB of an auditory filter corresponds to approximately one third of octave (Fletcher, 1940; Zwicker, 1961; Greenwood, 1961a; Scharf, 1970; Moore and Glasberg, 1983b) and (2) the present analysis counts 12 voices per octave, then one ERB unit should be associated with 4 voices. Because the TF masking kernel in Figure 11.1 covers a range of 15 ERB units (from ΔF = -5 to +10 ERB units), the ΔF axis should be divided into 61 voices. This was achieved by interpolating the ΔF axis at a sampling rate of 4 voices per ERB unit, based on a two-dimensional cubic spline fit along the TF plane. The resulting discrete masking kernel is represented in Figure 13.4 in the time-scale domain.

13.2.4 Implementation of the masking kernel in a time-scale filter

The discrete masking kernel above was used as a predictor of masking in the time-scale domain using two different approaches. The first consisted in convolving the masking kernel with the wavelet transform of the input signal so as to obtain a global masking threshold. The second, more pragmatic approach consisted in

[1] Because of the numerical divergence caused by the normalization term in Eq. (13.4), the signal reconstruction is made for frequencies in the range 40 Hz–20 kHz.

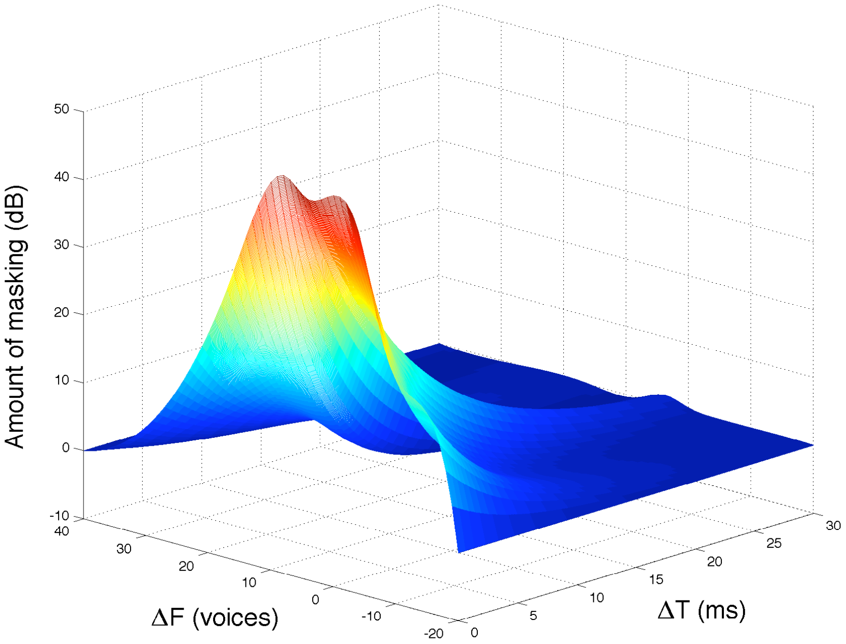


Figure 13.4: Representation of the discrete time-scale masking kernel implemented in the signal processing algorithm. The ΔF axis was sampled at 4 voices per ERB unit. The time axis was sampled at 44.1 kHz.

applying the masking kernel to prominent components or "local maskers" in the wavelet transform.

13.2.4.1 Notations and definition

We denote by $S_g(a,b)$, with $a = \{a_j; j = 0, \ldots, \mathcal{D}_o\mathcal{D}_v - 1\}$ the wavelet transform of the input signal $s(k)$. Unless otherwise stated, all signals were sampled at F_S = 44.1 kHz. The representation after thresholding (*i.e.*, from which components have been removed) is referred to as $\widetilde{S}_g(a,b)$. Accordingly, the output signal (reconstructed from $\widetilde{S}_g(a,b)$) is referred to as $\tilde{s}(k)$.

$\mathcal{M}(a,b)$ refers to the discrete masking kernel in dB (see Fig. 13.4), while $m(a,b)$ refers to the linear masking kernel defined by $m(a,b) = 10^{\mathcal{M}(a,b)/20}$ and normalized between 0 and 1, *i.e.*, $m(a,b) = \frac{m(a,b)}{\max\{m\}}$.

Given $f, g \in L^2(\mathrm{R})$ two functions of two discrete variables $k_1, k_2 \in \mathbb{Z}$, the two-dimensional convolution of f and g is

$$(f * g)(k_1,k_2) \ = \sum_{k_1'}\sum_{k_2'} f(k_1,k_2)\, g(k_1 - k_1', k_2 - k_2') \tag{13.8}$$

13.2.4.2 Convolution of the masking kernel in the time-scale domain

The global architecture of the time-scale convolution filter algorithm is presented in Figure 13.5. Note that a similar approach was used in the "irrelevance filter" algorithm developed by Balazs et al. (2010). The irrelevance filter allows removing the irrelevant TF components in the Gabor transforms of real-world sounds. However, the selection of irrelevant components in the cited algorithm is based on a model of simultaneous masking only using the so-called spreading function of masking in Equation (4.1).

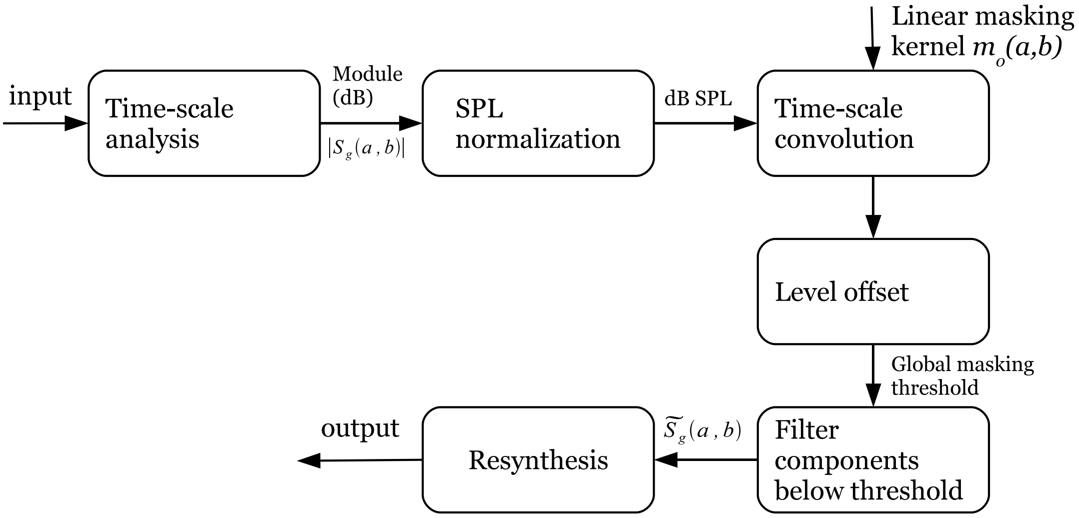


Figure 13.5: Global architecture of the time-scale convolution filter.

The two-dimensional convolution of the wavelet transform and the linear masking kernel is computed in the power domain as $\left(|S_g|^2 * m_0\right)(a,b)$ (see Eq. (13.8)) where

$$m_0(a,b) = \begin{cases} 0 & \text{for } a=0,\ b=0 \\ m(a,b) & \text{elsewhere} \end{cases}$$

The result of the convolution is then used to compute a global masking threshold $\mathbf{G}_M(a,b)$ as

$$\mathbf{G}_M(a,b) = 10\log\left[\left(|S_g|^2 * m_0\right)(a,b)\right] - \epsilon \quad \text{(in dB)} \tag{13.9}$$

where ϵ is a level offset parameter (specified in dB). The use of $m_0(a,b)$ (instead of $m(a,b)$) in the convolution product reflects the assumption that a given component at location (a_j, b_k) cannot influence the global masking threshold at the same location or, in other words, that a component cannot mask itself.

Finally, the modified representation is computed as

$$\widetilde{S}_g(a,b) = \begin{cases} S_g(a,b) & \text{if} & |S_g(a,b)| \geq \mathbf{G}_M(a,b) \quad \text{(in dB)} \\ 0 & \text{otherwise} & \end{cases}$$

which is equivalent to a thresholding operation in the time-scale domain.

Using the convolution implies a linear summation of energy in the time-scale domain, *i.e.*, a linear additivity of TF masking effects. Although a power-law additivity may be more appropriate in certain masker-target configurations (see Sec. 3.4 and Laback et al. 2008), we found worth trying this approach for the ease of implementation. A linear summation of energy results in that $\mathbf{G}_M$ takes the contributions of all components into account. However, it has to be considered that the implemented time-scale grid (see Fig. 13.1) with $\mathcal{D}_o = 9$, $\mathcal{D}_v$ =12 and $F_S =$

44.1 kHz provides a highly redundant signal representation. In such a representation, a single Gaussian-shaped sinusoid as used in psychoacoustical experiments is represented by a set of time-scale components with complex amplitude and phase relationships. Consequently, the linear energy summation of all these components might result in the overestimation of $\mathbf{G}_M(a,b)$. Ideally, the accurate prediction of $\mathbf{G}_M(a,b)$ would require that a Gaussian masker as used in Experiment 6 be represented by a single component in the time-scale domain. This could be achieved by using a sparser signal representation than the currently implemented one, *i.e.*, by opting for dependent time and scale parameters in the discretization (see, *e.g.*, Rioul and Vetterli, 1991; Vetterli and Kovačević, 1995).

Nonetheless, we evaluated the performance of the time-scale convolution filter by reproducing one of the experimental conditions measured in Experiment 6. We synthesized a test signal $s(t)$ that consisted of a sum of two Gaussian-shaped sinusoids with time and frequency shifts

$$s(t) = \underbrace{\mathfrak{g}_m(t)}_{\text{Masker}} + \underbrace{\mathfrak{g}_t(t - \Delta T)}_{\text{Target}} \tag{13.10}$$

with (see Eq. (5.1))

$$\begin{aligned} \mathfrak{g}_m(t) &= A_m \sin\left(2\pi f_m t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2} \\ \mathfrak{g}_t(t) &= A_t \sin\left(2\pi f_t t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2} \end{aligned}$$

and defined the following set of parameters for $s(t)$:

- $\Gamma = 600$;
- Masker: $f_m = 4$ kHz, $L_M = 80$ dB SPL [2];
- Target: $f_t = 6.3$ kHz (*i.e.*, $\Delta F = +4$ ERB units), $L_T = 50$ dB SPL;
- $\Delta T = 10$ ms.

The ϵ parameter of the filter was arbitrarily fixed to 10 dB (the lowest and highest ϵ values in Balazs et al. 2010 were 3 and 9 dB, respectively). The mean results from Experiment 6 with $\Delta T = 10$ ms and $\Delta F = +4$ ERB units indicate a target SPL at threshold of about 25 dB. Thus, the target component in $s(t)$ should not be masked, which means that $\mathfrak{g}_t(t)$ should not be filtered out from the wavelet transform of $s(t)$.

Figure 13.7 depicts the original scalogram of $s(t)$ (Fig. 13.7a) and the scalogram after thresholding (Fig. 13.7b) in dB SPL. It can be seen that the target was erased from the signal representation. The representation of the masker was also altered by the filter. To evaluate the amount of components filtered out from $S_g(a,b)$, we computed the binary representations associated with the input (Fig. 13.7c) and output (Fig. 13.7d) scalograms. These images comprise pixels '1' (in black) where $|S_g(a,b)|$ (respectively, $|\widetilde{S}_g(a,b)|$) > -10 dB SPL and pixels '0' (white) elsewhere. It can be seen that the time-scale convolution filter with $\epsilon = 10$ dB removed about 92% components in $|S_g(a,b)|$, which is huge. It has to be retained, however, that the scalogram after thresholding in Figure 13.7b is *not* the actual representation

[2] The SPL of the test signals was controlled by setting the signals amplitude $A = 10^{(L-92)/20}$ where L is the desired SPL and 92 dB corresponds to the amplitude of a full-scale signal (see Eq. (13.7)).

of the reconstructed signal, *i.e.*, this is not what we listen to after resynthesis. Because of the reproducing kernel, reconstructing the signal from the modified representation ($\widetilde{S}_g(a,b)$) restores some of the removed components. To illustrate the effect of $K_g(a',b',a,b)$ on $\widetilde{S}_g(a,b)$, the scalogram of $\tilde{s}(t)$, *i.e.*, the analysis of the reconstructed signal, is represented in Figure 13.7e. Computing the binary representation associated with Figure 13.7e indicated that about 23% removed components have been restored. Although the representation of $\tilde{s}(t)$ in Figure 13.7e is not strictly identical to the representation of $\mathfrak{g}_m(t)$ in Figure 13.7a), listening to $\tilde{s}(t)$ revealed no perceptual difference to $\mathfrak{g}_m(t)$.

Overall, these preliminary results suggest that an ϵ value of 10 dB is too low to compensate for the overestimation of $\mathbf{G}_M(a,b)$. Figure 13.6 shows $\mathbf{G}_M(a,b)$ (in dB SPL) resulting from the time-scale convolution of $|S_g(a,b)|^2$ and $m_0(a,b)$ with $\epsilon = 0$. The masking threshold at the target location ($\Delta T = 10$ ms, $\Delta F = +4$ ERB units) is about 80 dB SPL, *i.e.*, 55 dB above the actual masked threshold. Such an overestimation of $\mathbf{G}_M$ was expected given (1) the high redundancy of the signal representation and (2) the linear summation of energy in the time-scale domain.

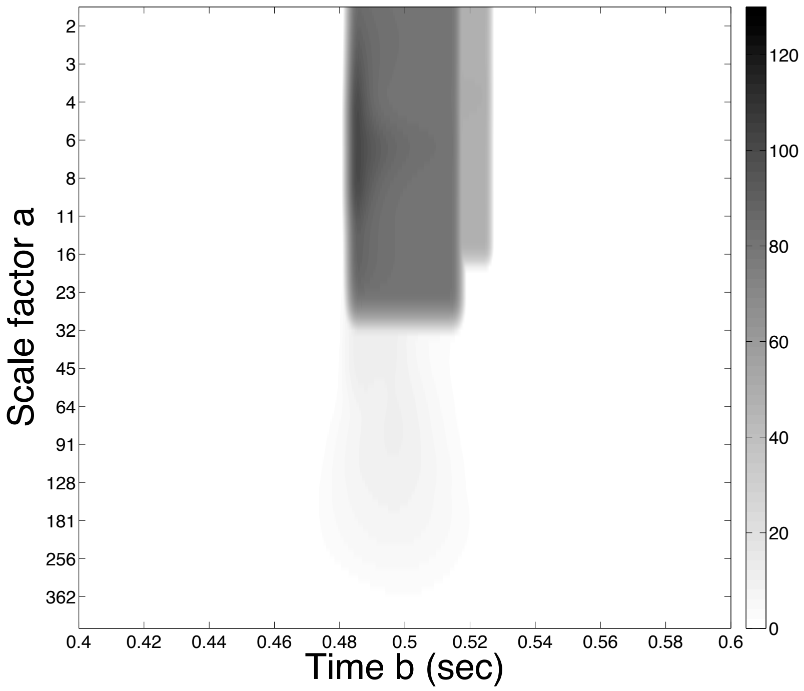
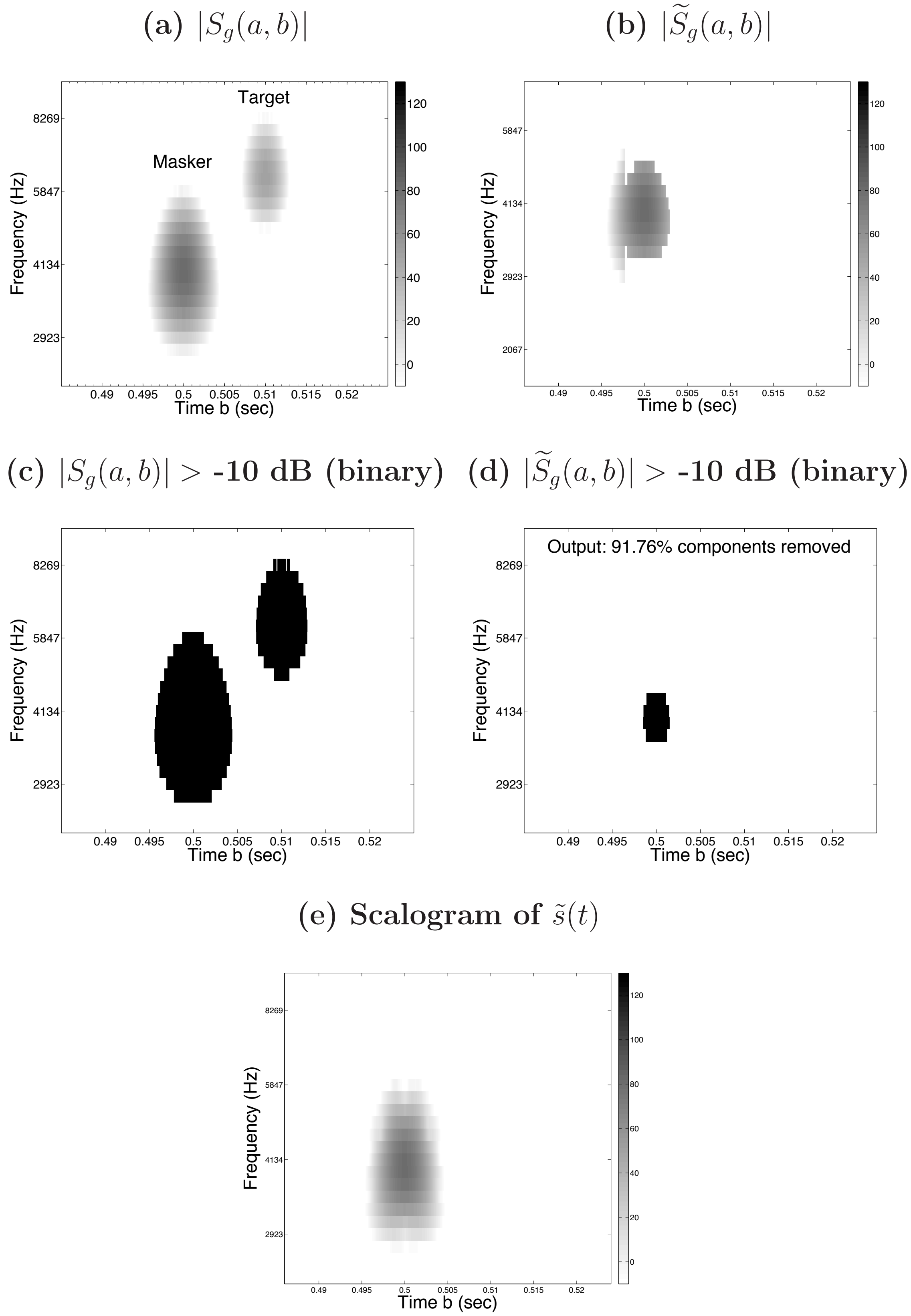


Figure 13.6: Global masking threshold $\mathbf{G}_M(a,b)$ with $\epsilon = 0$ (in dB SPL, see Eq. (13.9)) computed for test signal $s(t)$.



(a) $|S_g(a,b)|$

(b) $|\widetilde{S}_g(a,b)|$

(c) $|S_g(a,b)| >$ **-10 dB (binary)**

(d) $|\widetilde{S}_g(a,b)| >$ **-10 dB (binary)**

(e) **Scalogram of** $\tilde{s}(t)$

Figure 13.7: Scalograms (in dB SPL) of test signal $s(t)$ with $\Delta T = 10$ ms and $\Delta F = +4$ ERB units ($L_M = 80,\ L_T = 50$ dB SPL) (a) at the input and (b) at the output of the time-scale convolution filter with $\epsilon = 10$ dB. The corresponding binary representations are depicted in (c) and (d). (e) Scalogram (in dB SPL) of the reconstructed signal $\tilde{s}(t)$.

13.2.4.3 Local application of the masking kernel

We therefore tested a second, more pragmatic approach whose global architecture is described in Figure 13.8.

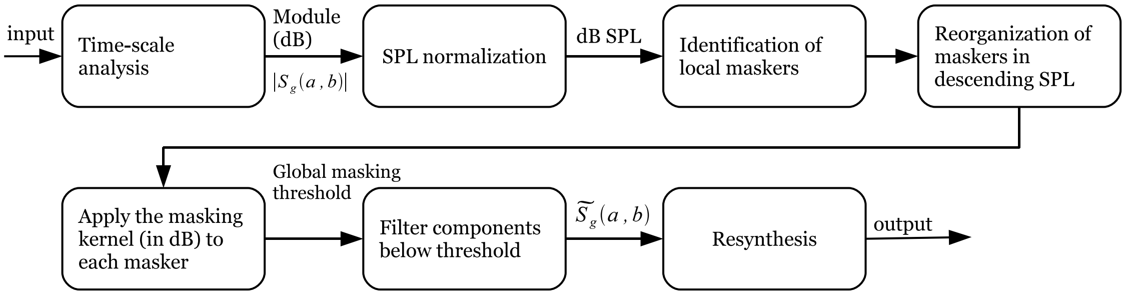

input
Time-scale analysis
Module (dB)
$|S_g(a,b)|$
SPL normalization
dB SPL
Identification of local maskers
Reorganization of maskers in descending SPL
Apply the masking kernel (in dB) to each masker
Global masking threshold
Filter components below threshold
$\widetilde{S}_g(a,b)$
Resynthesis
output

Figure 13.8: Global architecture of the "local" approach of the time-scale filter.

After computation and SPL normalization of the wavelet transform, the first step consisted in identifying local maskers, *i.e.*, local maxima in the transform which verify

$$|S_g(a,b)| \geq Tq(a,\cdot) + 60 \quad \text{(dB SPL)}$$

namely the components whose SPL emerges from the threshold in quiet by 60 dB. This selection rule follows from the masker level used in Experiment 6 ($L_M = 60$ dB SL). Let Ω_M denote the set of maskers constituted in step 1.

The second step consisted in the reorganization of maskers in descending SPL. Then, in the third step, the masking kernel $\mathcal{M}(a,b)$ (in dB) was applied to each masker, and the output wavelet transform was iteratively computed as

$$\widetilde{S}_g(a,b) = \begin{cases} S_g(a,b) & \text{if} \quad |S_g(a,b)| \geq Tq(a,\cdot) + \mathcal{M}(a,b) \quad \text{(dB SPL)} \\ 0 & \text{otherwise} \end{cases}$$

until Ω_M was empty.

Unlike the convolution filter, the local filter does not take the contributions of all components into account in the estimation of the global masking threshold. Instead, the masking kernel is applied to local components only. This should contribute to minimizing the overestimation of masking threshold due to the highly redundant signal representation. Figure 13.9 shows the "local" time-scale filter's response to test signal $s(t)$ in Equation (13.10) with the same parameters as above (*i.e.*, $\Delta T =$ 10 ms, $\Delta F = +4$ ERB units, $L_M = 80$, and $L_T = 50$ dB SPL). It can be seen that the target was not removed, consistent with the results from Experiment 6. However, the representation of the masker was roughly altered by the filter. Comparing the input (Fig. 13.9c) and output (Fig. 13.9d) binary representations indicates that about 45% components were filtered out by the local filter, *i.e.*, about twice less than by the convolution filter (see Fig. 13.7d).

To examine the signal reconstruction, Figure 13.9e shows the scalogram of $\tilde{s}(t)$. It can be seen that the masker components removed by the filter were restored by the reproducing kernel. Informal listening revealed no perceptual difference between $\tilde{s}(t)$ and $s(t)$, and the reconstruction error (*i.e.*, $\tilde{s}(t) - s(t)$) was below 10^{-4}.

To evaluate the performance of the local filter in conditions where the target is masked, we defined a second set of parameters for $s(t)$:

- $\Gamma = 600$;
- Masker: $f_m = 4$ kHz, $L_M = 80$ dB SPL;
- Target: $f_t = 3.2$ kHz (*i.e.*, $\Delta F =$ -2 ERB units), $L_T = 15$ dB SPL;
- $\Delta T = 5$ ms.

In these conditions, the mean results from Experiment 6 indicate a target SPL of about 17 dB. Therefore, $\mathfrak{g}_t(t)$ should be filtered out from $S_g(a, b)$. Figure 13.10a shows $|S_g(a, b)|$ and Figure 13.10b shows the scalogram after thresholding (in dB SPL). As expected, the target was erased from $S_g(a, b)$. The binary representations (Figs. 13.10c and 13.10d) indicate that about 57% components were removed. The scalogram of $\tilde{s}(t)$ is represented in Figure 13.10e. It can be seen that the reproducing kernel restored the masker components removed by the filter. Informal listening revealed no perceptual difference between $\tilde{s}(t)$ and $\mathfrak{g}_m(t)$.

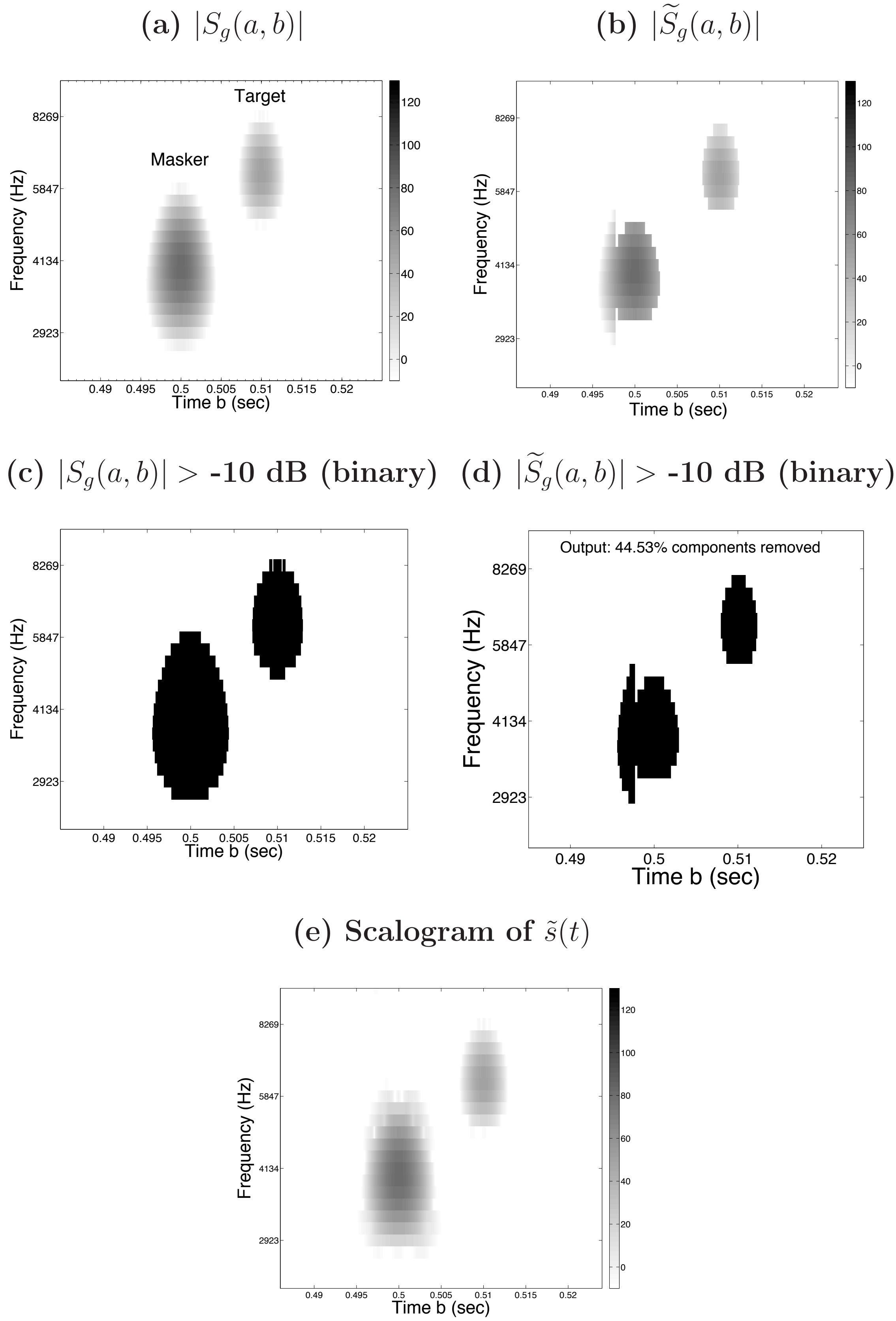
13.3 Discussion

We proposed two algorithms (referred to as the “convolution” and the “local” time-scale filters, respectively) for removing the perceptually irrelevant components in the wavelet transform of a real-world sound while causing no audible difference to the original sound after resynthesis. Preliminary results obtained with elementary sounds indicated that the local filter removes proper information in the signal representation (as predicted from experimental data), while the convolution filter systematically overestimates the masking threshold, and hence removes perceptually *relevant* information. It has to be considered that these two algorithms constitute very preliminary work on the implementation of TF masking data in a sound signal processing tool and require further advancements in signal processing. Therefore, we focus below on the possible improvements of the implementation rather than on the performance of the current algorithms.

First, the two approaches tested above consisted in computing a global masking threshold in the time-scale domain and removing the components below threshold. As discussed above, it is tricky to remove components in time-scale representations. Therefore, an alternative approach could consist in encoding the masked components on a smaller number of bits than the perceptually relevant components, as currently done in perceptual audio codecs such as MPEG-1 Layer III (see Sec. 4.1.2).

Second, it appeared that the determination of the masking threshold in both filters is roughly dependent on the discretization of the wavelet transform. The highly redundant signal representation we opted for in the present study indeed caused an overestimation of masking effects. In future works, a more appropriate discretization should be chosen so as to represent a Gaussian-shaped sinusoid with a shape factor $\alpha = 0.15$ by a single component or “atom” in the signal representation. Such a discretization could be, for example, a dyadic grid (Rioul and Vetterli, 1991; Flandrin, 1993; Vetterli and Kovačević, 1995). We could also opt for a different wavelets family, as proposed by Abolhassani and Salimpour (2008). Another possibility to avoid the constraint on the discretization could be to use an explicit function of TF masking instead of the discrete masking kernel to determine the global masking threshold for any time-scale coordinates. Such a function might provide faster implementation and better results than the masking kernel. For instance, the exponential masking function designed in Section 12.1.1 could be implemented.

Third, the current masking kernel was designed based on TF masking data for a single Gaussian masker. A collaborative study is currently underway at the Acoustics Research Institute, which investigates the additivity of masking arising from up to four Gaussian maskers shifted in time and frequency. It would be interesting to explore (1) the extent to which these new data on the additivity of TF masking can be implemented in the current algorithm, and (2) the degree to which the filter's performance is improved. Additionally, it would be interesting to consider the effect of masker level on the TF masking pattern's shape. Combining the results from Experiment 3 (which explored the level-dependency of simultaneous masking for a Gaussian masker, see Sec. 8.2) and literature data on the level-dependency of forward masking (see Sec. 3.3.2) may allow designing a level-dependent TF masking kernel or function.



(a) $|S_g(a,b)|$

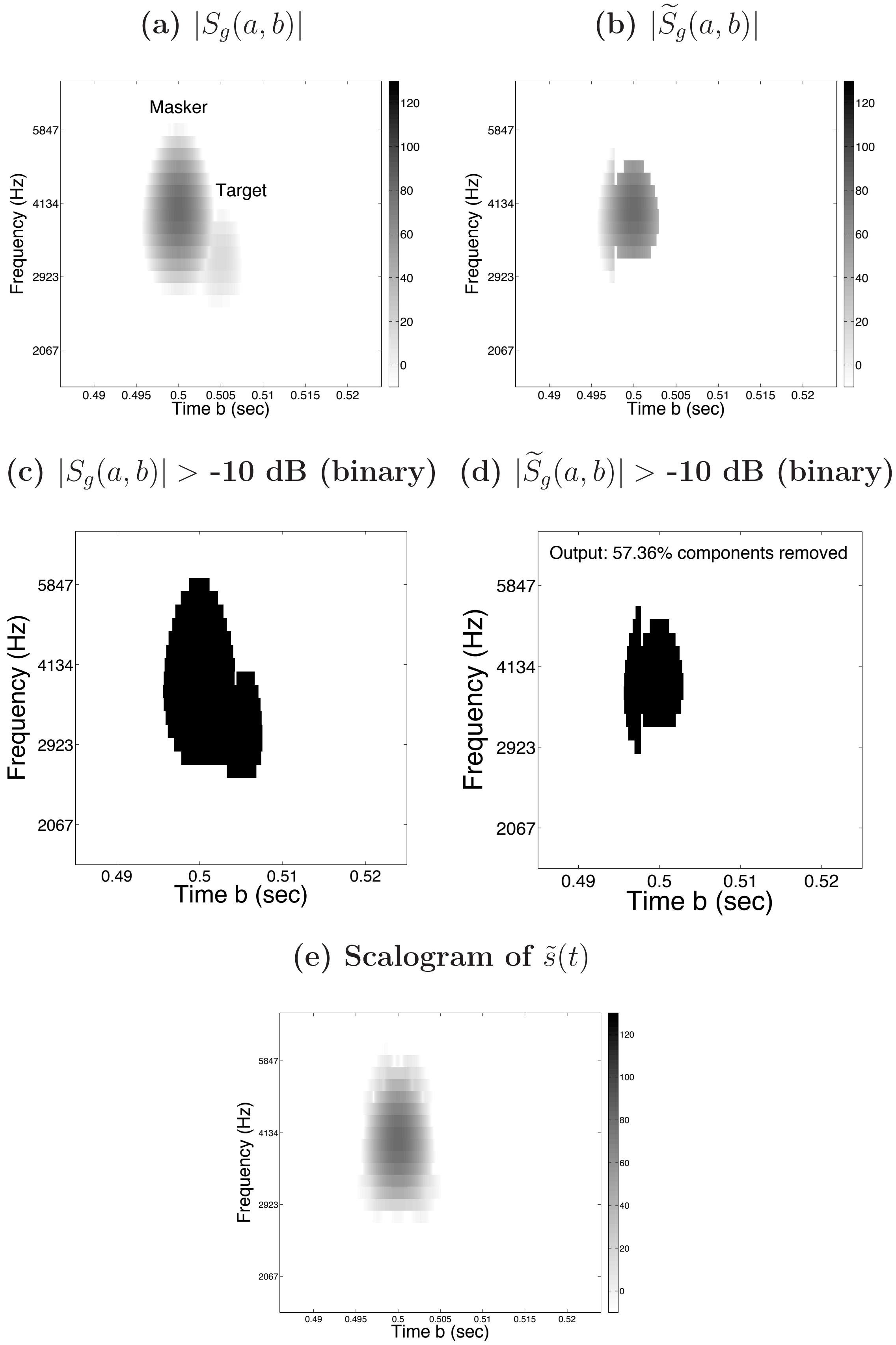
(b) $|\widetilde{S}_g(a,b)|$

(c) $|S_g(a,b)| >$ **-10 dB (binary)**

(d) $|\widetilde{S}_g(a,b)| >$ **-10 dB (binary)**

(e) Scalogram of $\tilde{s}(t)$

Figure 13.9: Scalograms (in dB SPL) of test signal $s(t)$ with $\Delta T = 10$ ms and $\Delta F = +4$ ERB units ($L_M = 80,\ L_T = 50$ dB SPL) (a) at the input and (b) at the output of the local time-scale filter. The corresponding binary representations are depicted in (c) and (d). (e) Scalogram (in dB SPL) of the reconstructed signal $\tilde{s}(t)$.


(a) $|S_g(a,b)|$

(b) $|\widetilde{S}_g(a,b)|$

(c) $|S_g(a,b)| >$ **-10 dB (binary)**

(d) $|\widetilde{S}_g(a,b)| >$ **-10 dB (binary)**

(e) **Scalogram of** $\tilde{s}(t)$

Figure 13.10: Same as Figure 13.9 but for test signal $s(t)$ with $\Delta T = 5$ ms and ΔF = -2 ERB units ($L_M = 80,\ L_T = 15$ dB SPL).

Part IV

CONCLUSIONS AND PERSPECTIVES

Chapter 14

Conclusions and perspectives

Conclusions

This dissertation presented psychoacoustical results of auditory masking in the time-frequency (TF) domain for stimuli with maximal concentration in the TF plane, namely Gaussian-shaped sinusoids with fixed bandwidth (ERB = 600 Hz) and duration (ERD = 1.7 ms; 0-amplitude duration = 9.6 ms). These stimuli minimize both the Heisenberg's uncertainty principle and the number of excited auditory TF observation windows. We used Gaussian stimuli to obtain a measure of the TF spread of masking produced by an "elementary" sound, *i.e.*, a short *and* narrowband signal. Any real-world sound can be decomposed into a set of elementary functions (or "atoms") well localized in the TF domain. Acquiring knowledge on the "basic" spread of TF masking produced by an atom would constitute an important advance towards describing and predicting auditory masking for complex sounds.

In the experiments, the Gaussian masker had a carrier frequency of 4 kHz and a level of 60 dB SL. Masker and target were separated either in frequency, in time, or both. The results for the simultaneous conditions (Exp. 2–4) allowed us to show that masking patterns for Gaussian maskers are consistent with those previously reported in the literature for sinusoidal maskers with comparable frequencies and levels and temporally controlled with windows limiting spectral broadening.

The results for the non-simultaneous conditions (Exp. 5) allowed us to show that (1) the decay of forward masking is a linear function of $\log(\Delta T)$, a result consistent with previous data for maskers with various spectro-temporal characteristics, and (2) the temporal spread of forward masking for a 9.6-ms Gaussian masker is narrower than that for maskers with comparable bandwidths, frequencies and levels but with longer durations.

The results for the TF conditions (Exp. 6) were in close agreement with the few preceding studies that have varied both ΔT and ΔF with long-duration sinusoidal maskers.

There are models of TF masking that are currently implemented in perceptual audio codecs (such as MP3). Most of these models simply assume a linear combination of simultaneous and forward masking functions to account for TF masking. However, given the highly non linear behavior of cochlear mechanics, such a simple combination of temporal and frequency masking functions is unlikely to provide an accurate representation of TF masking. To prove that the spread of TF masking cannot be deduced from masking data measured separately in the time and frequency domains, we tested two simple prediction schemes assuming a linear combination of temporal and frequency masking. Specifically, we assessed whether the results from Experiment 6 (TF masking) could be predicted by the results from Experiment 2 (frequency masking) and 5 (temporal masking). It appeared that the two prediction schemes failed in predicting the TF masking data from Experiment 6. This indicates that combining temporal and frequency masking data measured separately does not provide an accurate representation of TF masking. Consequently, audio coding algorithms using such an approach provide rather erroneous predictions of TF masking.

Overall, the TF masking data gathered in Experiment 6 provide a measure of the spread of TF masking produced by an "elementary" sound, *i.e.*, with maximal concentration in the TF domain. These new data constitute a crucial basis for the prediction of auditory masking for complex sounds. On that purpose, we attempted to develop a model of TF masking. An exponential function of the form $C(\Delta F)\, e^{-\Delta T/\lambda(\Delta F)}$ was proposed that could roughly predict the experimental data for ΔFs $\neq 0$ and ΔTs up to 10 ms. However, because the ΔF-dependency of parameters C and λ lies on a set of four equations (1^{st}- and 3^{rd}-order polynomials, respectively), this model was not implemented. Future works will consist in the development of an explicit function of TF masking which would be easier to implement than the current exponential function. For example, it would be interesting to examine the extent to which the amount of TF masking can be described by a tensor product of temporal and frequency masking functions. The achievement of a robust TF masking function would constitute an important advance in the field of perceptual audio coding. Indeed, replacing the currently implemented "spreading function" of masking (which only accounts for simultaneous masking) in audio coding algorithms by such a function would improve certain features in the efficiency of these algorithms (although maybe not their compression rate).

In an attempt to match the time-scale representations of audio signals to human auditory perception, we proposed two algorithms for removing the perceptually irrelevant components in the wavelet transform of a real-world sound while causing no audible difference after resynthesis. First, we applied the TF masking *kernel* (obtained by discretization of the mean TF masking pattern in Exp. 6) to local maxima in the wavelet transform. This provided good predictions of the masking effects between the components. The second approach was based on a two-dimensional convolution of the wavelet transform and the masking kernel. This algorithm systematically overestimated masking. The performance of the two algorithms was mainly evaluated based on preliminary results obtained with simple sounds (such as the Gaussian stimuli used in psychoacoustical experiments) and informal listening by the author. Because of time constraints, none of the tested approaches could be formally evaluated by listening tests. These two algorithms constitute a preliminary step in the implementation of TF masking data in a sound signal processing tool, and require further advancements in signal processing.

Perspectives

Future works will essentially concern the signal processing part of the study, namely the implementation of TF masking data in an algorithm aiming at matching the time-scale representations of audio signals to human auditory perception. As discussed in Section 13.3, this will require further efforts on the development of a TF masking function, on the discretization of the continuous wavelet transform chosen for signal decomposition, and on the manner to "remove" the components in the signal representation without causing any (audible) artifact due to the properties of time-scale representations.

It would also be interesting to consider the effect of masker level on the TF masking pattern's shape. Combining the results from Experiment 3 and literature data on the level-dependency of forward masking may allow designing a level-dependent TF masking function.

A study is currently underway at the *Acoustics Research Institute* (ARI) which investigates the additivity of masking arising from up to four Gaussian maskers distributed in the TF plane. Preliminary results indicated a rather complex additivity of TF masking. It would be interesting to explore the degree to which these new data can be incorporated in the TF masking model.

A major step will consist in the perceptual validation of the algorithm by conducting listening tests with real-world sounds. This shall be done with subjective listening tests (such as MUSHRA, "multiple stimuli with hidden reference and anchor", defined in ITU-R BS.1534-1, 2003) and/or objective listening tests (such as PEAQ, "perceptual evaluation of audio quality", defined in ITU-R BS.1387, 2001).

These investigations will mainly be conducted within the collaboration between the LMA and the ARI. The collaborative project on TF masking indeed follows up within a research project called *MulAC* ("Frame Multipliers: Theory and Application in Acoustics") managed by the ARI. This project is currently underway and lasts until the end of 2012.

Annexe A

Programme de valorisation des compétences : le Nouveau Chapitre de la Thèse

Sommaire

Le programme « Un Nouveau Chapitre de la Thèse$^{®}$ » (NCT) a été imaginé et mis en place par le CNRS, la région Ile de France et l'Association Bernard Gregory[1] (ABG) pour faciliter l'insertion professionnelle des docteurs. Cette formation, destinée aux doctorants en cours de dernière année de thèse, a pour objectif d'aider les futurs docteurs à faire le point sur les compétences et savoir-faire professionnels développés au cours de la préparation de leur doctorat et à se les approprier. L'une des originalités réside dans le fait que les doctorants bénéficient d'un accompagnement personnalisé par un « mentor », professionnel (consultant spécialiste du recrutement) extérieur au monde académique.

Depuis la mise en place du programme en 2002, ce sont près de 2000 doctorants issus d'environ 180 Ecoles Doctorales (ED) distinctes qui on participé à l'exercice. D'après une étude réalisée au premier trimestre 2007, l'impact du NCT se révèle encourageant dans le sens où il apparaît comme :

- un outil valorisant la thèse comme une expérience professionnelle susceptible de déboucher sur un large éventail de perspectives professionnelles,
- une formation débouchant sur une plus grande efficacité de la démarche de recherche d'emploi,
- une aide appréciable à une insertion réussie hors du monde académique, en Recherche ou dans d'autres domaines d'activité.

Alors en pleine réflexion sur mon « après-thèse » lorsque la campagne NCT 2009 a été lancée, j'ai été séduit par le programme et aie décidé de participer à cet exercice

[1] www.abg.asso.fr.

de réflexion et de prise de recul par rapport à ces trois années de Recherche passées au Laboratoire de Mécanique et d'Acoustique (LMA) à Marseille.

A.1 Cadre général et enjeux de la thèse

A.1.1 Le sujet de thèse

Mon projet de thèse porte sur l'étude du masquage sonore. Le masquage se définit comme l'élévation du seuil d'audibilité d'une source sonore (qu'on appellera la *cible*) en présence d'une autre source (le *masque*). Les phénomènes de masquage sonore ont été révélés au début du $XX^{\text{ème}}$ siècle et ont depuis fait l'objet de nombreuses études psychoacoustiques. Ces dernières ont montré que la quantité de masquage varie en fonction de la distance temporelle et/ou fréquentielle séparant les deux sources sonores et de leurs amplitudes. De nombreux modèles de masquage ont ainsi été développés, lesquels sont implémentés dans des outils d'analyse-synthèse et de compression audio tels que le *codage MP3*. Cependant, ces modèles se révèlent incomplets car ils ne prennent en considération que l'aspect fréquentiel et sont basés sur des mesures réalisées avec des sons entretenus temporellement (dans la grande majorité des études psychoacoustiques, le masque a toujours une durée d'au moins 100 millisecondes). Or, des études récentes indiquent que la perception auditive de signaux brefs diffère de celle de signaux de longue durée. Des travaux psychoacoustiques sur les effets de masquage existant avec des sons « étroits » dans les plans temporel *et* fréquentiel doivent donc être développés.

Ma thèse a donc pour objectif de proposer un nouveau modèle de masquage *temps-fréquence* adapté aux propriétés mathématiques des représentations temps-fréquence ou temps-échelle. Ces types de représentation étant très utilisés dans les outils d'analyse et de traitement des signaux audio, l'intégration du nouveau modèle de masquage dans de tels outils permettra de *représenter les sons de manière cohérente avec la perception de ceux-ci*. Ainsi, en ne traitant que l'information « utile » (du point de vue perceptif) contenue dans les signaux, on peut envisager une nette amélioration des algorithmes d'analyse, de traitement et de compression audio, que ce soit en termes d'efficacité, de robustesse et de rapidité. Les enjeux scientifiques et techniques de ma thèse sont donc d'une grande importance, à la fois pour la recherche et l'industrie audio. Si notre algorithme se révélait performant, un dépôt de brevet a même été évoqué en début de thèse, ce qui induirait des retombées économiques importantes pour le laboratoire et tous les acteurs du projet.

Les étapes de réalisation du projet se présentent comme suit :

1. Déterminer les signaux acoustiques les plus appropriés pour la réalisation d'un modèle de masquage temps-fréquence,
2. Effectuer des mesures psychoacoustiques du masquage sonore dans le plan temps-fréquence avec ces signaux,
3. Tenter de modéliser ces données empiriques,
4. Implémenter le modèle de masquage ainsi obtenu dans un outil d'analyse,
5. Valider l'algorithme par des tests perceptifs sur des sons complexes (*musique, langage*...).

A.1.2 Mon projet de recherche dans son contexte

Le projet « masquage sonore temps-fréquence » s'inscrit dans le cadre d'une collaboration entre quatre équipes de recherche réparties sur deux laboratoires : le Laboratoire de Mécanique et d'Acoustique (**LMA**) à Marseille, unité propre de recherche rattachée au département Sciences et Technologies de l'Information et de l'Ingénierie du CNRS, et l'« Acoustics Research Institute » (**ARI**) à Vienne, rattaché à l'Académie des Sciences Autrichienne. Au LMA, le projet implique les équipes « *Perception Auditive* » (PA) et « *Modélisation, Synthèse et Contrôle des Signaux Sonores et Musicaux* » (S2M). Les équipes homologues sont impliquées à ARI, précisément « Psychoacoustique » et « *Mathématiques et Traitement du Signal* ». L'intérêt de cette collaboration repose sur la complémentarité des compétences de chaque laboratoire. Bien que les objectifs scientifiques et l'expérience de chaque partie soient intimement liés, ils se révèlent suffisamment distincts pour favoriser l'émergence de points de vue novateurs inhérents à une volonté forte de *collaboration pluridisciplinaire*, en accord avec les besoins de ce projet.

L'équipe S2M a une grande expérience en traitement du signal audio, en modélisation et en synthèse sonore (domaine créé par Jean-Claude Risset, médaille d'or du CNRS en 1999). Son expertise a déjà été utilisée dans un grand nombre de domaines aussi bien fondamentaux qu'industriels (PSA Peugeot-Citroën, France Télécom, cinéma, facteurs d'instruments de musique). L'équipe a également développé un réseau de collaborations nationales et internationales parmi lesquelles on peut citer : l'Institut des Neurosciences Cognitives de la Méditerranée (Marseille), l'Institut de Recherche et de Coordination Acoustique Musique (Paris), les universités de Stanford et de Berkeley. De son côté, l'équipe PA est experte dans l'étude des traitements des sons par le système auditif (« psychoacoustique »). Elle travaille actuellement sur de nombreuses thématiques très impliquées dans les technologies de l'information et de la communication et l'environnement sonore. On citera entre autres le masquage, la sonie, la localisation auditive ou la caractérisation d'enceintes acoustiques. Dans ce contexte, l'équipe développe de nombreuses collaborations avec des laboratoires et des industriels (IRCAM, Genesis, Canon).

Le laboratoire ARI a une grande expérience dans le développement, l'implémentation et l'évaluation de modèles psychoacoustiques. ARI est actuellement coordinateur de plusieurs projets de recherche nationaux (projet NoiseDESCription, département linguistique de l'université de Vienne) et internationaux (société Müller BBM et l'université de Göttingen en Allemagne, l'« International Computer Science Institute » de Berkeley en Californie).

Fort de cette collaboration franco-autrichienne, le projet « masquage temps-fréquence » entre dans une perspective d'innovation et d'amélioration des outils de traitement audio. Bien que les bases de notre projet reposent sur les propriétés des signaux mis en œuvre dans une étude psychoacoustique antérieure (van Schijndel *et coll.*, 1999), l'utilisation de ces signaux dans l'étude du masquage sonore constitue une approche nouvelle. Plus encore, l'association de la théorie mathématique des représentations temps-fréquence aux résultats de la psychoacoustique devrait permettre d'aboutir à des connaissances et savoir-faire nouveaux en termes de traitement des signaux sonores. A ma connaissance, aucun autre laboratoire de recherche public ou privé ne travaille au développement de tels outils à ce jour. D'ailleurs, la présentation du projet lors de congrès nationaux et internationaux a suscité de nombreuses réactions - à la fois positives et négatives - de la part de chercheurs et représentants de l'industrie. Ceci nous conforte dans le sentiment « *novateur* » de notre projet.

Concrètement, le projet a été initié en décembre 2005. Les principaux objectifs de recherche ont été définis au cours de diverses réunions tenues à Marseille et à Vienne et le protocole expérimental de mesures psychoacoustiques a été mis en place début 2006. Un sujet de stage (niveau Master II) doté d'une possibilité de poursuite en thèse (sous réserve de financement) a été proposé. Ce dernier a été été rapidement pourvu et s'est déroulé au LMA de mars à juillet 2006. Il a permis la réalisation de mesures préliminaires. Seulement à l'issue du stage l'étudiant n'a pas souhaité poursuivre en thèse. Le sujet de thèse a donc été re-proposé au sein du LMA, c'est à ce moment-là que j'ai déposé ma candidature et commencé à prendre part aux discussions.

A.1.3 Ma place dans ce contexte

Dans le cadre de ma dernière année d'études de cycle « ingénieur » à l'Institut Supérieur de l'Electronique et du Numérique (ISEN) de Toulon, j'ai suivi les cours du Master II « Traitement du signal » à La Garde (83). J'ai ensuite effectué mon stage d'initiation à la recherche au LMA, au sein de l'équipe S2M. Ce dernier portait sur l'analyse-synthèse de sons d'impact. A ce moment là je n'étais pas encore déterminé quant à ma poursuite en thèse ou non. Mais, encouragé par mes encadrants de stage, ma curiosité et l'intérêt grandissant que je porte aux techniques de traitements audio, je me suis finalement intéressé aux sujets de thèse proposés en relation avec les aspects sonores et musicaux. C'est donc en mai 2006 que m'a été présenté le projet « masquage temps-fréquence ». Le contexte global du projet (sujet, laboratoires, collaboration franco-autrichienne, encadrants) m'a fortement séduit et j'ai rapidement posé ma candidature à une bourse ministère (aucune autre possibilité de financement ne s'offrait à moi étant donnés les courts délais impartis), que j'ai obtenue à la fin du mois de juin.

Le projet ayant été cadré avant mon arrivée je n'ai pas joué de rôle particulier dans la définition du sujet. Cependant, la planification initiale du projet (voir § A.2.1) a été décidée lors d'une concertation regroupant tous les acteurs principaux. D'autre part, il a été proposé de mettre en place un programme de co-tutelle de thèse franco-autrichienne, ceci afin d'obtenir des fonds supplémentaires pour les échanges entre la France et l'Autriche (voyages, séjours...), de me permettre d'obtenir un double diplôme de doctorat et de valoriser la collaboration entre les deux laboratoires. Je me suis occupé personnellement de la mise en place de ce programme qui a nécessité beaucoup d'investissement, procédures administratives obliges. La co-tutelle a finalement été validée courant 2007 mais nous n'avons pas réussi à obtenir les ressources financières demandées (motif(s) du refus non explicité(s) par l'administration).

A.1.4 Ressources mises à disposition du projet

La mobilisation de quatre équipes de recherche (réparties entre LMA et ARI) a permis de doter le projet de nombreuses compétences scientifiques et techniques. Sur le plan humain ce sont six chercheurs et un ingénieur qui m'ont accompagné tout au long de la thèse :

- Richard Kronland-Martinet (LMA-S2M), Directeur de Recherche et *co-directeur de thèse*
- Sølvi Ystad (LMA-S2M), Chargée de Recherche
- Sophie Savel (LMA-PA), Chargée de Recherche et *co-directrice de thèse*
- Sabine Meunier (LMA-PA), Chargée de Recherche
- Michel Jevaud (LMA-PA), Ingénieur d'Etude
- Peter Balazs (ARI-« Mathématiques et Traitement du signal »), Chargé de Recherche et *co-directeur de thèse*
- Bernhard Laback (ARI-« Psychoacoustique »), Chargé de Recherche

A ajouter à ces ressources, les services technique et administratif de chaque laboratoire (informatique, secrétariat, documentation).

Sur le plan technique, l'équipe PA a mis à ma disposition un ordinateur équipé des logiciels nécessaires à la programmation des expériences psychoacoustiques ainsi que les locaux et tout le matériel de génération et d'écoute des signaux de mesure. La seconde phase du projet (modélisation et implémentation des données psychoacoustiques, voir § A.1.1) a nécessité l'achat d'un ordinateur par l'équipe S2M. Cette dernière a également fourni les locaux et les logiciels nécessaires (licences détenues par le LMA). Lors des déplacements à Vienne, un ordinateur portable était fourni par l'équipe S2M en cas de nécessité.

A.2 Déroulement, gestion et coût de mon projet

A.2.1 Etude amont et cadrage du projet

Le projet ayant démarré avant mon arrivée, je n'ai pas participé aux études préliminaires. Notamment, le protocole expérimental de mesure et le choix des stimuli m'ont été imposés. D'autre part, la façon d'aborder la phase de modélisation dépendant fortement des résultats obtenus à l'étape 2 (voir § A.1.1), des hypothèses ont été formulées mais les efforts de cadrage en début de thèse se sont concentrés sur les mesures psychoacoustiques. Le planning de travail prévisionnel illustré en Figure A.1 a été décidé.

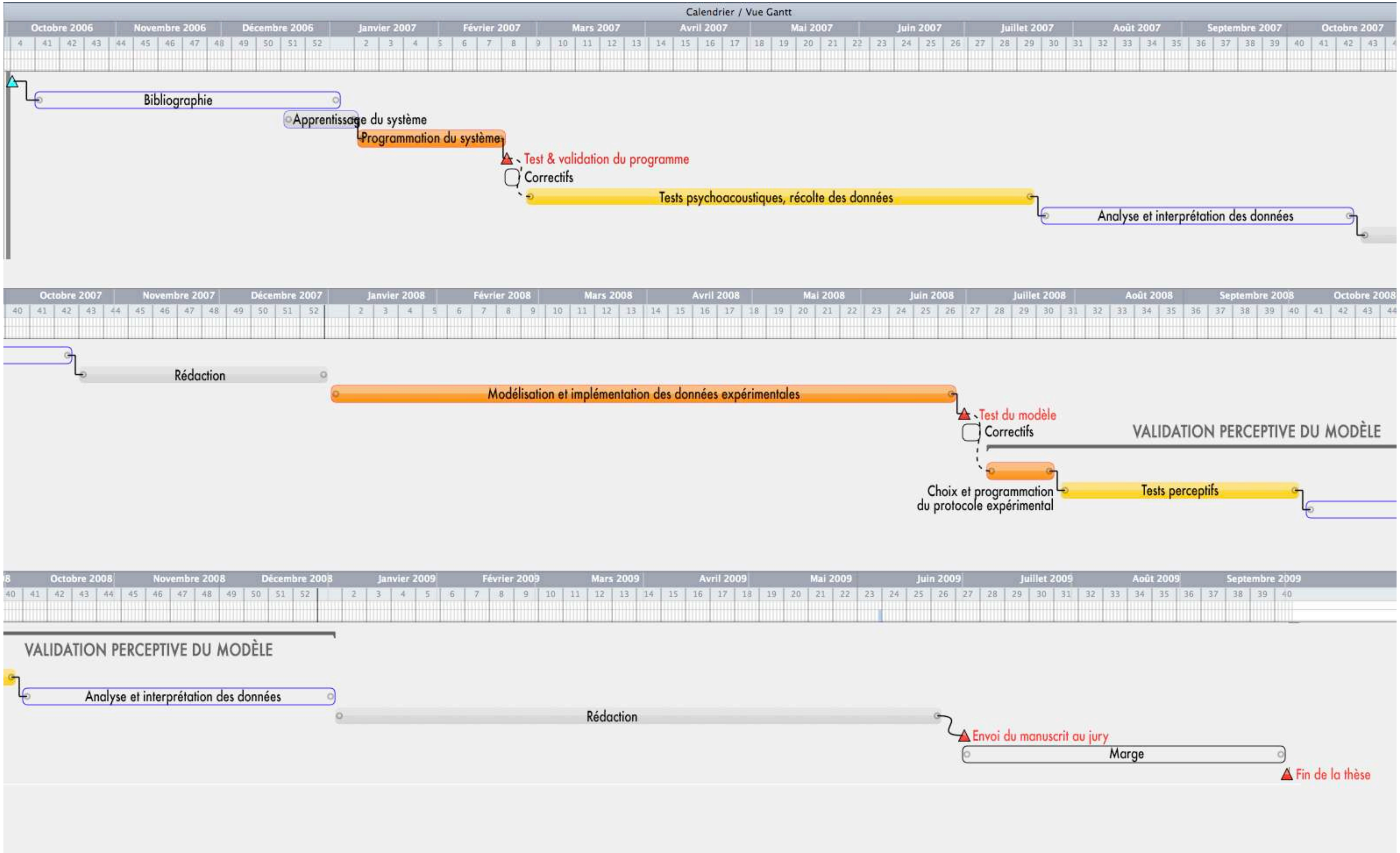


Figure A.1 – Diagramme de Gantt, planification initiale du projet.

Etant issu d'une formation principalement « traitement du signal », je ne disposais pas en début de thèse des pré-requis nécessaires en psychoacoustique. L'étude amont a donc consisté pour moi en une période bibliographique ayant pour objectifs :

1. me familiariser avec les méthodes de psychophysique,
2. comprendre les principes de fonctionnement du système auditif et en particulier ceux impliqués dans le masquage,
3. tenter de formuler des hypothèses quant aux résultats de nos mesures.

Ensuite, la mise en place du protocole expérimental et de la campagne de mesures ont pu débuter. Les risques liés à cette phase de la thèse, ainsi que les actions correctives associées, sont présentés dans le Tableau A.1.

Risques	Actions correctives	Outils d'évaluation
Problème technique sur chaîne de mesure	Remplacement de l'étage défaillant par équivalent (2 systèmes disponibles *in situ*)	Oscilloscope, écoute
Cabine audiométrique indisponible	Changer de salle (3 salles *in situ*)	Calendrier d'occupation des salles
Travaux sur le site (gêne sonore)	Report des mesures ou changement de cabine	Contacter la Direction du laboratoire
Sujet malentendant	Vérifier audiogramme avant mesure	Audiomètre
Tâche difficile pour les sujets	Changer la procédure adaptative	Ecart-type, tracé des résultats
Résultats incohérents	Vérifier les mesures ou si défaillance technique	Bibliographie

Table A.1 – Résultats de l'analyse de risques liés à la phase de mesures psychoacoustiques (étape 2, voir § A.1.1).

Des problèmes non évalués en amont sont cependant survenus pendant cette phase. Ils sont détaillés dans le paragraphe suivant. Quant à la phase de modélisation/implémentation des données de masquage (étapes 3 et 4), l'analyse de risques est présentée dans le Tableau A.2.

Enfin, la phase de validation du modèle étant constituée de tests perceptifs, les risques évalués dans le Tableau A.1 sont applicables à cette étape, bien que les protocoles de mise en œuvre soient différents.

A.2.2 Financement du projet

Le financement du projet a été réparti sur plusieurs organismes (voir Tab. A.3 pour le détail de la répartition) :

- les ressources propres aux équipes PA et S2M,

Risques	Actions correctives	Outils d'évaluation
Problème informatique	Remplacement ordinateur	Technicien, service informatique
Problème réseau, licence logiciel indisponible	Opter pour logiciel équivalent « libre » (sans licence)	Message d'erreur
Perte d'information « utile » (qualité sonore dégradée)	Etalonnage du modèle	Tests perceptifs, Calcul du résidu sonore

Table A.2 – Résultats de l'analyse de risques liés à la phase de modélisation/implémentation des données de masquage (étapes 3 et 4, voir § A.1.1).

- les laboratoires (LMA et ARI),
- le programme de recherche *SenSons*[2]. Ce programme, coordonné par Sølvi Ystad et financé par l'Agence Nationale pour la Recherche, regroupe plusieurs projets de recherche dont le projet « masquage temps-fréquence »,
- la Société Française d'Acoustique (SFA),
- les universités d'Aix-Marseille I et de Vienne (Autriche) dans le cadre de la mise en place d'un programme de co-tutelle de thèse franco-autrichienne. Programme qui a malheureusement échoué,
- l'Ecole Doctorale « Physique, modélisation et sciences pour l'ingénieur » (ED353), notamment pour la prise en charge des formations.

A.2.3 Conduite du projet

A.2.3.1 Gestion de l'avancement du projet

Le projet impliquant de nombreuses personnes réparties sur deux laboratoires différents, plusieurs types de point d'avancement ont été mis en place, choisis en fonction des décisions à prendre à un moment précis :

- Sous réserve de financement disponible, une à deux réunions franco-autrichienne étaient organisées par an, à ARI ou au LMA. Ces réunions avaient pour but de faire un point complet sur l'avancement du projet, sous la forme d'une présentation par chaque partie. Il n'a pas toujours été possible de réunir les six chercheurs impliqués dans le projet mais, dans ce cas, la présence d'au moins un responsable par équipe était souhaitée. Ces réunions se sont avérées très dynamiques et enrichissantes car de nombreux points importants y ont été débattus et défendus lors de désaccords. De plus, les décisions importantes pour ma thèse ont été prises à ces occasions. A noter que ma parole avait un certain poids dans ces discussions.
- Des points d'avancement intermédiaires ont été organisés au sein des équipes du LMA et de ARI. Ces réunions ne regroupaient pas l'ensemble des acteurs puisqu'elles ciblaient un aspect particulier de la thèse (psychoacoustique ou

[2] www.sensons.cnrs-mrs.fr.

modélisation). Après chacun de ces points, un compte-rendu était rédigé et communiqué à tous les chercheurs par courrier électronique.

- Des rencontres régulières (tous les deux mois environ) étaient organisées entre mes directeurs de thèse au LMA (Richard Kronland-Martinet et Sophie Savel) et moi. Il s'agissait de faire des points rapides sur l'avancée des mesures. D'autre part, ces personnes étaient souvent disponibles pour pouvoir m'aider lorsque le besoin se faisait sentir. S'ils n'étaient pas présents « physiquement » au laboratoire, on échangeait par courrier électronique.
- Pendant l'analyse et la rédaction des données psychoacoustiques, des points d'avancements ont été mis en place tous les quinze jours environ avec Sophie Savel, ceci afin d'éviter que je parte dans de mauvaises directions dans l'analyse. Ces points ont cependant été mis en place un peu trop tard. J'ai un peu tardé à alerter Sophie de mes difficultés à analyser correctement les données.

A noter également qu'un espace de discussion privé (forum) a été mis en place sur Internet. Ainsi, chaque membre du projet pouvait à tout moment poser une question, discuter un point particulier ou revenir sur une décision. Cet espace était aussi réservé à l'organisation des prochaines réunions. Cet outil s'est révélé bien pratique.

A.2.3.2 Problèmes rencontrés et solutions apportées

Au cours des mesures psychoacoustiques (étape 2), divers problèmes sont survenus dont certains d'entre eux n'avaient pas été pris en compte dans l'analyse de risques initiale. Ces problèmes peuvent être listés en trois catégories : problèmes techniques, problèmes liés à la mesure (protocole expérimental) et problèmes de méthode.

Des problèmes techniques sont survenus avant et pendant les mesures. Le système de génération des signaux utilisé par l'équipe PA (acquis en 2006) m'a été présenté comme doté d'une puissance de calcul et de mémoire suffisantes pour pouvoir réaliser une génération temps réel des stimuli. Or, lors des premiers tests il s'est avéré que le système n'était pas capable de procéder à une telle génération, le processeur saturait. Avec l'aide de l'ingénieur PA (Michel Jevaud) et après avoir pris contact avec le service après-vente du système, il m'a été confirmé que ce modèle n'était pas assez puissant. Ses capacités ont été surestimées. Deux options s'offraient donc à nous :

1. acquérir un système plus puissant, ce qui aurait induit des coûts supplémentaires,
2. changer le mode de génération.

J'ai opté pour la solution (2). J'ai donc dû ré-écrire l'algorithme de génération, ce qui a induit des délais supplémentaires (*un mois environ*). D'autre problèmes techniques sont survenus pendant les mesures, un atténuateur et un sommateur de signaux ayant montré des signes de faiblesse. Ils ont été remplacés par du matériel équivalent, conformément aux prévisions.

Un problème de mesure prévu dans l'analyse de risques (voir Tab. A.1), est survenu pendant la phase de tests préliminaires sur 4 sujets. Ces derniers n'arrivaient pas à résoudre la tâche de masquage correctement (les mesures ne se stabilisaient pas). Il a donc fallu modifier la méthode de mesure de façon a fournir aux sujets un indice supplémentaire. Ceci a nécessité de reprendre les mesures avec la nouvelle méthode, soit des délais supplémentaires (*un mois environ*).

Un problème de méthode et d'expérience en psychoacoustique s'est révélé en cours d'analyse des données. En fait, une fois les données récoltées, je me suis retrouvé face à de multiples courbes que je ne savais pas vraiment analyser et interpréter. L'erreur, de ma part, a été de ne pas alerter Sophie Savel à temps. Le problème s'est révélé lors d'une réunion (à laquelle assistaient des personnes internes et externes au projet) pendant laquelle je devais présenter mes résultats. Je me suis alors retrouvé face à des diapositives que je n'ai pu clairement analyser. Cette expérience - peu agréable - aura au moins servi à mettre la lumière sur ce problème et prendre les mesures adéquates. Ces dernières ont consisté à reprendre la littérature sur le masquage et nos données expérimentales et à les confronter. Après une période de vacances tout est rentré dans l'ordre. Ce sont alors des délais supplémentaires qui ont été imputés au projet (*4 mois environ*).

Les retards accumulés tout au long de la phase 2 s'élèvent donc à six mois. A cela s'ajoute le temps de rédaction des résultats expérimentaux, qui ont nécessité pas moins de six mois de travail. Une expérience antérieure en psychoacoustique aurait permis de réduire le temps nécessaire à la rédaction mais, en contrepartie de ce retard, j'avoue avoir beaucoup appris de cette étape. La phase de modélisation démarre donc avec un an de retard (avril-mai 2009), ce qui va avoir comme conséquences de devoir faire des concessions par la suite. Tout dépendra du temps qu'il faudra pour proposer une première version de l'algorithme de traitement (estimé à 1,5 mois). L'étalonnage et la validation du modèle seront certainement simplifiés dans le cadre de ma thèse. *Mes travaux serviront de base à la suite du projet.*

J'ai également eu à fonctionner dans le cadre d'une problématique « relationnelle » liée à des divergences de choix de méthode dans la mise en place du protocole expérimental des mesures psychoacoustiques. J'ai été capable, par l'argumentation, l'écoute et la reformulation, de fonctionner comme un pont relationnel entre ces deux chercheurs.

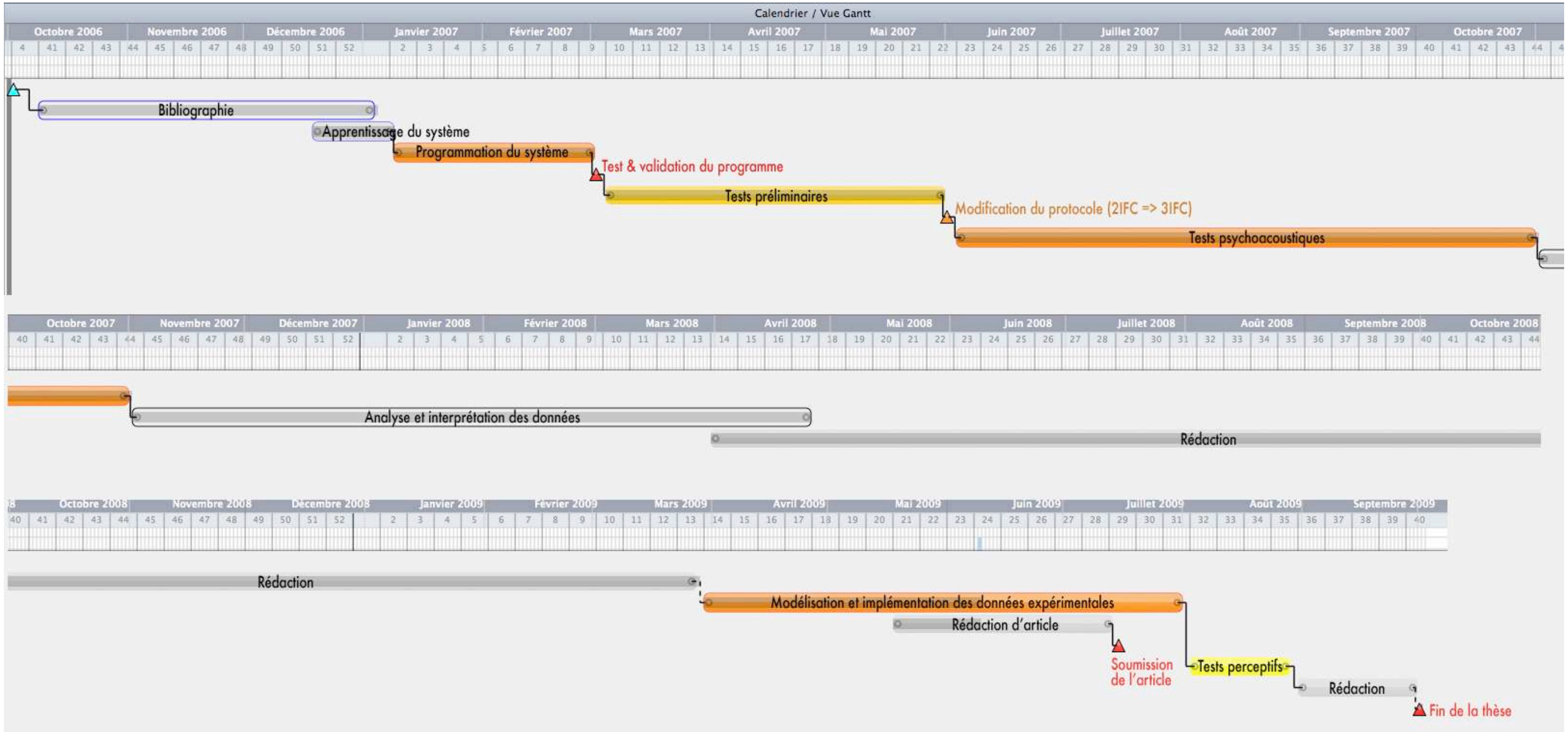


Figure A.2 – Diagramme de Gantt mis à jour, planification réelle du projet.

A.2.4 Evaluation et prise en charge des coûts du projet

Les coûts du projet ainsi que la répartition de leur prise en charge par les différents organismes de financement sont détaillés dans le Tableau A.3. Les figures A.3 et A.4 montrent respectivement la répartition du coût total du projet entre les différents motifs de dépense et la répartition de cette charge financière entre les différentes ressources.

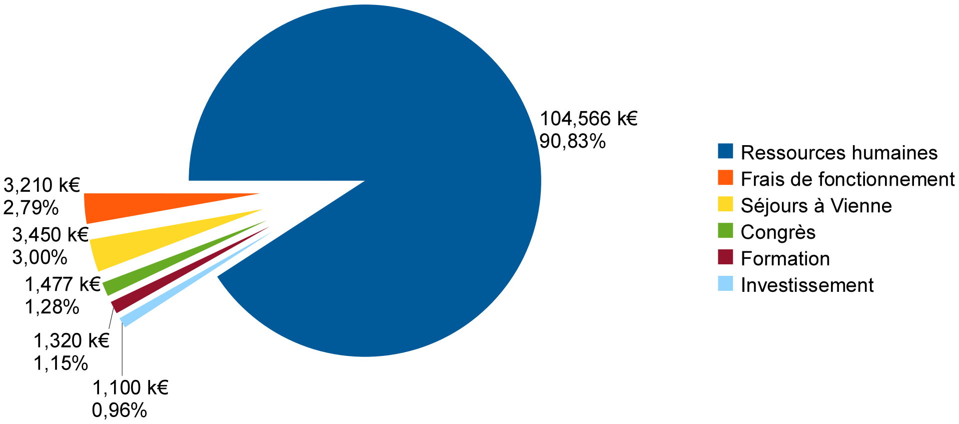


Figure A.3 – Etat de la répartition du coût total (115,123 k€) entre les différents motifs de dépense associés au projet.

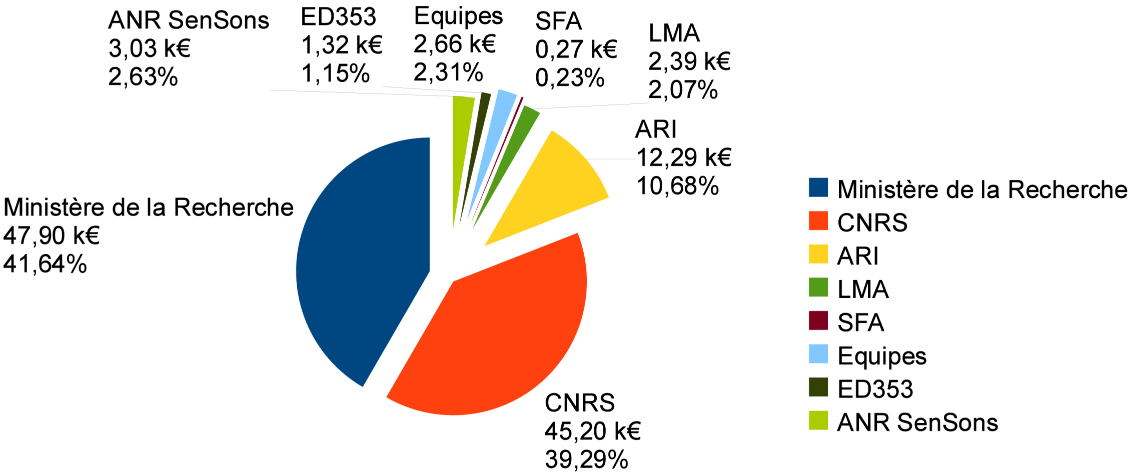


Figure A.4 – Répartition de la charge financière entre les différents organismes de financement.

Ressources humaines					
	Salaire brut	Charges sociales	Utilisation	Coût	Financement
	(€/mois)	(€/mois)	(mois)	(k€)	
Doctorant	1663,22	332,64	36	47,974	Ministère de la Recherche
Encadrants					
Directeur de Recherche	3861,50	772,30	4	12,356	CNRS
Chargée de Recherche	2240,00	448,00	8	14,336	CNRS
Chargé de Recherche	2847,50	569,50	3	6,834	ARI
Chargée de Recherche	2577,83	515,57	3	6,187	CNRS
Chargée de Recherche	2847,50	569,50	3	6,834	CNRS
Chargé de Recherche	2847,50	569,50	2	4,556	ARI
Ingénieur d'Etude	2047,64	409,53	2,5	4,095	CNRS
Secrétaire	1741,86	348,37	1	1,394	CNRS
Sous-total				**104,566**	
Frais de fonctionnement					
Subventions repas		6 € par jour/110 jours		0,660	LMA
Locaux		Services communs		-	CNRS
Equipement psychoacoustique		Frais de maintenance et d'usure		0,700	PA
Licences informatiques					
Matlab®		Jeton utilisé 150 jours		1,725	LMA
KaleidaGraph®		Licence 1 poste		0,065	PA
Bureautique				0,060	PA/S2M
Sous-total				**3,210**	
Séjours à Vienne (AT)					
Octobre 2006		5 jours		0,850	*SenSons*
Juillet 2007		5 jours		0,850	*SenSons*
Octobre 2007		5 jours		0,850	*SenSons*
Septembre 2008		10 jours		0,500	ARI
Novembre 2008		14 jours		0,400	ARI
Sous-total				**3,450**	
Congrès					
JJCAAS'06		Lyon (FR)		0,060	SFA
Journées GPS		Lyon (FR)		0,100	PA
JJCAAS'08		Toulouse (FR)		0,060	SFA
Acoustics'08		Paris (FR)		0,780	PA/SFA
PhD Workshop		Nottingham (UK)		0,477	*SenSons*
Sous-total				**1,477**	
Formation					
Vulgarisation scientifique				0,300	ED353
Programmation C++				0,120	ED353
Nouveau Chapitre de la Thèse				0,900	ED353
Sous-total				**1,320**	
Investissement					
Equipement informatique		Achat d'un ordinateur		1,100	S2M
TOTAL				**115,123**	

Table A.3 – Evaluation et prise en charge des coûts du projet (en k€) par les différents organismes de financement.

A.3 Compétences acquises pendant la thèse

Au-delà de l'aspect « spécialisation scientifique » que procure le doctorat, ce dernier permet d'acquérir et/ou d'affiner de nombreux champs de compétences. Avec un minimum de recul, l'analyse de ces trois années de thèse m'a permis de dégager plusieurs traits caractéristiques de ma personnalité. En parallèle de la thèse je suis également président de l'association METAL C.O.M.M.A.N.D.[3] (loi 1901) organisatrice de concerts et guitariste dans un groupe de métal. Ces deux expériences personnelles liées au grand intérêt que je porte à la musique m'ont aussi beaucoup apporté en termes de compétences et de savoir-faire.

A.3.1 Compétences relationnelles

Le fait de travailler au sein d'une collaboration franco-autrichienne regroupant six chercheurs m'a permis de développer de nombreuses compétences relationnelles, en français et en anglais :

- l'organisation et l'animation de réunions, avec les qualités que cela requiert ; savoir écouter et donner la parole aux participants, respecter les avis de chacun, reformuler et questionner, gérer les conflits,
- l'intégration au sein d'une équipe ; tenir compte de l'avancement des travaux, établir une liaison de confiance avec les co-équipiers, savoir à qui s'adresser en cas de problème,
- gérer aussi les relations avec les ingénieurs et autres personnes extérieures au projet, savoir faire appel à eux en cas de besoin tout en respectant leur domaine de compétences,
- les relations avec les participants aux tests psychoacoustiques (non rémunérés) qui m'ont accordé beaucoup de leur temps. En contrepartie je me devais de leur expliquer les objectifs et principes de ces mesures, les tenir informés de l'avancement du projet auquel ils ont gracieusement participé,
- en parallèle de la thèse, le fait d'avoir été représentant des doctorants du LMA pendant un an m'a appris à bien faire circuler l'information, de la Direction vers les doctorants et vice versa. Dans ce cadre, ma participation aux conseils de laboratoire a, entre autres, renforcé mes compétences dans la tenue de réunion,
- enfin, dans le cadre de l'association j'ai appris à gérer les relations avec les partenaires internes et externes (sponsors).

A.3.2 Communication externe

En tant que président d'une structure organisatrice de concerts, j'ai dû apprendre à :

- communiquer sur l'association et promouvoir les événements organisés : rédaction de communiqués de presse, relations avec les médias et autres structures administratives (mairie, préfecture, DRAC, SACEM. . .),
- rechercher des partenaires et sponsors ; rédiger des dossiers de communication,

[3] **METAL** **C**oncerts **O**rganisation **M**anagement **M**ontpellier - **A**ssociation **N**ovation **D**istribution.

- gérer de l'événementiel : réseau de contacts, organisation, logistique, demandes d'autorisations, location de matériel,
- négocier des tarifs de prestation artistique,
- vendre des produits artistiques en ligne (*site Internet*) et sur des stands.

A.3.3 Compétences scientifiques et techniques

Une majeure partie de cette thèse ayant été consacrée à la psychoacoustique, un domaine que je ne connaissais pas avant d'intégrer ce projet, j'ai donc énormément appris de cette thématique :

- les méthodes psychophysiques de mesure des sensations auditives,
- le fonctionnement du système auditif (physiologie),
- la mise en place de protocoles expérimentaux et toutes les étapes associées (programmation, vérifications, étalonnage, pré-tests, récolte des données. . .),
- l'analyse et l'interprétation de données psychoacoustiques (méthodologie),
- les analyses statistiques,
- la présentation orale et écrite de résultats psychoacoustiques.

La modélisation des données faisant appel aux acquis de ma formation initiale, cette partie n'a fait que renforcer mes compétences dans ce domaine.

Sur le plan technique j'ai acquis de nombreuses compétences liées à la thèse, aux formations proposées par l'ED353 ainsi qu'à mes intérêts personnel (principalement : musique) :

- langages de programmation : Pascal, C++, Action Script, HTML, PHP, LaTeX
- logiciels : Delphi 6, Statistica 5, Xcode, Matlab, Max/MSP, Maple 11, Visual C++
- outils de gestion de projet : Merlin, Projector, Open Proj, Microsoft Project
- programmes d'informatique musicale : Audacity, Guitar Pro, Logic Pro
- environnements : Apple Mac OS X, Microsoft Windows XP/2000

A.3.4 Compétences linguistiques

Le fait de communiquer souvent en anglais de façons orale et écrite a beaucoup contribué à l'approfondissement des connaissances et de la maîtrise de cette langue. Je peux également préciser :

- obtention du « *First Certificate of English* » (FCE) de l'université de Cambridge en 2004,
- notions d'espagnol,
- en cours d'apprentissage du chinois ; séjour d'un mois en Chine en 2008.

A.3.5 Compétences méthodologiques en gestion de projet

En analysant ces trois années de recherche comme un projet cadré sur trois ans, je peux en tirer les compétences acquises suivantes :

- identification des phases relatives à la bonne gestion d'un projet : analyser le contexte, fixer les objectifs et identifier les enjeux et les contraintes,
- importance de l'étude de faisabilité,

- en fonction des ressources disponibles (humaines, matérielles et financières), identifier les étapes nécessaires à la réalisation du projet et en déduire un planning prévisionnel. Cela implique une analyse des risques et des actions correctives associées, avec les conséquences sur les coûts et les délais,
- respecter au maximum les délais,
- préparer un budget prévisionnel (*par exemple*, demande de bourse ou de subvention) et le respecter.

A.3.6 Compétences stratégiques et environnementales

En liaison avec les compétences relationnelles, j'ai développé beaucoup de compétences stratégiques et environnementale telles que :

- une grande capacité d'adaptation face à un système nouveau, découvrir rapidement « qui fait quoi »,
- savoir s'adresser aux personnes les mieux placées dans le système en fonction des besoins pour ne pas perdre trop de temps,
- savoir se créer un réseau de contacts,
- savoir rapidement réajuster sa stratégie face à un problème nouveau ou une situation imprévue,
- savoir percevoir et acquérir les différentes approches et modes d'analyse face à des problèmes d'origines variées ; par exemple psychoacoustique et « signal » dans mon cas,
- les discussions avec les employés et ma participation aux conseils de laboratoire m'ont beaucoup appris de l'organisation interne du CNRS et du code du travail,
- j'ai enfin acquis pas mal d'expérience dans la gestions de tâches administratives, que ce soit en tant que doctorant ou président d'association (dossiers de bourse, subventions, règlement intérieur, co-tutelle, assurance. . .).

A.3.7 Compétences méthodologiques en management

Le fait de travailler en équipe m'a également beaucoup apporté en termes de management :

- animation de réunions, gérer des groupes de travail ou équipes complémentaires,
- gestion des conflits en interne, savoir jouer le rôle de médiateur,
- savoir s'adapter aux susceptibilités des personnes et les analyser en termes d'impact sur les délais,
- prise de décision en accord avec les autres membres de l'équipe (participatif) ou non (directif),
- savoir déléguer,
- travailler dans un univers multi-culturel et pluridisciplinaire,
- dans le cadre de l'association : management d'artistes ; consiste à aider, guider le groupe dans ses démarches de communication et trouver des contrats de prestation.

A.3.8 Savoir-faire pédagogique

Le travail de chercheur ne consistant pas simplement à découvrir des phénomènes mais aussi à les expliquer, j'ai développé des savoir-faire pédagogiques tels que :

- acquérir un esprit de synthèse,
- savoir justifier et argumenter ses choix et actions de recherche,
- acquérir des informations et les inventorier, les classifier de manière à en faire un savoir ; ce qui a notamment été le cas lors de l'étude bibliographique psychoacoustique. Je ne connaissais pas ce domaine et ai dû apprendre à le maîtriser rapidement,
- lors de présentations écrites ou orales, savoir adapter son article ou discours au public ciblé. Ceci nécessite parfois de savoir « vulgariser » ses travaux de recherche,
- savoir transmettre ses compétences ; j'ai notamment pu expérimenter ce savoir-faire lors de vacations à l'Institut Universitaire Technologique de Luminy ou j'ai été chargé de TP (en traitement du signal) pendant deux années consécutives,
- enfin, au sein de l'association j'ai accueilli un stagiaire (école de communication) pour une période de trois mois. Il était chargé de démarcher la presse et de rédiger les communiqués.

A.3.9 Qualités personnelles

A ces compétences j'ajouterai le développement des qualités personnelles suivantes :

- autonomie de travail ; quoique bien encadré, j'ai souvent travaillé seul,
- capacité d'innovation, de création ; j'ai tenté de proposer de nouvelles interfaces pour les tests psychoacoustiques ainsi que pour la modélisation,
- patient ; face aux moments d'hésitation, de remise en question j'ai su temporiser et analyser la situation calmement afin de prendre la décision la plus adéquate,
- capacité d'investissement,
- curieux, aime découvrir des cultures différentes.

A.4 Résultats, impact de la thèse

A.4.1 Pour les laboratoires, pour la Recherche

Tout d'abord, ma thèse a permis de consolider les liens établis entre le LMA et ARI et d'approfondir la collaboration. Actuellement, de nombreuses discussions sont en cours quant à la mise en place de nouveaux projets communs aux deux laboratoires. D'autre part, au sein même du LMA, ma thèse a permis de lier les équipes PA et S2M qui, jusque-là, n'avaient pas collaboré sur un projet d'une telle envergure.

En termes de résultats, l'issue de ma thèse va permettre à tous les acteurs du projet de poursuivre les travaux bien avancés mais non achevés à ce jour. Notamment, le modèle pourra être amélioré (temps de calcul, performances) et évalué de manière plus générale et sur plus de sujets. Une continuité est donc assurée. Le résultat le plus important reste tout de même l'apport d'un modèle de masquage temps-fréquence, non disponible avant mon arrivée. Ce modèle va donner lieu à trois publications (dont deux en cours de rédaction) destinées à des journaux spécialisés (acoustique et traitement du signal). Si elles sont effectivement acceptées et publiées, le LMA et ARI bénéficieront d'une certaine renommée dans les domaines concernés.

D'un point de vue appliqué, ces nouvelles données psychoacoustiques (modèle de masquage) pourront être utilisées dans de nombreuses applications, la plus intuitive étant l'amélioration possible des codeurs audio (type mp3). Si, à terme, ceci s'avérait possible, mes travaux pourraient avoir un impact économique pour les laboratoires (dépôt de brevet) et la Société.

A.4.2 Pour moi-même

Au-delà des nombreuses compétences (voir Sec. A.3) acquises pendant ces trois années, ce projet de thèse m'a fait découvrir le monde de la Recherche et m'a surtout permis de réfléchir quant à mon orientation professionnelle. Après l'obtention de mon Master en traitement du signal et de mon diplôme d'ingénieur ISEN, je disposais de compétences générales en informatique et traitement du signal. La thèse m'a donc permis d'appliquer ces compétences au traitement des signaux audio et de me former en psychoacoustique, me spécialisant ainsi au domaine des sons. Je suis maintenant confronté au choix que tout doctorant doit faire à l'issue de sa thèse, à savoir :

1. Poursuivre dans la voie de la Recherche,
2. S'orienter vers l'enseignement,
3. Retourner dans le milieu de l'industrie.

N'étant pas attiré par l'enseignement, mon choix repose donc sur les solutions (1) et (3). J'avoue que cette décision est un peu complexe car je souhaite rester d'une part très fidèle au domaine de la Recherche (fondamentale et appliquée) et, d'autre part, je ressens aussi le désir de m'impliquer dans le développement de l'industrie audio. A partir de là, et avec l'aide de la réflexion personnelle que constitue le NCT, j'ai pu dégager deux types de profils professionnels adaptés à ma personnalité et à mes désirs. Ces profils constituent en quelque sorte l'issue de ce NCT et sont développés dans la dernière partie de ce document.

A.5 Conclusion : les apports du NCT

A l'issue de ce document, je peux dégager les conclusions suivantes : le NCT m'a permis de prendre énormément de recul par rapport à ces trois années de thèse en la considérant comme un projet à part entière. Notamment, l'analyse des compétences acquises et/ou approfondies pendant la thèse, ainsi que des problèmes rencontrés et de l'attitude adoptée face à ces situations, m'ont permis de mieux cerner ma capacité à travailler et vivre dans un contexte d'équipe et de collaboration. Ainsi, j'ai pu regrouper ces compétences dans deux profils professionnels envisageables après la thèse :

1. Poursuivre vers un post-doctorat me permettrait de me consacrer pleinement à la Recherche et de découvrir un nouvel environnement (laboratoire, situation

géographique, culture. . .). La recherche de ce post-doctorat pourrait s'orienter soit dans la même thématique que la thèse (masquage sonore, modélisation de données psychoacoustiques). Ce sera alors pour moi l'occasion d'approfondir les compétences acquises dans ces domaines. Mais je peux également chercher un post-doctorat dans une thématique connexe mais différente de celle de la thèse, ce qui sera alors l'occasion d'élargir mon champ de compétences, d'acquérir une autre spécialisation. A l'issue de ce post-doctorat, si je ne souhaite pas évoluer dans le domaine de la Recherche publique il me sera possible de rejoindre le domaine privé, à condition d'avoir ciblé au préalable des industriels (ou laboratoires privés) intéressés par mon sujet de post-doctorat.

2. Rejoindre le milieu industriel, et plus particulièrement l'industrie audio. Dans ce milieu, je souhaite (dans la mesure du possible) évoluer dans un contexte d'équipe et d'ambiance conviviale. Au sein de l'entreprise, je souhaite garder contact avec la Recherche, très appliquée dans ce milieu, tout en ayant des responsabilités. A partir de là, je peux identifier deux profils de poste :
   - Ingénieur recherche & développement : me permettrait de conserver une activité de Recherche appliquée et de mettre en œuvre mes compétences scientifiques et techniques, relationnelles, stratégiques ainsi que mon savoir-faire pédagogique (voir § A.3).
   - Chef de projet : bien que plus détaché des activités techniques, ce profil à responsabilité me permettrait de mettre en œuvre mes compétences acquises en gestion de projet, en management et en communication externe ainsi que mes compétences relationnelles, stratégiques et environnementales et mon savoir-faire pédagogique.

Je suis donc en train d'approfondir ces deux parcours et de voir les possibilités qui s'offrent à moi pour chacun d'eux. Notamment, je travaille actuellement à la rédaction de CVs adaptés à chaque profil. Pour m'aider dans cette démarche, je compte fortement sur les conseils de mes directeurs de thèse ainsi que mon mentor NCT, Catherine Echenne-Placa, que je tiens à remercier.

List of Figures

List of Tables

Bibliography

Abolhassani, M. D. and Salimpour, Y. (2008). A human auditory tuning curves matched wavelet function. In *Proceedings of the IEEE International conference on Engineering in Medicine and Biology Society, Vancouver*, pages 2956–2959.

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (2001). Human middle-ear sound transfer function and cochlear input impedance. *Hearing research*, 152(1-2):100–109.

ANSI S3.4 (2007). Procedure for the computation of loudness of steady sounds. Technical Report S3.4–2007, American National Standards Institute, New York.

ANSI S3.6 (1996). Specifications for audiometers. Technical Report S3.6–1996, American National Standards Institute, New York.

Aramaki, M. and Kronland-Martinet, R. (2006). Analysis-synthesis of impact sounds by real time dynamic filtering. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2):695–705.

Aran, J. M. (1988). Electrophysiologie de la cochlée. In *Physiologie de la cochlée*, Série Audition, pages 116–154. INSERM/SFA.

Ashihara, K. (2006). Combination tone: Absent but audible component. *Acoustical Science and Technology*, 27(6):332–335.

Bacon, S. P., Boden, L. N., Lee, J., and Repovsch, J. L. (1999). Growth of simultaneous masking for $f_m < f_s$: Effects of overall frequency and level. *The Journal of the Acoustical Society of America*, 106(1):341–350.

Bacon, S. P., Repovsch-Duffey, J. L., and Liu, L. (2002). Effects of signal delay on auditory filter shapes derived from psychophysical tuning curves and notched-noise data obtained in simultaneous masking. *The Journal of the Acoustical Society of America*, 112(1):227–237.

Bacon, S. P. and Viemeister, N. F. (1985). Simultaneous masking by gated and continuous sinusoidal maskers. *The Journal of the Acoustical Society of America*, 78(4):1220–1230.

Balazs, P., Laback, B., Eckel, G., and Deutsch, W. A. (2010). Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. *IEEE Transactions on Audio, Speech and Language Processing*, 18(1):34–49.

Bian, L. and Chen, S. (2008). Comparing the optimal signal conditions for recording cubic and quadratic distortion product otoacoustic emissions. *The Journal of the Acoustical Society of America*, 124(6):3739–3750.

Bilger, R. C. and Hirsh, I. J. (1956). Masking of tones by bands of noise. *The Journal of the Acoustical Society of America*, 28(4):623–630.

Boullet, I. (2005). *La sonie des sons impulsionnels: perception, mesures et modèles.* Doctorat de mécanique, Université de la Méditerranée Aix-Marseille II, France. (In French).

Carlyon, R. and Moore, B. (1984). Intensity discrimination: A severe departure from weber's law. *The Journal of the Acoustical Society of America*, 76(5):1369–1376.

Carlyon, R. P. (1988). The development and decline of forward masking. *Hearing Research*, 32:65–79.

Cokely, C. G. and Humes, L. E. (1993). Two experiments on the temporal boundaries for the nonlinear additivity of masking. *The Journal of the Acoustical Society of America*, 94(5):2553–2559.

Cooper, N. and Yates, G. (1994). Nonlinear input-output functions derived from the response of guinea-pig cochlear nerve fibers: Variations with characteristic frequency. *Hearing Research*, 78:221–234.

Davis, H. (1983). An active process in cochlear mechanics. *Hearing Research*, 9(1):79–90.

Delgutte, B. (1990). Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 87(2):791–809.

Depalle, P. and Hélie, T. (1997). Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE. Mohonk, NY, USA.

Dolan, T. G. and Small, A. M. J. (1984). Frequency effects in backward masking. *The Journal of the Acoustical Society of America*, 75(3):932–936.

Duifhuis, H. (1973). Consequences of peripheral frequency selectivity for nonsimultaneous masking. *The Journal of the Acoustical Society of America*, 54(6):1471–1488.

Eddins, D. A. and Green, D. M. (1995). Temporal integration and temporal resolution. In Moore, B. C. J., editor, *Hearing*, Handbook of perception and cognition, chapter 6, pages 207–242. Academic Press, San Diego, CA, USA, 2$^{\text{nd}}$ edition.

Eddins, D. A., Hall, J. W., and Grose, J. H. (1992). The detection of temporal gaps as a function of frequency region and absolute noise bandwidth. *The Journal of the Acoustical Society of America*, 91(2):1069–1077.

Egan, J. P. and Hake, H. W. (1950). On the masking pattern of a simple auditory stimulus. *The Journal of the Acoustical Society of America*, 22(5):622–630.

Ehmer, R. H. (1959a). Masking by tones *vs* noise bands. *The Journal of the Acoustical Society of America*, 31(9):1253–1256.

Ehmer, R. H. (1959b). Masking patterns of tones. *The Journal of the Acoustical Society of America*, 31(8):1115–1120.

Elliott, L. L. (1962). Backward and forward masking of probe tones of different frequencies. *The Journal of the Acoustical Society of America*, 34(8):1116–1117.

Elliott, L. L. (1965). Changes in the simultaneous masked threshold of brief tones. *The Journal of the Acoustical Society of America*, 38(5):738–746.

Elliott, L. L. (1967). Development of auditory narrow-band frequency contours. *The Journal of the Acoustical Society of America*, 42(1):143–153.

Fahey, P. and Allen, J. (1985). Nonlinear phenomena as observed in the ear canal and at the auditory nerve. *The Journal of the Acoustical Society of America*, 77(2):599–612.

Fastl, H. (1976). Temporal masking effects: I. Broad band noise masker. *Acustica*, 35(5):287–302.

Fastl, H. (1977). Temporal masking effects: II. Critical band noise masker. *Acustica*, 36(5):317–330.

Fastl, H. (1979). Temporal masking effects: III. Pure tone masker. *Acta Acustica*, 43(5):282–294.

Flandrin, P. (1993). *Temps-fréquence*. Traitement du signal. Hermès, Paris, France. (In French).

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, 12(1):47–65.

Florentine, M. (1986). Level discrimination of tones as a function of duration. *The Journal of the Acoustical Society of America*, 79(3):792–798.

Florentine, M., Fastl, H., and Buus, S. (1988). Temporal integration in normal hearing, cochlear impairment, and impairment simulated by masking. *The Journal of the Acoustical Society of America*, 84(1):195–203.

Garner, W. and Miller, G. (1947). The masked threshold of pure tones as a function of duration. *The Journal of Experimental Psychology*, 37:293–303.

Garner, W. R. (1947). The effect of frequency spectrum on temporal integration of energy in the ear. *The Journal of the Acoustical Society of America*, 19(5):808–815.

Gelfand, S. A. (1998). *Hearing, an introduction to psychological and physiological acoustics*. Marcel Dekker, New York, USA, 3$^{\text{rd}}$ edition.

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138.

Glasberg, B. R. and Moore, B. C. J. (2000). Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *The Journal of the Acoustical Society of America*, 108(5):2318–2328.

Goldstein, J. L. (1967). Auditory nonlinearity. *The Journal of the Acoustical Society of America*, 41(3):676–689.

Grantham, D. W. and Yost, W. A. (1982). Measures of intensity discrimination. *The Journal of the Acoustical Society of America*, 72(2):406–410.

Green, D. M. (1969). Masking with continuous and pulsed sinusoids. *The Journal of the Acoustical Society of America*, 46(4):939–946.

Green, D. M. (1973). Temporal acuity as a function of frequency. *The Journal of the Acoustical Society of America*, 54(2):373–379.

Greenwood, D. D. (1961a). Auditory masking and the critical band. *The Journal of the Acoustical Society of America*, 33(4):484–502.

Greenwood, D. D. (1961b). Critical bandwidth and the frequency coordinates of the basilar membrane. *The Journal of the Acoustical Society of America*, 33(10):1344–1356.

Gröchening, K. (2001). *Foundations of time-frequency analysis*. Applied and numerical harmonic analysis. Birkhaüser, Boston, MA, United States.

Hall, J. L. (1997). Asymmetry of masking revisited: Generalization of masker and probe bandwidth. *The Journal of the Acoustical Society of America*, 101(2):1023–1033.

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. In *IEEE Proc-78*, volume 66, pages 51–83.

Hartmann, W. M. and Wolf, E. M. (2009). Matching the waveform and the temporal window in the creation of experimental signals. *The Journal of the Acoustical Society of America*, 126(5):2580–2588.

He, X. and Scordilis, M. S. (2008). Psychoacoustic music analysis based on the discrete wavelet packet transform. *Research Letters in Signal Processing*, 2008(4):1–5.

Hempstock, T. I., Bryan, M. E., and Tempest, W. (1964). A redetermination of quiet thresholds as a function of stimulus duration. *The Journal of Sound and Vibration*, 1(4):365–380.

Huang, Y.-H. and Chiueh, T.-D. (2002). A new audio coding scheme using a forward masking model and perceptually weighted vector quantization. *IEEE Transactions on Speech and Audio Processing*, 10(5):325–335.

Humes, L. E. and Jesteadt, W. (1989). Models of the additivity of masking. *The Journal of the Acoustical Society of America*, 85(3):1285–1294.

Humes, L. E., Lee, L. W., and Jesteadt, W. (1992). Two experiments on the spectral boundary conditions for nonlinear additivity of simultaneous masking. *The Journal of the Acoustical Society of America*, 92(5):2598–2606.

ISO 532B (1975). Method for calculating loudness level. Technical Report 532B, International Organization for Standardization.

ITU-R BS.1387 (2001). Method for objective measurements of perceived audio quality. Recommendation BS.1387, International Telecommunication Union.

ITU-R BS.1534-1 (2003). Method for the subjective assessment of intermediate quality levels of coding systems. Recommendation BS.1534-1, International Telecommunication Union.

Jeffress, L. A. (1975). Masking of tone by tone as a function of duration. *The Journal of the Acoustical Society of America*, 58(2):399–403.

Jepsen, M., Ewert, S. D., and Dau, T. (2008). A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438.

Jesteadt, W., Bacon, S. P., and Lehman, J. R. (1982). Forward masking as a function of frequency, masker level, and signal delay. *The Journal of the Acoustical Society of America*, 71(4):950–962.

Kidd Jr., G. and Feth, L. L. (1981). Patterns of residual masking. *Hearing Research*, 5:49–67.

Kidd Jr., G. and Feth, L. L. (1982). Effects of masker duration in pure-tone forward masking. *The Journal of the Acoustical Society of America*, 72(5):1384–1386.

Killion, M. C. (1978). Revised estimate of minimum audible pressure: Where is the "missing 6 db"? *The Journal of the Acoustical Society of America*, 63(5):1501–1508.

Laback, B., Balazs, P., Toupin, G., Necciari, T., Savel, S., Meunier, S., Ystad, S., and Kronland-Martinet, R. (2008). Additivity of auditory masking using gaussian-shaped tones. In *Proceedings of the Acoustics'08 meeting*, pages 3889–3894, Paris, France.

Leshowitz, B. (1971). Measurement of the two-click threshold. *The Journal of the Acoustical Society of America*, 49(2):462–466.

Leshowitz, B. and Wightman, F. L. (1971). On-frequency masking with continuous sinusoids. *The Journal of the Acoustical Society of America*, 49(4):1180–1190.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2):467–477.

Liberman, M. C. (1978). Auditory-nerve response from cats raised in low-noise chamber. *The Journal of the Acoustical Society of America*, 63(2):442–455.

Liberman, M. C. and Guinan, J. J. J. (1998). Feedback control of the auditory periphery: Anti-masking effects of middle ear muscles *vs.* olivocochlear efferents. *The Journal of Communication Disorders*, 31(6):471–483.

Lincoln, B. (1998). An experimental high fidelity perceptual audio coder. MUS420 Project, Standford University, Stanford, CA, USA.

Lopez-Poveda, E. A. and Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, 110(6):3107–3118.

Lopez-Poveda, E. A., Plack, C. J., and Meddis, R. (2003). Cochlear nonlinearity between 500 and 8000 hz in listeners with normal hearing. *The Journal of the Acoustical Society of America*, 113(2):951–960.

Lüscher, E. and Zwislocki, J. (1949). Adaptation of the ear to sound stimuli. *The Journal of the Acoustical Society of America*, 21(2):135–139.

Lutfi, R. A. (1988). Interpreting measures of frequency selectivity: Is forward masking special? *The Journal of the Acoustical Society of America*, 83(1):163–177.

Lutfi, R. A. and Patterson, R. D. (1984). On the growth of masking asymmetry with stimulus intensity. *The Journal of the Acoustical Society of America*, 76(3):739–745.

McFadden, D. (1986). The curious half-octave shift: Evidence for a basalward migration of the traveling-wave envelope with increasing intensity. In Salvi, R. J., Henderson, D., Hamernik, R., and Coletti, V., editors, *Basic and applied aspects of noise-induced hearing loss*, pages 295–312. Plenum Publishing, New York, NY, United States.

McFadden, D. and Champlin, C. A. (1990). Reductions in overshoot during aspirin use. *The Journal of the Acoustical Society of America*, 87(6):2634–2642.

Meddis, R. and O'Mard, L. P. (2005). A computer model of the auditory-nerve response to forward-masking stimuli. *The Journal of the Acoustical Society of America*, 117(6):3787–3798.

Moore, B. C. J. (1981). Interactions of masker bandwidth with signal duration and delay in forward masking. *The Journal of the Acoustical Society of America*, 70(1):62–68.

Moore, B. C. J. (1985). Additivity of simultaneous masking, revisited. *The Journal of the Acoustical Society of America*, 78(2):488–494.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing.* Academic Press, London, UK, 5th edition.

Moore, B. C. J., Alcántara, J. I., and Dau, T. (1998). Masking patterns for sinusoidal and narrow-band noise maskers. *The Journal of the Acoustical Society of America*, 104(2):1023–1038.

Moore, B. C. J. and Glasberg, B. R. (1982). Interpreting the role of suppression in psychophysical tuning curves. *The Journal of the Acoustical Society of America*, 72(5):1374–1379.

Moore, B. C. J. and Glasberg, B. R. (1983a). Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise. *The Journal of the Acoustical Society of America*, 73(4):1249–1259.

Moore, B. C. J. and Glasberg, B. R. (1983b). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753.

Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (1988). The shape of the ear's temporal window. *The Journal of the Acoustical Society of America*, 83(3):1102–1116.

Munson, W. and Gardner, M. B. (1950). Loudness patterns – a new approach. *The Journal of the Acoustical Society of America*, 22(2):177–190.

Najaf-Zadeh, H., Lahdili, H., Thibault, L., and Lavoie, M. C. (2003). Use of auditory temporal masking in the MPEG psychoacoustic model 2. In *Proceedings of the 114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Convention paper 5840.

Neff, D. L. (1985). Stimulus parameters governing confusion effects in forward masking. *The Journal of the Acoustical Society of America*, 78(6):1966–1976.

Nizami, L., Reimer, J. F., and Jesteadt, W. (2001). The intensity-difference limen for gaussian-enveloped stimuli as a function of level: Tones and broadband noise. *The Journal of the Acoustical Society of America*, 110(5):2505–2515.

Nizami, L. and Schneider, B. A. (1999). The fine structure of the recovering auditory detection threshold. *The Journal of the Acoustical Society of America*, 106(2):1187–1190.

Nizami, L. and Schneider, B. A. (2000). The periodicity of forward-masked auditory pip-detection thresholds, predicted from the output power of the auditory filter during ringing. *Hearing ResearchH*, 144:168–174.

O'Donovan, J. J. and Furlong, D. J. (2005). Perceptually motivated time-frequency analysis. *The Journal of the Acoustical Society of America*, 117(1):250–262.

Oh, E. and Lutfi, R. A. (1997). Effect of number and frequency spacing of masker components on multitone masking (A). *The Journal of the Acoustical Society of America*, 101(5):3148–3148.

Oxenham, A. J. (2001). Forward masking: Adaptation or integration? *The Journal of the Acoustical Society of America*, 109(2):732–741.

Oxenham, A. J. and Bacon, S. P. (2003). Cochlear compression: Perceptual measures and implications for normal and impaired hearing. *Ear and Hearing*, 24(5):352–366.

Oxenham, A. J. and Plack, C. J. (2000). Effects of masker frequency and duration in forward masking: Further evidence for the influence of peripheral nonlinearity. *Hearing Research*, 150(1–2):258–266.

Painter, T. and Spanias, A. (2000). Perceptual coding of digital audio. In *Proceedings of the IEEE*, volume 88, pages 451–515.

Palmer, A. R. (1995). Neural signal processing. In Moore, B. C. J., editor, *Hearing*, Handbook of perception and cognition, chapter 3, pages 75–121. Academic Press, San Diego, CA, USA, 2$^{\text{nd}}$ edition.

Pang, X. D. and Guinan, J. J. J. (1997). Effects of stapedius-muscle contractions on the masking of auditory-nerve responses. *The Journal of the Acoustical Society of America*, 102(6):3579–3586.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, 59(3):640–654.

Patterson, R. D., Allerhand, M. H., and Giguère, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894.

Patterson, R. D. and Moore, B. C. J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In *Frequency selectivity in hearing*, pages 123–177. Academic Press, London, UK.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *The Journal of the Acoustical Society of America*, 72(6):1788–1803.

Penner, M. J. (1974). Effect of masker duration and masker level on forward and backward masking. *The Journal of the Acoustical Society of America*, 56(1):179–182.

Penner, M. J. (1977). Detection of temporal gaps in noise as a measure of the decay of auditory sensation. *The Journal of the Acoustical Society of America*, 61(2):552–557.

Penner, M. J. (1980). The coding of intensity and the interaction of forward and backward masking. *The Journal of the Acoustical Society of America*, 67(2):608–616.

Pielemeier, W. J. and Wakefield, G. H. (1996). A high-resolution time–frequency representation for musical instrument signals. *The Journal of the Acoustical Society of America*, 99(4):2382–2396.

Plack, C. J. and Moore, B. C. J. (1990). Temporal window shape as a function of frequency and level. *The Journal of the Acoustical Society of America*, 87(5):2178–2187.

Plomp, R. (1964). Rate of decay of auditory sensation. *The Journal of the Acoustical Society of America*, 36(2):277–282.

Plomp, R. (1965). Detectability threshold for combination tones. *The Journal of the Acoustical Society of America*, 37(6):1110–1123.

Plomp, R. and Bouman, M. A. (1959). Relation between hearing threshold and duration for tone pulses. *The Journal of the Acoustical Society of America*, 31(6):749–758.

Pujol, R. (1990). Le traitement du son dans l'oreille interne. *Pour la Science*, 154:20–29.

Rhode, W. S. and Cooper, N. P. (1993). Two-tone suppression and distortion production on the basilar membrane in the hook region of cat and guinea pig cochleae. *Hearing Research*, 66:31–45.

Rioul, O. and Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8:14–38.

Robinson, D. E. and Watson, C. S. (1973). Psychophysical methods in modern psychoacoustics. In Tobias, J. V., editor, *Foundations of modern auditory theory*, volume 2, pages 99–131. Academic Press, New York, USA.

Rodenburg, M. (1977). Investigation of temporal effects with amplitude modulated signals. In Evans, E. F. and Wilson, J. P., editors, *Psychophysics and physiology of hearing*, chapter Temporal factors in hearing, pages 429–437. Academic Press, London, UK.

Rodriguez, J., Neely, S. T., Patra, H., Kopun, J., Jesteadt, W., Tan, H., and Gorga, M. P. (2010). The role of suppression in psychophysical tone-on-tone masking. *The Journal of the Acoustical Society of America*, 127(1):361–369.

Romand, R., Aschoff, A., Ehret, G., de Ribaupierre, F., and Rouiller, E. (1992). *Le système auditif central, anatomie et physiologie.* Série Audition. INSERM/SFA.

Ronken, D. A. (1970). Monaural detection of a phase difference between clicks. *The Journal of the Acoustical Society of America*, 47(2):1091–1099.

Ruggero, M. A. and Rich, N. C. (1991). Furosemide alters organ of corti mechanics: Evidence for feedback of outer hair cells upon the basilar membrane. *The Journal of Neuroscience*, 11(4):1057–1067.

Ruggero, M. A., Rich, N. C., Recio, A., Shyamla, N. S., and Robles, L. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America*, 101(4):2151–2163.

Sachs, M. B. and Kiang, N. Y. S. (1968). Two-tone inhibition in auditory nerve fibers. *The Journal of the Acoustical Society of America*, 43(5):1120–1128.

Savel, S. and Bacon, S. P. (2002). Effectiveness of narrow-band versus tonal off-frequency maskers. *The Journal of the Acoustical Society of America*, 114(1):380–385.

Savel, S. and Bacon, S. P. (2004). Temporal effects in simultaneous masking with on- and off-frequency noise maskers: Effects of signal frequency and masker level. *The Journal of the Acoustical Society of America*, 115(4):1674–1683.

Scharf, B. (1970). Critical bands. In Tobias, J. V., editor, *Foundations of modern auditory theory*, volume 1, pages 159–202. Academic Press, New York, USA.

Shailer, M. J. and Moore, B. C. J. (1983). Gap detection as a function of frequency, bandwidth, and level. *The Journal of the Acoustical Society of America*, 74(2):467–473.

Shailer, M. J. and Moore, B. C. J. (1985). Detection of temporal gaps in bandlimited noise: Effects of variations in bandwidth and signal-to-masker ratio. *The Journal of the Acoustical Society of America*, 77(2):635–639.

Shailer, M. J. and Moore, B. C. J. (1987). Gap detection and the auditory filter: Phase effects using sinusoidal stimuli. *The Journal of the Acoustical Society of America*, 81(4):1110–1117.

Skoglund, J. and Kleijn, W. B. (2000). On time-frequency masking in voiced speech. *IEEE Transactions on Speech and Audio Processing*, 8(4):361–369.

Smith, R. and Zwislocki, J. J. (1975). Short-term adaptation and incremental responses of single auditory-nerve fibers. *Biological Cybernetics*, 17(3):169–182.

Smoorenburg, G. F. (1972). Audibility region of combination tones. *The Journal of the Acoustical Society of America*, 52(2):603–614.

Soderquist, D., Carstens, A., and Frank, G. (1981). Backward, simultaneous, and forward masking as a function of signal delay and frequency. *The Journal of Auditory Research*, 21:227–245.

Soendergaard, P. (2010). The Linear Time-Frequency Analysis Toolbox (LTFAT). Available at http://sourceforge.net/projects/ltfat.

Strickland, E. A. (2001). The relationship between frequency selectivity and overshoot. *The Journal of the Acoustical Society of America*, 109(5):2062–2073.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., and Meddis, R. (2003). Adaptation in a revised inner-hair cell model. *The Journal of the Acoustical Society of America*, 113(2):893–901.

Swets, J. A., Green, D. M., and Tanner, W. P. J. (1962). On the width of critical bands. *The Journal of the Acoustical Society of America*, 34(1):108–113.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, 1:155–182.

Thornton, A. (1972). PSM studies II. Post-stimulatory masked thresholds as a function of probe tone duration. *Journal of Sound and Vibration*, 22(2):183–191.

Tooley, M. H. (2006). *Electronic circuits: Fundamentals and applications*. Newnes, Oxford, UK, 3$^{\text{rd}}$ edition.

Vafin, R., Andersen, S. V., and Kleijn, W. B. (2000). Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'00)*, volume 2, pages 901–904, Istanbul, Turkey.

van de Par, S., Koppens, J., Kohlrausch, A., and Oomen, W. (2008). A new perceptual model for audio coding based on spectro-temporal masking. In *Proceedings of the 124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Convention paper 7336.

van den Brink, W. A. C. and Houtgast, T. (1990). Spectro-temporal integration in signal detection. *The Journal of the Acoustical Society of America*, 88(4):1703–1711.

van Schijndel, N. H., Houtgast, T., and Festen, J. M. (1999). Intensity discrimination of Gaussian-windowed tones: Indications for the shape of the auditory frequency-time window. *The Journal of the Acoustical Society of America*, 105(6):3425–3435.

Vetterli, M. and Kovačević, J. (1995). *Wavelets and subband coding*. Prentice Hall PTR, Englewood Cliffs, New Jersey.

Viemeister, N. F. (1977). Temporal factors in audition: A system analysis approach. In Evans, E. F. and Wilson, J. P., editors, *Psychophysics and physiology of hearing*, chapter Temporal factors in hearing, pages 419–428. Academic Press, London, UK.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, 66(5):1364–1380.

Viemeister, N. F. and Plack, C. J. (1993). Time analysis. In Yost, W. A., Popper, A. N., and Fay, R. R., editors, *Human psychophysics*, volume 3 of *Handbook of Auditory Research*, pages 75–121. Springer-Verlag, 2$^{\text{nd}}$ edition.

Viemeister, N. F. and Wakefield, G. (1991). Temporal integration and multiple looks. *The Journal of the Acoustical Society of America*, 90(2):858–865.

Vogten, L. L. M. (1978a). Low-level pure-tone masking: A comparison of "tuning curves" obtained with simultaneous and forward masking. *The Journal of the Acoustical Society of America*, 63(5):1520–1527.

Vogten, L. L. M. (1978b). Simultaneous pure-tone masking: The dependence of masking asymmetries on intensity. *The Journal of the Acoustical Society of America*, 63(5):1509–1519.

von Békésy, G. (1960). *Experiments in hearing.* McGraw Hill, Oxford, UK.

Weber, D. L. and Moore, B. C. J. (1981). Forward masking by sinusoidal and noise maskers. *The Journal of the Acoustical Society of America*, 69(5):1402–1409.

Wegel, R. L. and Lane, C. E. (1924). The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physical Review*, 23(2):266–285.

Widin, G. P. and Viemeister, N. F. (1979a). Intensive and temporal effects in pure-tone forward masking. *The Journal of the Acoustical Society of America*, 66(2):388–395.

Widin, G. P. and Viemeister, N. F. (1979b). Short-term spectral effects in pure-tone forward masking. *The Journal of the Acoustical Society of America*, 66(2):396–399.

Wright, B. A. (1997). Detectability of simultaneously masked signals as a function of masker bandwidth and configuration for different signal delays. *The Journal of the Acoustical Society of America*, 101(1):420–429.

Wright, B. A. and Dai, H. (1994). Detection of unexpected tones with short and long durations. *The Journal of the Acoustical Society of America*, 95(2):931–938.

Yasin, I. and Plack, C. J. (2005). The role of suppression in the upward spread of masking. *The Journal of the Association for Research in Otolaryngology*, 6(4):368–377.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen) (L). *The Journal of the Acoustical Society of America*, 33(2):248.

Zwicker, E. (1965). Temporal effects in simultaneous masking and loudness. *The Journal of the Acoustical Society of America*, 38(1):132–141.

Zwicker, E. (1984). Dependence of post-masking on masker duration and its relation to temporal effects in loudness. *The Journal of the Acoustical Society of America*, 75(1):219–223.

Zwicker, E. (1985). Suppression and $(2f_1-f_2)$-difference tones in a nonlinear cochlear preprocessing model with active feedback. *The Journal of the Acoustical Society of America*, 80(1):163–176.

Zwicker, E. and Fastl, H. (1972). On the development of the critical band. *The Journal of the Acoustical Society of America*, 52(2):699–702.

Zwicker, E. and Feldtkeller, R. (1999). *The ear as a communication receiver.* Acoustical Society of America, New York, USA, 2$^{\text{nd}}$ edition.

Zwicker, E. and Jaroszewski, A. (1982). Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels. *The Journal of the Acoustical Society of America*, 71(6):1508–1512.

Zwislocki, J. (1960). Theory of temporal auditory summation. *The Journal of the Acoustical Society of America*, 32(8):1046–1060.

Zwislocki, J., Pirodda, E., and Rubin, H. (1959). On some poststimulatory effects at the threshold of audibility. *The Journal of the Acoustical Society of America*, 31(1):9–14.