TROISIÈME PARTIE

Analyse de scènes auditives

Chapitre 5

Analyse de scènes auditives computationnelle

5.1. Introduction

Jusqu'à une époque récente, l'audition s'intéressait à la perception de qualités telles que la hauteur, la sonie, le timbre, etc., d'un son émis par une *source unique*. L'expérimentation psychoacoustique a mis en évidence les relations entre les caractéristiques physiques du son et les sensations qu'il évoque, et permis d'entrevoir les mécanismes physiologiques qui font passer de l'un à l'autre. Des modèles de traitement auditif ont été élaborés, qui opèrent à partir de l'onde acoustique ou de son spectre.

Malheureusement, les sources qui nous entourent émettent rarement de façon isolée. Nous évoluons dans une cacophonie de voix, sons et bruits superposés, dont le spectre collectif est bien différent de celui d'une source unique. Chaque oreille reçoit des ondes provenant d'une multitude de sources. Néanmoins, on peut souvent porter son attention sur une source particulière et juger de sa sonie, de sa hauteur, de son timbre, voire comprendre ce qui est dit lorsqu'il s'agit de parole, même en la présence de sons concurrents. Les modèles classiques, conçus pour traiter une source isolée, ne sont pas suffisants pour expliquer la perception de sources multiples.

Helmholtz déjà se demandait comment on pouvait percevoir les qualités individuelles d'instruments qui jouent ensemble [HEL 77]. Mais il faut attendre le travail de Bregman pour que l'*analyse de scènes auditives* (Auditory Scene Analysis, ou ASA) devienne un sujet d'étude à part entière [BRE 90]. Pour Bregman, le problème de l'émergence de sources subjectives (flux, ou streams) est primordial, puisqu'elle précède logiquement la détermination de leurs qualités individuelles. L'ASA de Bregman

---

Chapitre rédigé par Alain DE CHEVEIGNÉ.

est une transposition dans le domaine de l'audition des principes de l'analyse de scènes visuelles.

Figure 5.1. *Analyse de scène visuelle. A gauche, les fragments sont inorganisés. A droite, la présence d'une forme masquante permet leur regroupement perceptif. L'ASA cherche des principes analogues pour l'organisation du monde sonore (d'après [BRE 90]).*

Avec le développement de l'informatique et de l'intelligence artificielle sont apparues des tentatives d'*analyse de scène auditive computationnelle*(CASA) [BRO 92a, COO 91, ELL 96, LYO 83, MEL 91, WAN 95, WEI 85]. Les modèles CASA ont la double ambition d'aider à comprendre les processus perceptifs et de résoudre des problèmes pratiques, par exemple éliminer le bruit dans un système de reconnaissance de la parole. L'influence de la vision computationnelle, notamment les travaux de Marr [MAR 82], a joué un rôle déterminant.

La notion de *modèle CASA* souffre d'une certaine ambiguïté. En ce qui concerne la modélisation de processus perceptifs, il n'est pas aisé de situer la frontière entre les modèles CASA et les autres modèles, d'autant que la modélisation computationnelle est devenue commune dans de nombreux domaines. En tant que méthode de traitement du signal, la spécificité ou les avantages des modèles CASA par rapport à d'autres techniques ne sont pas toujours évidents. En cherchant à être *et* un bon modèle auditif *et* une technique utile, le modèle CASA court le risque de n'être ni l'un ni l'autre. Néanmoins, l'approche CASA peut être fructueuse à condition de bien différencier ses rôles de modèle et méthode, notamment au moment de leur évaluation. L'insistance à construire un système complet (et donc complexe) est un bon antidote à la dérive réductionniste des modèles psychoacoustiques. Du point de vue pratique, les applications telles que la reconnaissance de la parole ont grand besoin de reproduire les capacités de tolérance au bruit du système auditif. Des développements intéressants sont issus récemment de l'approche CASA, en particulier la *théorie des données manquantes* (*Missing Feature Theory*) [COO 94, COO 97, LIP 97, MOR 98].

5.2. Principes de ASA

5.2.1. *Fusion versus scission : le choix d'une représentation*

En ASA, on a l'habitude de parler de *groupement* (ou fusion) et *séparation* (ou scission) de traits acoustiques. En cas de fusion, les traits sont attribués à une même source ou *flux sonore*, en cas de scisson ils sont répartis entre plusieurs sources. Pour que ces mots aient un sens, il faut supposer une représentation interne peuplée d'indices du monde sonore, dans laquelle les indices de chaque source sont séparables de ceux des autres sources. On peut entendre par là une représentation du stimulus physique dans le domaine temps, le domaine fréquence, ou l'une des nombreuses représentations temps-fréquence. On peut aussi se référer à une représentation physiologique – canaux fréquentiels issus de la cochlée, réseau de coïncidence neuronal, etc. – dans laquelle le système auditif puiserait des éléments à attribuer à chaque source.

Les psychoacousticiens emploient, en fait, une troisième représentation lorsqu'ils décrivent un stimulus en termes de paramètres de synthèse (durée, amplitude, fréquence ou phase instantanée de chaque composante). Ce n'est pas vraiment une représentation temps-fréquence au sens du traitement du signal, puisque aucune représentation de ce type ne permet une description aussi précise sur les deux axes temps et fréquence. A titre d'exemple, imaginons un stimulus contenant plusieurs composantes sinusoïdales modulées en fréquence. Au moment de la synthèse, la fréquence instantanée est parfaitement spécifiée, mais il n'existe pas de méthode générale pour retrouver ces paramètres à partir du stimulus. Une analyse temps-fréquence pourra fournir une estimation approchée, mais elle n'est pas unique, et en tout cas pas exactement conforme à la description idéalisée du psychoacousticien.

C'est là une source de confusion considérable. Les *principes de l'ASA* ont été énoncés par les psychoacousticiens en termes de paramètres de synthèse. Le modèle CASA, lui, n'a pas accès à cette représentation idéalisée, et doit se contenter de ce qu'il peut extraire du signal. Nombre de « bonnes idées » en termes d'une représentation idéalisée se dégonflent lorsqu'on les applique dans la pratique. C'est l'un des mérites de l'approche CASA que de révéler ces difficultés.

5.2.2. *Traits de groupement simultané*

En gardant à l'esprit la mise en garde précédente, considérons un stimulus « constitué » d'un certain nombre de composantes. On pourrait s'attendre à ce que le système auditif les attribue à la même source, comme ferait un sonomètre ou un système de reconnaissance de la parole. Notre expérience nous prouve que ce n'est pas toujours le cas : on peut souvent « séparer les composantes » du stimulus et en attribuer une partie à chaque source. Se pose alors la question suivante : puisque dans certains cas les composantes de sources distinctes sont séparables (scission), qu'est-ce qui parfois les

retient ensemble pour représenter une même source (fusion) ? Fusion et scission sont les deux faces d'une même pièce. Quels traits acoustiques favorisent l'une ou l'autre ?

Harmonicité. Une relation harmonique entre composantes favorise leur fusion. C'est le cas lorsque le stimulus est périodique (parole voisée, certains sons d'instruments). Dans le cas contraire (la « polypériodicité » de [MAR 91]), le stimulus paraît contenir plusieurs sources. Des voyelles ou voix concurrentes sont plus faciles à comprendre si elles suivent des séries harmoniques distinctes, c'est-à-dire si leurs fréquences fondamentales (F_0) sont différentes.

Cohérence d'enveloppe, synchronicité d'attaque. Si des partiels démarrent ensemble et leur amplitude évolue de façon cohérente, ils tendent à fusionner. Une asynchronie d'attaque favorise au contraire la scission. C'est un exemple du principe plus général de *destin commun*.

Corrélation binaurale. Si les composantes d'une source ont toutes la même relation binaurale, leur fusion est favorisée. Une différence de relation binaurale entre cible et masqueur favorise la perception de la cible.

Modulation cohérente de fréquence. Il s'agit d'un autre exemple du principe de destin commun. Si l'on imagine une représentation spectro-temporelle de façon graphique, des composantes dont la modulation est cohérente devraient former une « figure », et se distinguer de composantes immobiles ou dont la modulation serait incohérente.

Tous ces traits ont été proposés et implémentés avec plus ou moins de bonheur dans des systèmes CASA.

5.2.3. *Traits de groupement séquentiel*

Comme pour le groupement simultané, on pourrait imaginer que les sons qui se suivent au cours du temps soient toujours attribués à la même source (fusion). Il n'en est rien : dans certains cas le système auditif divise une succession de sons en plusieurs flux distincts (scission). Chaque flux semble alors évoluer de façon indépendante. Chacun peut être choisi et « isolé » par l'attention. On peut distinguer l'ordre des sons à l'intérieur d'un flux, mais pas d'un flux à l'autre. Ce phénomène est exploité dans les fugues de Bach pour créer plusieurs lignes mélodiques par le jeu d'un seul instrument. Parmi les traits qui déterminent fusion et scission, on note :

– la proximité fréquentielle. Une succession de sons purs dont les fréquenes sont proches tendent à fusionner en un flux. Elles forment des flux distincts si les fréquences sont éloignées ;

– le caractère répétitif. La tendance à la séparation est renforcée par la durée et le caractère répétitif des stimuli ;

– le taux de répétition. La présentation d'une succession de sons à un rythme rapide favorise leur scission. Le ralentissement du rythme favorise la fusion ;

– la similarité de timbre. Une succession de sons de même timbre tend à fusionner. Des sons de timbre très différent ont du mal à fusionner, et il est difficile de distinguer leur ordre temporel.

Sur la base de cette liste, on pourrait s'attendre à ce que les nombreuses discontinuités d'amplitude, timbre, etc., de la parole l'empêchent d'être perçue comme un flux cohérent. Paradoxalement, il n'en est rien : une voix garde sa cohérence malgré ces discontinuités.

5.2.4. *Schémas*

Les traits précédents, qui dépendent du signal, relèvent de ce que l'on appelle le groupement *primitif*. Les mécanismes de groupement primitif sont automatiques et involontaires, et ne dépendent pas de l'apprentissage ou du contexte cognitif. Mais il existe aussi des situations où le groupement s'appuie sur des *schémas* appris, sur des régularités abstraites, ou sur l'état d'esprit du sujet. La distinction primitif/schéma est à rapprocher de celle entre processus *bottom-up* et *top-down* en intelligence artificielle.

5.2.5. *Illusion de continuité, restauration phonémique*

Lorsque l'on superpose un bruit court à un ton continu, le ton semble continuer « derrière » le bruit. Il en est de même si le ton est interrompu pendant le bruit, à condition que ce dernier soit assez fort. C'est l'*illusion de continuité*. Le même phénomène se produit avec de la parole. Si l'on remplace un phonème par un bruit assez fort, le phonème absent est perçu comme présent. C'est la *restauration phonémique*. Le phonème « restauré » peut varier selon le contexte (par exemple le stimulus « *eel » devient « wheel », « peel », « meal », etc., selon le contexte sémantique). Chose curieuse, une fois la séquence restaurée, il est presqu'impossible de dire lequel parmi ses phonèmes était manquant.

5.3. Principes de CASA

5.3.1. *Création d'une représentation*

L'analogie avec l'analyse de scène visuelle, sur laquelle s'appuie l'ASA, suppose l'existence d'une « representation » d'une richesse comparable à l'espace 3-D des objets ou 2-D des images (Marr utilise le terme de 2 1/2-D pour qualifier la représentation enrichie fournie par la vision binoculaire et autres mécanismes de perception de la profondeur [MAR 82]). L'onde acoustique étant de dimensionnalité faible, le modèle CASA commence par synthétiser une représentation plus riche.

5.3.1.1. *Filtre cochléaire*

Le modèle CASA typique commence par un banc de filtres. En principe, ils se veulent conformes à ce que l'on sait du filtrage cochléaire. En pratique, il y a une grande diversité selon que le concepteur aura privilégié un modèle physique de cochlée, la conformité aux données psychophysiques ou physiologiques, la facilité d'implémentation, etc. Actuellement, le filtre le plus populaire est du type *gammatone*, réaliste et facile à implémenter [COO 91, HOL 88, PAT 92, SLA 93]. Les filtres sont généralement de largeur constante (en Hz) jusqu'à 1 kHz et de largeur proportionnelle à leur fréquence centrale au-delà. Un délai supplémentaire est souvent ajouté aux sorties des canaux pour compenser les différences de délai de groupe et « aligner » les réponses impulsionnelles.

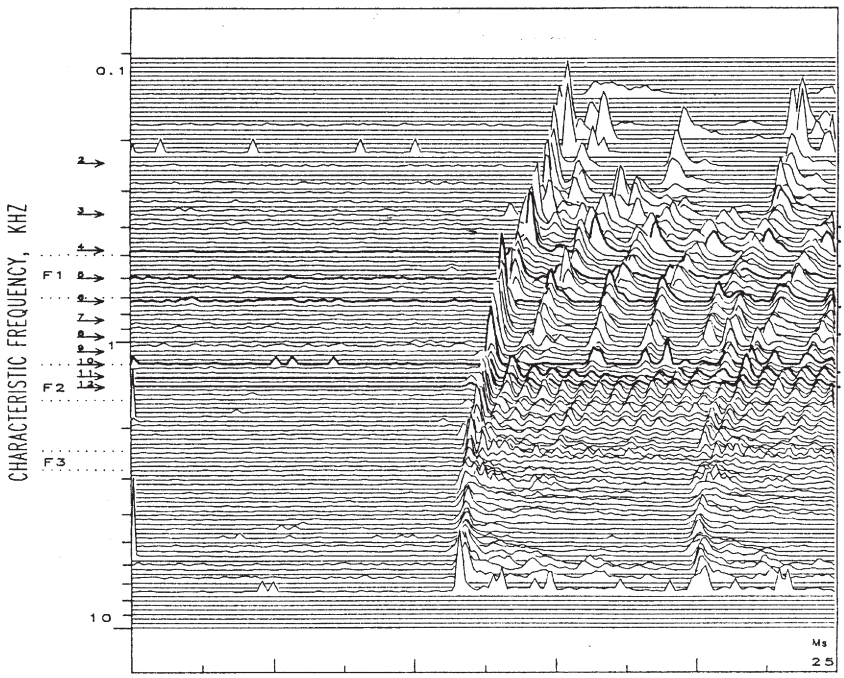

Figure 5.2. *Activité d'une population de fibres du nerf auditif du chat en réponse à la syllabe synthétique « da ». Les modèles de filtrage et transduction cochléaire cherchent à reproduire ce type de réponse (en faisant l'économie du chat). Le retard progressif des canaux de basse fréquence (en haut), dû au temps de propagation dans la cochlée, est souvent compensé dans les modèles (d'après [SHA 85]).*

5.3.1.2. *Transduction*

La vibration mécanique de la membrane basilaire détermine la *probabilité* de décharge des fibres du nerf auditif qui innervent les cellules ciliées internes.

Ce processus est modélisé de façon plus ou moins réaliste selon les modèles :

– une probabilité étant positive, la transduction a des propriétés proches d'un redresseur simple alternance ;

– elle a aussi des propriétés compressives, qu'on peut modéliser par une simple non-linéarité instantanée (log, racine cubique, etc.), ou par un mécanisme adaptatif : commande automatique de gain [HOL 90, LYO 82, LYO 84, PAT 92, SEN 85] ou modèle de cellule ciliée [MED 86, MED 88]) ;

– dans les modèles de Lyon et de Holdsworth, le gain de chaque canal varie en fonction de l'activité dans une région temporelle (passé récent), et spectrale (canaux voisins). Cette dernière propriété n'a pas de justification physiologique au niveau périphérique, mais elle a l'effet bénéfique de renforcer le contraste de la représentation le long de la dimension spectrale (c'est un exemple de confusion entre modèle et méthode). D'autres modèles vont plus loin et incorporent un mécanisme explicite de différentiation spectrale et/ou temporelle, dont un exemple est le LIN (*Lateral Inhibitory Network*) de [SHA 85] ;

– la transduction non linéaire est généralement suivie d'un filtrage passe-bas (lissage temporel). Selon les modèles, ce filtrage est soit léger (faible constante de temps) pour représenter la perte de synchronisation que l'on observe physiologiquement à hautes fréquences (entre 1 et 5 kHz), soit plus sévère de façon à éliminer la structure périodique de la parole voisée et obtenir une estimation du spectre stable au cours du temps.

La sortie du module filtre/transduction peut être vue, soit comme une succession de spectres à court terme, soit comme un ensemble de canaux parallèles portant chacun une version filtrée du signal. La représentation est de dimensionnalité élevée, premier pas vers un substrat propice à l'analyse de scènes.

5.3.1.3. *Affinage du pattern spectro-temporel*

Néanmoins, la sortie du module filtre/transduction n'a pas les caractéristiques idéales de la représentation qui a servi à la synthèse (voir paragraphe 5.2.1). Par rapport à cette représentation idéalisée, elle peut sembler manquer de résolution fréquentielle, ou temporelle, ou les deux. On a cité le LIN de Shamma qui renforce le contraste spectral [SHA 85]. Deng propose la corrélation croisée entre canaux voisins pour renforcer la représentation des formants [DEN 88]. Les *synchrony strands* de Cooke produisent une représentation proche d'une somme de sinusoïdes, propice à l'application des principes ASA (continuité de chaque *strand*, destin commun, harmonicité, etc.) [COO 91]. Ces techniques peuvent s'interpréter comme des tentatives d'extraire du signal une représentation proche de celle, idéalisée, qu'utilisent les psychoacousticiens pour énoncer les principes de l'ASA.

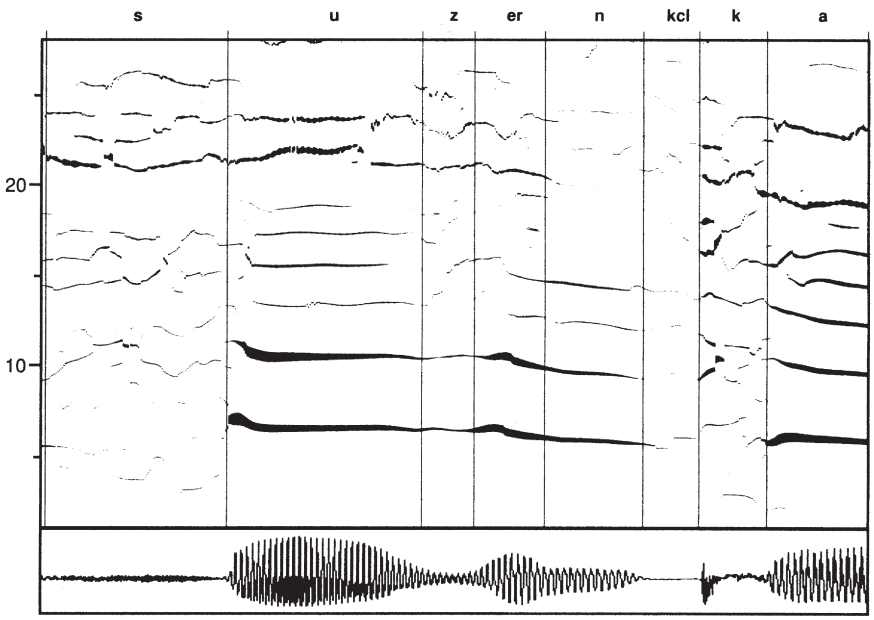

Figure 5.3. *Représentation de type «* synchrony strand *» en réponse à une portion de parole (en bas). Dans les basses fréquences chaque «* strand *» correspond à une harmonique, dans les hautes fréquences il correspond à un formant (d'après [COO91]).*

5.3.1.4. *Dimensions supplémentaires*

Si le lissage temporel n'est pas trop sévère, la *structure temporelle* de chaque canal issu du module filtre/transduction permet d'enrichir la représentation de dimensions supplémentaires. Lyon, en s'inspirant du modèle d'interaction binaurale de Jeffress, a proposé de calculer la fonction de *corrélation croisée* entre canaux issus des deux oreilles [LYO 83]. Cette représentation présente une dimension supplémentaire par rapport à une représentation temps-fréquence : le délai interaural. Des maxima peuvent apparaître à différentes positions le long de cet axe, correspondant aux azimuts des différentes sources. Lyon échantillonne la représentation, par coupes parallèles à l'axe des fréquences, pour isoler telle ou telle source [LYO 83]. Des tentatives similaires ont été faites depuis [BOD 96, PAT 96].

Une autre dimension apparaît si on calcule la fonction d'*autocorrélation* dans chaque canal. Cette idée a été proposée à l'origine par Licklider pour estimer la période dans un modèle de perception de la hauteur [LIC 59]. En réponse à un stimulus périodique (tel que de la parole voisée), des maxima surgissent à des positions correspondant à la période du son, ou à des multiples de cette période. Ce principe peut être exploité pour séparer les corrélats de voix concurrentes. En réponse à *plusieurs* stimuli périodiques (voix), certains canaux pourront être dominés par une voix, d'autres par une

autre. En sélectionnant les canaux selon les périodes qui les dominent, on peut isoler les voix. Proposée par Weintraub [WEI 85], cette idée a été reprise par Mellinger [MEL 91], Meddis et Hewitt [MED 92], Brown [BRO 92a], Lea [LEA 92] et Ellis [ELL 96].

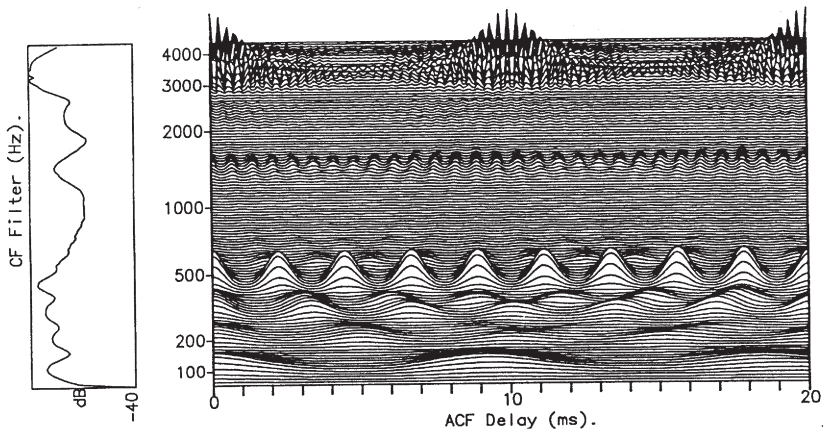

Figure 5.4. *Pattern d'autocorrélation, en réponse à un mélange de voyelles (/i/ à 100 Hz et /o/ à 112 Hz). Chaque trait correspond à un canal issu du filtre périphérique. Chaque canal sera affecté à une voyelle en fonction de la périodicité qui le domine (d'après [LEA 92]).*

L'autocorrélation analyse chaque canal avec une résolution temporelle fine, capable de résoudre la périodicité des formants de la parole. Une résolution aussi fine n'est pas toujours utile, d'autant que la structure fine reflète aussi la résonance des filtres cochléaires, qui ont peu à nous dire sur le signal. Un *lissage temporel*permet de se débarrasser de la structure fine et ne retenir (on l'espère) que les modulations qui reflètent la période fondamentale. Celles-ci peuvent alors être évaluées par autocorrélation ou par d'autres méthodes : passages par zéro [COO 91], transformée de Fourier [MEY 96, MEY 97]. Le *spectre de modulation* de paramètres (physiologiques, LPC, cepstraux, etc.) considérés comme suites temporelles est l'objet de beaucoup d'intérêt récemment, notamment en reconnaissance de la parole [GRE 96, HER 94, KAN 98, NAD 97].

D'autres transformations sont la carte de transition fréquentielle (frequency transition map) de Brown ou les onset maps de Mellinger, Brown ou Ellis, qui ont pour but de repérer les changements temporels brusques pouvant signaler le début d'un son [BRO 92a, ELL 96, MEL 91].

Chaque dimension supplémentaire enrichit la représentation. Si la pression acoustique à une oreille est fonction d'*une* dimension (le temps), l'ensemble des canaux périphériques est fonction de *deux* (temps, fréquence). Avec la corrélation binaurale et

l'autocorrélation (ou spectre de modulation) on arrive à *quatre* dimensions : temps, fréquence, délai interaural, fréquence de modulation. Cette « explosion dimensionnelle » est motivée par l'espoir que les indices de sons concurrents seront *séparables* si la dimensionnalité est suffisamment élevée.

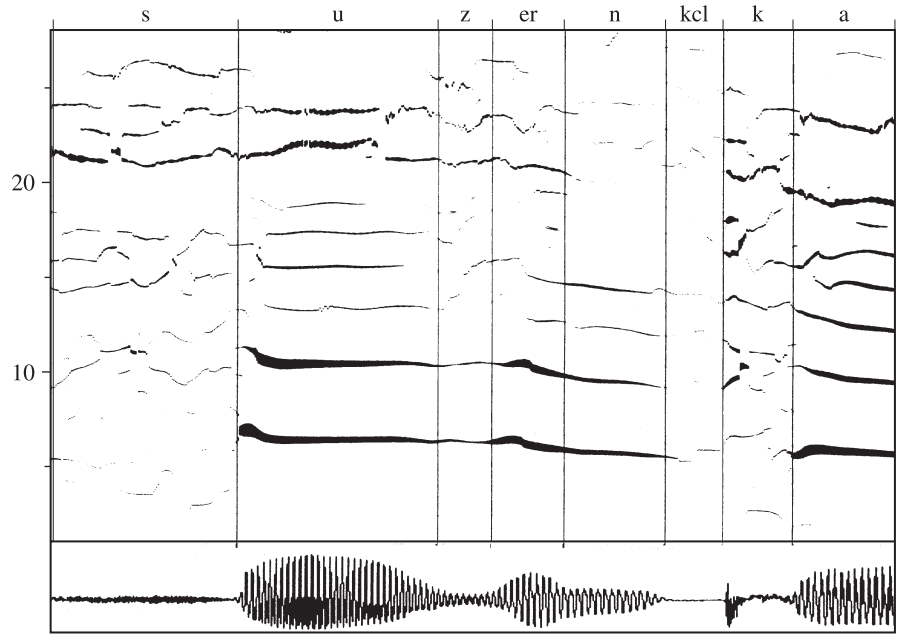

Figure 5.5. *Carte de transition fréquentielle en réponse à de la parole (même portion que pour Cooke, ci-dessus). Les flèches indiquent l'orientation estimée par un banc de filtres d'orientation spectro-temporels (d'après [BRO 93]).*

5.3.1.5. *Abstractions élémentaires*

La plupart des modèles CASA démarrent avec une représentation riche et peu contrainte (paragraphe précédent), et tentent ensuite d'organiser l'information en objets élémentaires, par exemple en suivant les principes ASA. Les synchrony strands de Cooke [COO 91] sont le résultat de l'application d'une contrainte de continuité temporelle aux composantes de la représentation spectrale. Le principe de groupement par harmonicité se traduit par les periodicity groups de Cooke et Brown [COO 92] ou les wefts (trames) de Ellis [ELL 96]. Le principe de synchronicité d'attaque est utilisé par Brown pour former les objets auditifs [BRO 92a].

5.3.1.6. *Organisation d'ordre supérieur*

L'organisation se poursuit de manière hiérarchique, en principe jusqu'à la partition de toute l'information en sources. Certains modèles utilisent un processus purement

ascendant (data driven), d'autres revendiquent une stratégie plus complexe (top-down), faisant appel à des techniques d'intelligence artificielle [ELL 96, GOD 97, KAS 97, NAK 97]. L'inconvénient de stratégies complexes est double : elles sont opaques, et elles tendent à réagir de manière « catastrophique » (dans le sens ou une petite perturbation des conditions à l'entrée du système peut produire un grand changement de son état). Elles sont néanmoins indispensables pour gérer l'ensemble des sources d'informations et hypothèses qui interviennent dans l'organisation d'une scène auditive.

5.3.1.7. *Schémas*

La plupart des systèmes CASA sont du type *data-driven* et s'appuyant sur des principes ASA de type *primitif*. Les approches du type *top-down*, s'appuyant sur des principes ASA de type *schémas* sont plus rares. A signaler la proposition d'Ellis d'utiliser un système de reconnaissance de la parole pour guider l'analyse de scène auditive [ELL 97]. Lorsque la partie parole de la scène est reconnue, les limites de sa contribution à la scène peuvent être précisées, et le reste de la scène analysé de façon plus fine.

5.3.1.8. *Le problème des composantes partagées*

Quelle que soit la richesse et la dimensionnalité de la représentation de base, il arrive que l'appartenance d'un élément soit ambiguë. Les stratégies divergent selon qu'on décide alors de l'attribuer à une seule des sources (principe d'allocation exclusive), aux deux (attribution duplex) ou à aucune. On peut aussi essayer de *scinder* l'élément, par exemple selon des critères de continuité fréquentielle ou temporelle [WEI 85]. D'un certain point de vue, une telle scission est un aveu d'échec de la représentation, qui a échoué à partitionner l'information acoustique en éléments atomiques attribuables à chaque source.

5.3.1.9. *Le problème des composantes manquantes*

Des raisons théoriques (que malheureusement la pratique confirme) nous disent qu'il est impossible d'aboutir à une séparation parfaite dans tous les cas. Par exemple, des composantes trop proches en fréquence seront confondues et attribuées à une source au détriment de l'autre. De telles portions, masquées ou d'appartenance incertaine, manqueront à la représentation d'une source séparée. Il y a deux façons d'aborder le problème :

1) recréer l'information manquante par interpolation ou extrapolation à partir du contexte acoustique ou cognitif [ELL 96, MAS 97] ;

2) marquer la portion comme manquante, et l'ignorer dans la suite des opérations, par exemple en l'affectant d'un poids nul lors de la reconnaissance de formes [COO 97, LIP 98, MOR 98].

La première, parfois motivée par une interprétation un peu trop littérale de la notion de *restauration phonémique*, se justifie si l'on veut opérer une resynthèse. La seconde est préférable dans une application de reconnaissance de la parole.

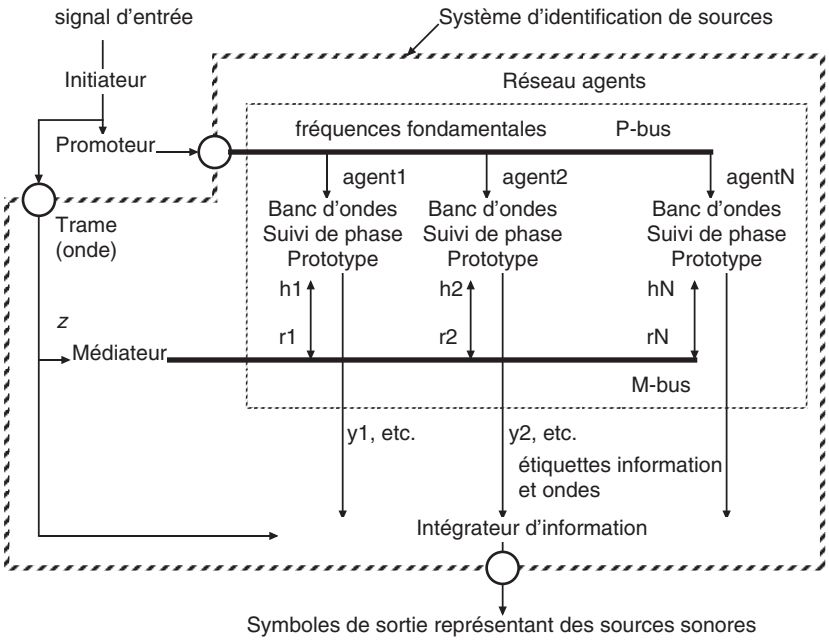

Figure 5.6. *Un exemple de structure d'un système CASA : le système Ipanema d'analyse musicale. Chaque agent est spécialisé dans la traque d'un aspect du signal, sous le contrôle du « médiateur » (d'après [KAS 96]).*

5.4. Critique de l'approche CASA

L'approche CASA est fertile, mais elle présente des faiblesses et écueils qu'il faut identifier si on veut les éviter.

5.4.1. *Limites de l'ASA*

L'ASA est fondée sur l'idée que la scène auditive peut se traiter comme une scène visuelle, et que les principes Gestalt qui régissent l'une peuvent se transposer à l'autre, moyennant le choix d'une représentation adéquate, et quelques aménagements pour tenir compte des spécificités du domaine acoustique. Cette idée mène dans certains cas à des intuitions erronées.

Prenons par exemple l'harmonicité, un principe de groupement-clé en ASA, très utilisé dans les modèles CASA. L'ASA voudrait que la régularité spectrale ou temporelle d'une cible harmonique en fasse une figure, facile à distinguer du fond (inharmonique, ou de F_0 différente). L'harmonicité d'une cible lui confèrerait une sorte de texture qui faciliterait son identification. Il n'en est rien : de nombreuses expériences ont montré

que l'harmonicité du *fond* (masqueur) facilite la ségrégation, mais celle de la cible n'a guère d'effet [DEC 95, DEC 97b, LEA 92, SUM 92c]. On peut aussi montrer que l'harmonicité de la cible est d'une utilité limitée pour séparer des voix concurrentes dans une tâche de reconnaissance de la parole, et moindre que celle de l'interférence [DEC 93b, DEC 94].

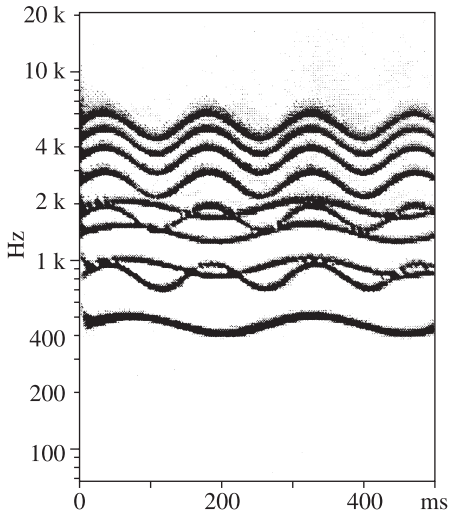

Figure 5.7. *L'incohérence de modulation de fréquence entre deux sources concurrentes est exploitée par le système d'analyse musicale de [MEL 91]. Les expériences psychoacoustiques ont pourtant montré que cette information n'est pas utilisée par le système auditif. Il s'agit d'un exemple où un système CASA exploite un principe Gestalt (destin commun des fréquences) qui n'est pas en accord avec la réalité perceptive.*

Autre exemple, le principe Gestalt de *destin commun* voudrait qu'un spectre fait de composantes qui bougent en parallèle (modulation de fréquence cohérente) forment une figure particulièrement facile à distinguer d'un fond statique ou modulé de façon différente. La modulation de fréquence d'une cible, de façon incohérente par rapport au fond, devrait ainsi faciliter son identification. Encore une fois, il n'en est rien : l'expérience montre que la modulation de fréquence n'a guère d'effet autre que celui, éventuel, de la différence de F_0 instantanée qu'elle induit [CAR 94, DAR 95, DEM 90, MAR 97, MCA 89, SUM 92b].

Encore un exemple, la qualité de la corrélation binaurale d'une cible détermine la précision de sa localisation. On pourrait penser que cela facilite sa ségrégation, quelle que soit la nature du fond. Encore une fois, il n'en est rien : la ségrégation dépend de la corrélation binaurale du *masqueur* et non de la cible. Un son masquant bien corrélé est facile à éliminer [COL 95, DUR 63]. Chose curieuse, il n'est pas nécessaire que cette corrélation soit cohérente entre les différents canaux fréquentiels [CUL 95].

5.4.2. *Limites de la notion de représentation séparable*

Comme signalé plus haut, l'enrichissement de la représentation et la multiplication de ses dimensions ne suffisent pas toujours à rendre les corrélats des différentes sources séparables. De nombreux auteurs se sont trouvés confrontés à la nécessité de scinder des éléments ou canaux [COO 91, ELL 96, PAR 76, WEI 85]. On peut alors se demander si l'étape de représentation séparable est nécessaire. Par exemple, les auteurs de [DEC 97b] ont montré que le modèle de Meddis et Hewitt [MED 92] ne pouvait pas expliquer tous les effets de différence de F_0 sur la ségrégation des voyelles. Ce modèle se fonde sur une représentation séparable du type autocorrélogramme (dimensions fréquence X délai X temps). En revanche, un modèle opérant sur la structure temporelle des décharges nerveuses dans chaque canal fréquentiel rend bien compte des phénomènes de ségrégation [DEC 97b]. Ce modèle est plus performant mais n'utilise pas une représentation séparable. Autre exemple, l'estimation des périodes de sons simultanés (par exemple, les notes d'instruments qui jouent ensemble) peut se faire sans avoir recours à une représentation séparable du type temps-fréquence, autocorrélation, etc. [DEC 93a, DEC 98].

Les représentations temps-fréquence-corrélation, etc., que l'on retrouve dans la plupart des modèles CASA ne sont ni une panacée ni un passage obligé pour effectuer des tâches d'organisation auditive.

5.4.3. *Ni modèle, ni méthode ?*

L'approche CASA offre un riche champ de liberté pour l'expérimentation d'idées, de modèles et de méthodes nouveaux. Ce n'est pas sans danger. Au mieux, le praticien CASA sera au courant de ce qui se fait en audition (psychoacoustique, physiologie) et parfaitement en prise avec le domaine d'application. Au pire, il ne sera ni l'un ni l'autre. Souvent, on voit défendre une approche peu réaliste au nom de l'efficacité, ou une méthode inefficace sous prétexte que « c'est comme ça que fait l'oreille ».

La modélisation (computationnelle ou autre) est florissante en théorie de l'audition, et il n'est pas toujours facile de situer la spécificité du modèle CASA. Inversement, il existe de nombreuses techniques de séparation de sources, réduction de bruit, etc. (en particulier du type *séparation aveugle*) qui ne relèvent pas du cadre CASA. Elles ne ressemblent pas forcément aux mécanismes perceptifs, mais il n'est pas sûr qu'elles soient moins efficaces pour autant.

5.5. Perspectives intéressantes

Malgré ces faiblesses, l'approche CASA continue à contribuer à la compréhension des mécanismes perceptifs, et à l'élaboration d'idées nouvelles en traitement du signal. Quatre évolutions récentes sont intéressantes.

5.5.1. *Théorie des données manquantes (*Missing Feature Theory*)*

Il est des situations où un système CASA (ou autre) n'arrive pas à restaurer une partie d'un signal cible. Les données correspondantes sont manquantes. Leur remplacement par une valeur nulle perturberait l'exploitation du pattern (par exemple dans un système de reconnaissance de la parole). Une valeur moyenne vaut à peine mieux. Dans certains cas, l'interpolation ou l'extrapolation à partir du contexte peut se justifier. Cependant, la solution optimale, dans une tâche de reconnaissance de formes, consiste à ignorer les données manquantes en leur affectant un poids nul [AHM 93, COO 94, COO 96, COO 97, DEC 93b, GRE 96, LIPP 97, MOR 98].

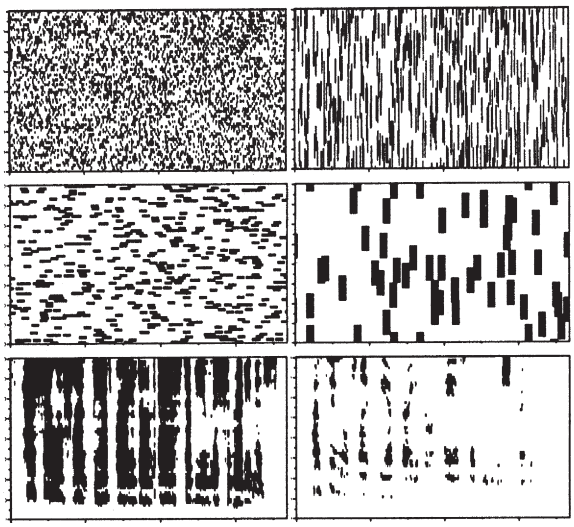

Figure 5.8. *Masques spectro-temporels utilisés dans les expériences sur les données manquantes. Les portions noires correspondent à l'information présente, le reste est absent. Les taux de reconnaissance restent élevés même pour un taux de suppression de 80 %. Il est aussi possible de faire l'apprentissage sur des données incomplètes (d'après [COO 96]).*

Dans cette approche, le module CASA a la responsabilité de fournir au module de reconnaissance une *carte de fiabilité*. Ce dernier doit être à même de l'exploiter, ce qui ne va pas sans poser quelques problèmes dans la pratique. Par exemple, beaucoup de systèmes de reconnaissance exploitent des paramètres *cepstraux*, dont l'avantage est d'être distribués de façon orthogonale et de permettre l'utilisation, par le modèle HMM (modèle de Markov caché), d'une matrice de covariance diagonale. Une carte de fiabilité dans le domaine *spectral* ne peut pas directement être exploitée par un tel système. L'utilisation de paramètres spectraux plutôt que cepstraux pose d'autres problèmes [MOR 96].

La maîtrise des techniques *données manquantes* est sans doute une clé de l'utilisation pratique de l'approche CASA. Elles peuvent aussi avoir une utilité plus large, par exemple pour l'intégration d'informations de modalités différentes. Par exemple, un système de reconnaissance audio-visuelle a intérêt à attribuer un poids faible à l'image lorsque le locuteur tourne la tête et la bouche n'est plus visible. Au contraire, il faut attribuer un poids faible au son lorsque la parole est masquée par un bruit.

5.5.2. *Le principe d'annulation*

Traditionnellement, l'ASA utilise la structure des sons cibles (par exemple, leur périodicité) pour les extraire d'un environnement non structuré, ou structuré différemment. Or, on s'est aperçu que cette approche n'est pas forcément très efficace et que ce n'est souvent pas ainsi que procède le système auditif. Prenons le cas de deux microphones, captant deux sources dont l'azimut est distinct. Un système exploitant la position de la cible arrivera au mieux (par beam-forming) à une réduction de rapport signal sur bruit de 6 dB, alors qu'un système exploitant la position de l'interférence peut aboutir à un rapport signal sur bruit infini (même si, en pratique, l'amélioration est moindre en cas de réverbération ou masqueurs multiples). De façon analogue, un système exploitant la périodicité de la cible pour la renforcer fonctionnera moins bien qu'un système exploitant celle du fond pour l'annuler [DEC 93a, DEC 93b]. Le système auditif exploite la périodicité du fond plutôt que celle de la cible [DEC 95, DEC 97c, LEA 92, SUM 92c]. Le critère d'annulation est proche de celui employé par les techniques de séparation aveugle. L'analyse de scènes par annulations successives est une caractéristique du système de Nakatani [NAK 95a, NAK 95b, NAK 97].

L'annulation offre dans certains cas un taux de rejet infini (amélioration infinie du taux cible/fond), mais elle introduit en général une distorsion de la cible. Par exemple, les composantes partagées ou masquées sont supprimées. Les techniques de données manquantes sont utiles dans ce cas.

5.5.3. *L'intégration multimodale*

Le développement de la reconnaissance de la parole multimodale laisse entrevoir une analyse de scènes multimodale, qui serait plus que la simple juxtaposition de modules d'analyse de scènes visuelles et auditives [OKU 99a]. Là encore, les techniques de données manquantes promettent d'être utiles pour l'intégration de données modales de fiabilité variable.

5.5.4. *Synthèse de scènes auditives : mesure de transparence*

On peut aborder la question de l'ASA d'un angle radicalement différent, celui du concepteur d'une scène sonore. Lorsque des matériaux sonores sont assemblés par

mixage, il peut arriver qu'un ingrédient soit particulièrement masquant et contribue à rendre confuse la scène sonore. La prise en compte des divers paramètres révélés par l'ASA permet de prédire le degré masquant d'une source en fonction de ses caractéristiques physiques. Une mesure de transparence sonore serait utile pour le concepteur pour bien choisir ses ingrédients. Une telle mesure a été proposé pour la future norme MPEG7 de description de données multimédias [DEC 99b].

5.6. Bibliographie

[AHM 93] AHMAD S., TRESP V., « Some solutions to the missing feature problem in vision », dans S.J. Hanson, J.D. Cowan et C.L. Giles (dir.), *Advances in Neural Information Processing Systems 5*, p. 393-400, San Mateo, Morgan Kaufmann, 1993.

[ASS 90] ASSMANN P.F., SUMMERfiELD Q., « Modeling the perception of concurrent vowels : Vowels with different fundamental frequencies », *J. Acoust. Soc. Am.*, 88, p. 680-697, 1990.

[BER 95] BERTHOMMIER F., MEYER G., « Source separation by a functional model of amplitude demodulation », *Proc. of ESCA Eurospeech*, p. 135-138, 1995.

[BOD 96] BODDEN M., RATEIKSHEK K., « Noise-robust speech recognition based on a binaural auditory model », *Proc. of Workshop on the auditory basis of speech perception*, Keele, p. 291-296, 1996.

[BRE 90] BREGMAN A.S., *Auditory scene analysis*, MIT Press, Cambridge, 1990.

[BRO 82] BROKX J.P.L., NOOTEBOOM S.G., « Intonation and the perceptual separation of simultaneous voices », *Journal of Phonetics*, 10, p. 23-36, 1982.

[BRO 92a] BROWN G.J., Computational auditory scene analysis : a representational approach, Sheffield, Department of Computer Science, these de doctorat non publiée, 1992.

[BRO 92b] BROWN G.J., COOKE M.P., « Computational auditory scene analysis : grouping sound sources using common pitch contours », *Proc. Inst. of Acoust.*, 14, p. 439-446, 1992.

[BRO 93] BROWN G.J., COOKE M., « Physiologically-motivated signal representations for computational auditory scene analysis », dans M. Cooke, S. Beet et M. Crawford (dir.), *Visual representations of speech signals*, Chichester, John Wiley and Sons, p. 181-188, 1993.

[CAR 94] CARLYON R., « Further evidence against an accross-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components », *J. Acoust. Soc. Am.*, 95, p. 949-961, 1994.

[CHE 53] CHERRY E.C., « Some experiments on the recognition of speech with one, and with two ears », *J. Acoust. Soc. Am.*, 25, p. 975-979, 1953.

[COL 95] COLBURN H.S., « Computational models of binaural processing », dans H. Hawkins, T. McMullin, A.N. Popper et R.R. Fay (dir.), *Auditory Computation*, New York, Springer-Verlag, p. 332-400, 1995.

[COO 91] COOKE M.P., Modeling auditory processing and organisation, Sheffield, Department of Computer Science, thèse non publiée, 1991.

[COO 93] COOKE M.P., BROWN G.J., « Computational auditory scene analysis : exploiting principles of perceived continuity », *Speech Comm.*, 13, p. 391-399, 1993.

[COO 94] COOKE M., GREEN P., ANDERSON C., ABBERLEY D., Recognition of occluded speech by hidden markov models, University of Sheffield Department of Computer Science, Technical report, TR-94-05-01, 1994.

[COO 96] COOKE M., MORRIS A., GREEN P., « Recognising occluded speech », *Proc. Workshop on the Auditory basis of Speech Perception*, Keele, p. 297-300, 1996.

[COO 97] COOKE M., MORRIS A., GREEN P., « Missing data techniques for robust speech recognition », *Proc. ICASSP*, p. 863-866, 1997.

[COO 99] COOKE M., ELLIS D.P.W., « The auditory organization of speech in listeners and machines », *Speech Communication* (soumis), 1999.

[CUL 95] CULLING, J.F., SUMMERfIELD Q., « Perceptual segregation of concurrent speech sounds : absence of across-frequency grouping by common interaural delay », *J. Acoust. Soc. Am.*, 98, p. 785-797, 1995.

[DAR 95] DARWIN C.J., CARLYON R.P., « Auditory grouping », dans B.C.J. Moore (dir.), *Handbook of perception and cognition : Hearing*, New York, Academic Press, p. 387-424, 1995.

[DEC 93a] DE CHEVEIGNÉ A., « Separation of concurrent harmonic sounds : Fundamental frequency estimation and a time-domain cancellation model of auditory processing », *J. Acoust. Soc. Am.*, 93, p. 3271-3290, 1993.

[DEC 93b] DE CHEVEIGNÉ A., Time-domain comb filtering for speech separation, ATR Human Information Processing Laboratories technical report, TR-H-016, 1993.

[DEC 94] DE CHEVEIGNÉ A., KAWAHARA H., AIKAWA K., LEA A., « Speech separation for speech recognition », *Journal de Physique*, IV 4, C5-545-C5-548, 1994.

[DEC 95] DE CHEVEIGNÉ A., MCADAMS S., LAROCHE J., ROSENBERG M., « Identification of concurrent harmonic and inharmonic vowels : A test of the theory of harmonic cancellation and enhancement », *J. Acoust. Soc. Am.*, 97, p. 3736-3748, 1995.

[DEC 97a] DE CHEVEIGNÉ A., « Concurrent vowel identification III : A neural model of harmonic interference cancellation », *J. Acoust. Soc. Am.*, 101, p. 2857-2865, 1997.

[DEC 97b] DE CHEVEIGNÉ A., KAWAHARA H., TSUZAKI M., AIKAWA K., « Concurrent vowel identification I : Effects of relative level and F0 difference », *J. Acoust. Soc. Am.*, 101, p. 2839-2847, 1997.

[DEC 97c] DE CHEVEIGNÉ A., MCADAMS S., MARIN C., « Concurrent vowel identification II : Effects of phase, harmonicity and task », *J. Acoust.Soc. Am.*, 101, p. 2848-2856, 1997.

[DEC 98] DE CHEVEIGNÉ A., « Cancellation model of pitch perception », *J. Acoust. Soc. Am.*, 103, p. 1261-1271, 1998.

[DEC 99a] DE CHEVEIGNÉ A., KAWAHARA H., « Multiple period estimation and pitch perception model », *Speech Communication*, 27, p. 175-185, 1999.

[DEC 99b] DE CHEVEIGNÉ A., SMITH B., A 'sound transparency' descriptor ISO/IEC JTC1/SC29/WG11, MPEG99/m5199, 1999.

[DUR 63] DURLACH, N.I., « Equalization and cancellation theory of binaural masking-level differences », *J. Acoust. Soc. Am.*, 35, p. 1206-1218, 1963.

[ELL 96] ELLIS D., Prediction-driven computational auditory scene analysis, MIT, thèse non publiée, 1996.

[ELL 97] ELLIS D.P.W., « Computational auditory scene analysis exploiting speech-recognition knowledge », *Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio*, Mohonk, 1997.

[GRE 95] GREEN P.D., COOKE M.P., CRAWFORD M.D., « Auditory scene analysis and hidden markov model recognition of speech in noise », *Proc. IEEE-ICASSP*, p. 401-404, 1995.

[GRE 97] GREENBERG, « Understanding speech understanding : towards a unified theory of speech perception », *Proc. ESCA Workshop on the auditory basis of speech perception*, Keele, p. 1-8, 1997.

[HAR 96] HARTMANN W.M., « Pitch, periodicity, and auditory organization », *J. Acoust. Soc. Am.*, 100, p. 3491-3502, 1996.

[HEL 77] HELMHOLTZ H. v., (1877). *On the sensations of tone (English translation A.J. Ellis, 1954)*, New York, Dover.

[HER 94] HERMANSKY H., MORGAN N., « RASTA processing of speech », *IEEE trans Speech and Audio Process.*, 2, p. 578-589, 1994.

[HOL 88] HOLDSWORTH J., NIMMO-SMITH I., PATTERSON R.D., RICE P., Implementing a GammaTone filter bank, MRC Applied Psychology Unit technical report, SVOS final report, annex C, 1998.

[HOL 90] HOLDSWORTH J., Two dimensional adaptive thresholding, APU AAM-HAP Report technical report, vol. 1, annex 4, 1990.

[HOL 92] HOLDSWORTH J., SCHWARTZ J.-L., BERTHOMMIER F., PATTERSON R.D., « A multi-representation model for auditory processing of sounds », dans Y. Cazals, L. Demany et K. Horner (dir.), *Auditory physiology and perception*, Oxford, Pergamon Press, p. 447-453, 1992.

[JOR 98] JORIS P.X., YIN T. C. T., « Envelope coding in the lateral superior olive. III. Comparison with afferent pathways », *J. Neurophysiol.*, 79, p. 253-269, 1998.

[KAN 98] KANADERA N., HERMANSKY H., ARAI T., « On properties of the modulation spectrum for robust automatic speech recognition », *Proc. IEEE-ICASSP*, p. 613-616, 1998.

[LEA 92] LEA A., Auditory models of vowel perception, Nottingham University, thèse non publiée, 1992.

[LIC 59] LICKLIDER J.C.R., « Three auditory theories », S. Koch (dir.), *Psychology, a study of a science*, New York, McGraw-Hill, I, p. 41-144, 1959.

[LIP 97] LIPPMANN R.P., CARLSON B.A., « Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise », *Proc. ESCA Eurospeech*, KN-37-40, 1997.

[LYO 83] LYON, R.F., « A computational model of binaural localization and separation », W. Richards (dir.), *Natural computation*, Cambridge, Mass, MIT Press, p. 319-327, 1983.

[LYO 84] LYON R., « Computational models of neural auditory processing », *Proc. IEEE ICASSP*, 36.1.(1-4), 1984.

[LYO 91] LYON R., « Automatic gain control in cochlear mechanics », dans P. Dallos, C.D.Geisler, J.W. Mathews (dir.), *Mechanics and biophysics of hearing*, M.A. Ruggero et C.R.Steele, New York, Springer-Verlag, 1991.

[MAR 82] MARR D., « Representing and computing visual information », dans P.H. Winston et R.H. Brown, *Artificial Intelligence : an MIT perspective*, Cambridge, Mass, MIT Press, 2, p. 17-82, 1982.

[MAR 97] MARIN C., DE CHEVEIGNÉ A., « Rôle de la modulation de fréquence dans la séparation de voyelles », *Proc. Congrès Français d'Acoustique*, p. 527-530, 1997.

[MCA 84] MCADAMS S., Spectral fusion, spectral parsing, and the formation of auditory images, Stanford University, thèse non publiée, 1984.

[MCA 89] MCADAMS S., « Segregation of concurrent sounds. I : Effects of frequency modulation coherence », *J. Acoust. Soc. Am.*, 86, p. 2148-2159, 1989.

[MED 88] MEDDIS R., « Simulation of auditory-neural transduction : further studies », *J. Acoust. Soc. Am.*, 83, p. 1056-1063, 1988.

[MED 92] MEDDIS R., HEWITT M.J., « Modeling the identification of concurrent vowels with different fundamental frequencies », *J. Acoust. Soc. Am.*, 91, p. 233-245, 1992.

[MEL 91] MELLINGER D.K., Event formation and separation in musical sound, Stanford Center for computer research in music and acoustics, thèse non publiée, 1991.

[MEY 96] MEYER G., BERTHOMMIER F., « Vowel segregation with amplitude modulation maps : a re-evaluation of place and place-time models », *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, p. 212-215, 1996.

[MEY 97] MEYER G.F., PLANTE F., BERTHOMMIER F., « Segregation of concurrent speech with the reassigned spectrum », *Proc. IEEE ICASSP*, p. 1203-1206, 1997.

[MOR 98] MORRIS A.C., COOKE M.P., GREEN P.D., « Some solutions to the missing feature problem in data classification, with application to noise robust ASR », *Proc. ICASSP*, p. 737-740, 1998.

[NAD 97] NADEU C., PACHÈS-LEAL P., JUANG B.-H., « Filtering the time sequences of spectral parameters for speech recognition », *Speech Comm.*, 22, p. 315-332, 1997.

[NAK 95a] NAKATANI T., OKUNO H.G., KAWABATA T., « Residue-driven architecture for computational auditory scene analysis », *Proc. IJCAI*, p. 165-172, 1995.

[NAK 95b] NAKATANI T., GOTO M., ITO T., OKUNO H.G., « Multi-agent based binaural sound stream segregation », *Proc. IJCAI Workshop on Computational Auditory Scene Analysis*, p. 84-91, 1995.

[NAK 96] NAKATANI T., GOTO M., OKUNO H. G., « Localization by harmonic structure and its application to harmonic stream segregation », *Proc. IEEE ICASSP*, p. 653-656, 1996.

[NAK 97] NAKATANI T., KASHINO K., OKUNO J.G., « Integration of speech stream and music stream segregations based on a sound ontology », *Proc. IJCAI Workshop on computational auditory scene analysis*, Nagoya, p. 25-32, 1997.

[OKU 99a] OKUNO H.G., NAKAGAWA Y., KITANO H., « Incorporating visual information into sound source separation », *Proc. International workshop on Computational Auditory Scene Analysis*, 1999.

[OKU 99b] OKUNO H.G., IKEDA S., NAKATANI T., « Combining independant component analysis and sound stream segregation », *Proc. International workshop on computational auditory scene analysis*, 1999.

[PAR 76] PARSONS, T.W., « Separation of speech from interfering speech by means of harmonic selection », *J. Acoust. Soc. Am.*, 60, p. 911-918, 1976.

[PAT 92] PATTERSON R.D., ROBINSON K., HOLDSWORTH J., MCKEOWN D., ZHANG C., ALLERHAND M., « Complex sounds and auditory images », dans Y. Cazals, K. Horner et L. Demany (dir.), *Auditory physiology and perception*, Oxford, Pergamon Press, p. 429-446, 1992.

[PAT 96] PATTERSON R., ANDERSON T.R., FRANCIS K., « Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition », *Proc. Workshop on the auditory basis of speech perception*, Keele, p. 245-252, 1996.

[ROS 97] ROSENTHAL D.F., OKUNO H.G., *Computational auditory scene analysis*, Lawrence Erlbaum, 1997.

[SCH 83] SCHEFFERS M.T.M., Sifting vowels, Gröningen University, Thèse, 1983.

[SEN 85] SENEFF S., Pitch and spectral analysis of speech based on an auditory synchrony model, MIT, thèse non publiée (technical report 504), 1985.

[SHA 85] SHAMMA S.A., « Speech processing in the auditory system I : The representation of speech sounds in the responses of the auditory nerve », *J. Acoust. Soc. Am.*, 78, p. 1612-1621, 1985.

[SLA 93] SLANEY M., An efficient implementation of the Patterson-Holdsworth auditory filter bank, Apple Computer technical report, 35, 1993.

[SLA 95] SLANEY, M., « A critique of pure audition », *Proc. Computational auditory scene analysis workshop*, IJCAI, Montreal, 1995.

[SUM 90] SUMMERFIELD Q., LEA A., MARSHALL D., « Modelling auditory scene analysis : strategies for source segregation using autocorrelograms », *Proc. Institute of Acoustics*, 12, p. 507-514, 1990.

[SUM 92a] SUMMERFIELD Q., CULLING J. F., « Auditory segregation of competing voices : absence of effects of FM or AM coherence », *Phil. Trans. R. Soc. Lond.*, B 336, p. 357-366, 1992.

[SUM 92b] SUMMERFIELD Q., « Roles of harmonicity and coherent frequency modulation in auditory grouping », dans M.E.H. Schouten (dir.), *The auditory processing of speech : from sounds to words*, Berlin, Mouton de Gruyter, p. 157-166, 1992.

[SUM 92c] SUMMERFIELD Q., CULLING J.F., « Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency », *Proc. 124th meeting of the ASA*, 2317(A), 1992.

[WAN 95] WANG A.L.-C., Instantaneous and frequency-warped signal processing techniques for auditory source separation, CCRMA (Stanford University), thèse non publiée, 1995.

[WAR 70] WARREN R.M., « Perceptual restoration of missing speech sounds », *Science*, 167, p. 392-393, 1970.

[WAR 72] WARREN R.M., OBUSEK C.J., ACKROFF J. M., « Auditory induction : perceptual synthesis of absent sounds », Science, 176, p. 1149-1151, 1972.

[WEI 85] WEINTRAUB M., A theory and computational model of auditory monaural sound separation, Stanford University, thèse non publiée, 1985.

[YOS 96] YOST W.A., DYE R.H., SHEFT S., « A simulated “cocktail party” with up to three sound sources », *Perception and Psychophysics*, 58, p. 1026-1036, 1996.