# Analyse de Scènes Auditives Computationnelle

Alain de Cheveigné (CNRS / Ircam)

**Résumé**

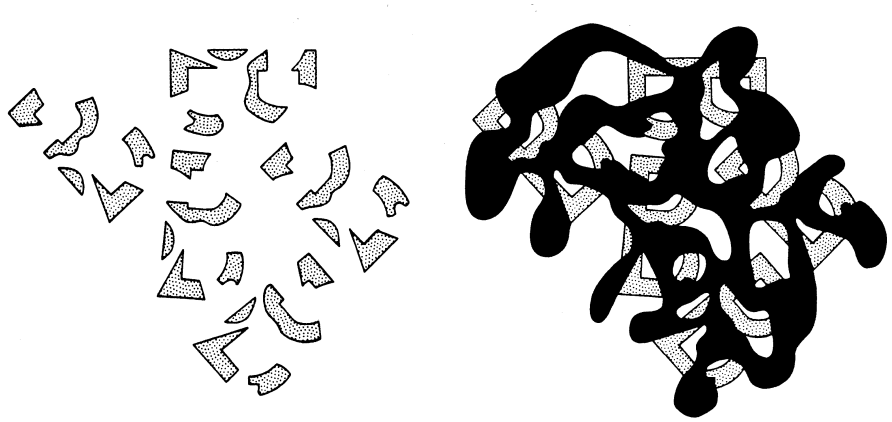
L'*Analyse de Scènes Auditives* (ASA) est la structuration par le système auditif d'un environnement sonore complexe. Les indices acoustiques de sources multiples se superposent dans le champ acoustique échantillonné par les deux oreilles. L'ASA les sépare perceptivement et les attribue aux sources qui les ont produites. Elle contribue ainsi au modèle que se fait l'individu du monde qui l'entoure, qui guide son action et facilite sa survie. L'*Analyse de Scènes Auditives Computationnelle* (CASA) tente de faire la même chose par des moyens computationnels, soit en tant que modèle des processus biologiques, soit dans le but d'une application pratique (par exemple dans un système de reconnaissance de la parole).

Les propriétés de l'ASA sont d'abord énumérées brièvement, puis les principes qui guident les modèles CASA. La première étape d'un modèle CASA est généralement d'extraire une représentation multidimensionnelle qui puisse servir de substrat à l'analyse de la "scène". La représentation est généralement du type temps-fréquence, parfois augmentée de dimensions telles que l'autocorrélation monaurale ou la corrélation interaurale. Les éléments de la représentation sont structurés de façon hiérarchique, en commençant par le bas ("bottom-up") pour simuler les processus ASA dits "primitifs", ou par le haut ("top-down") pour simuler les processus dits "à base de schémas". Des techniques d'intelligence artificielle sont utilisées pour gérer les données et le processus de structuration.

Je termine par une critique de l'approche CASA : en tant que modèle des processus perceptifs, elle n'est pas toujours en accord avec les données physiologiques ou psychoacoustiques. En tant que méthode de traitement du signal acoustique, elle n'est pas toujours à la hauteur de ses ambitions. J'indique enfin un certain nombre de directions prometteuses, issues de travaux récents.

## 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



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 Bregman, 1990).

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 (1877) déjà se demandait comment on pouvait percevoir les qualités individuelles d'instruments qui jouent ensemble. Mais il faut attendre le travail de Bregman (1990) pour que l'*Analyse de Scène Auditives* (Auditory Scene Analysis, ou ASA) devienne un sujet d'étude à part entière. 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 est une transposition dans le domaine de l'audition des principes de l'analyse de scènes visuelles.

Avec le développement de l'Informatique et de l'Intelligence Artificielle sont apparues des tentatives d'*Analyse de Scène Auditive Computationnelle* (CASA) (Lyon, 1983 ; Weintraub, 1985 ; Cooke, 1991 ; Mellinger, 1991 ; Brown,

1992 ; Wang, 1995 ; Ellis, 1996). 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 (1982), 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) (Cooke et al., 1994, 1997 ; Morris et al., 1998 ; Lippmann, 1997).

# 1 Principes de ASA

## 1.1 Fusion vs. 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 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. À 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.

## 1.2 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 : 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 Marin, 1991), 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 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.

## 1.3 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 fréquences 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.

## 1.4 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 IA.

### 1.5 Illusion de continuité, restauration phonémique

Lorsqu'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 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.

## 2 Principes de CASA

### 2.1 Création d'une représentation

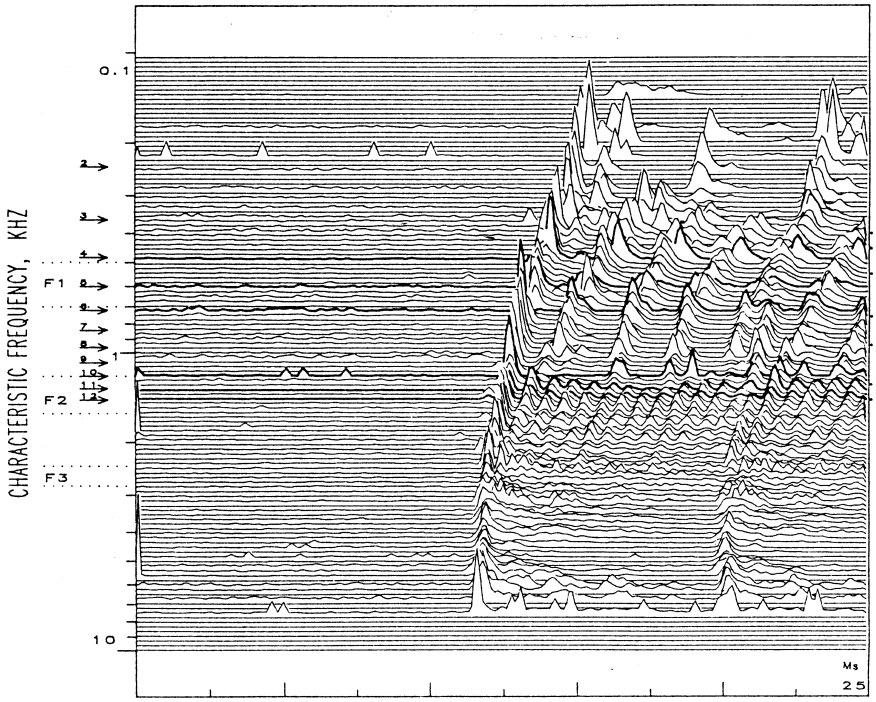
L'analogie avec l'analyse de scène visuelle, sur laquelle s'appuie l'ASA, suppose l'existence d'une "représentation" d'une richesse comparable à l'espace 3-D des objets ou 2-D des images (Marr, 1982, utilise le terme de 2 1/2-D pour qualifier la représentation enrichie fournie par la vision binoculaire et autres mécanismes de perception de la profondeur). L'onde acoustique étant de dimensionnalité faible, le modèle CASA commence par synthétiser une représentation plus riche.

#### 2.1.1 Filtre "cochléaire"

Le modèle CASA typique commence par un banc de filtres. En principe ils se veulent conformes à ce qu'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 (Holdsworth et al., 1988 ; Cooke, 1991 ; Patterson et al., 1992 ; Slaney, 1993). 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.

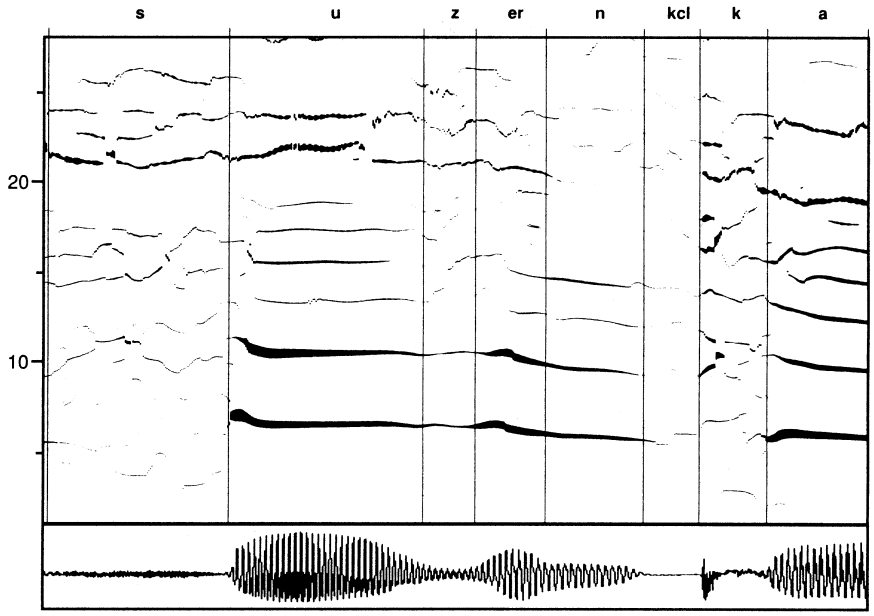
#### 2.1.2 Transduction

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 :




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 Shamma, 1985).

- 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 (Lyon 1982, 1984 ; Seneff, 1985 ; Holdsworth, 1990, Patterson et al. 1992) ou modèle de cellule ciliée (Meddis, 1986, 1988).
- 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




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 Cooke, 1991).

exemple est le LIN (lateral inhibitory network) de Shamma (1985).

- 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 qu'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 se voir 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.

### 2.1.3 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 (Sect. **??**). Par rap-

port à cette représentation idéalisée, elle peut sembler manquer de résolution fréquentielle, ou temporelle, ou les deux. On a cité le LIN de Shamma (1985) qui renforce le contraste spectral. Deng (1988) propose la corrélation croisée entre canaux voisins pour renforcer la représentation des formants. Les "synchrony strands" de Cooke (1991) produisent une représentation proche d'une somme de sinusoïdes, propice à l'application des principes ASA (continuité de chaque "strand", destin commun, harmonicité, etc.). 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.

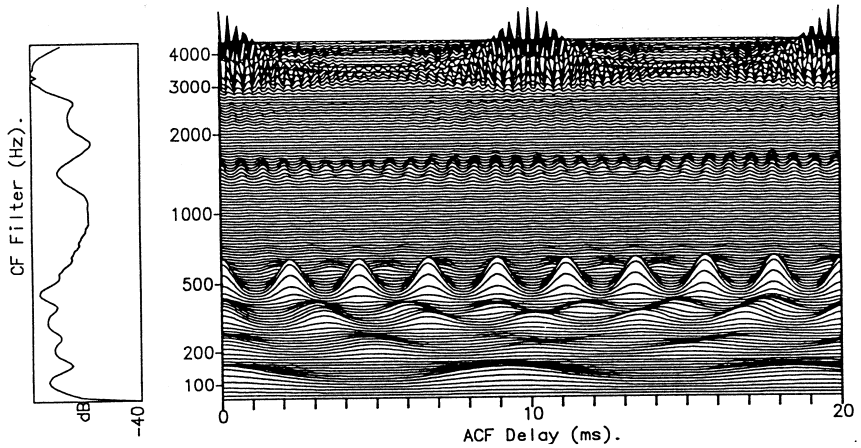
### 2.1.4 Dimensions supplémentaires

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 (1983), 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. 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 (1983) échantillonne la représentation, par coupes parallèles à l'axe des fréquences, pour isoler telle ou telle source. Des tentatives similaires ont été faites depuis (Bodden et al., 1996 ; Patterson et al., 1996).

Une autre dimension apparaît si on calcule la fonction d'*autocorrélation* dans chaque canal. Cette idée a été proposée à l'origine par Licklider (1959) pour estimer la période dans un modèle de perception de la hauteur. 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 (1985) cette idée a été reprise par Mellinger (1991), Meddis et Hewitt (1992), Brown (1992), Lea (1992), Ellis (1996).

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 d'intéressant à nous dire sur le signal. Un *lissage temporel* permet de se débarrasser de la structure fine et ne retenir (on espère) que les modulations qui reflètent la période fondamentale. Celles-ci peuvent alors être évaluées par autocorrélation ou

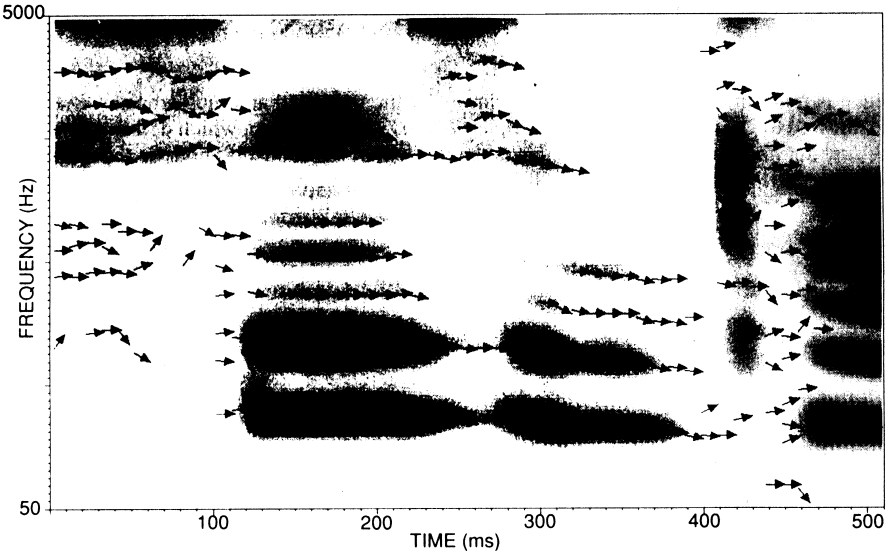



Pattern d'autocorrélation, en réponse à un mélange de voyelles (/i/ à 100 Hz et /o/ à 112 Hz). Chaque trait correspond à un canal issu du filtre périphérique. Chaque canal sera affecté à une voyelle en fonction de la périodicité qui le domine (d'après Lea, 1992).

par d'autres méthodes : passages par zéro (Cooke, 1991), transformée de Fourier (Meyer et al., 1996, 1997). 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 (Hermansky et al., 1994 ; Greenberg, 1996 ; Nadeu et al. 1997 ; Kanedera et al., 1998).

D'autres transformations sont la carte de transition fréquentielle (frequency transition map) de Brown (1992) ou les "onset maps" de Mellinger (1991), Brown (1992) et Ellis (1996), qui ont pour but de repérer les changements temporels brusques pouvant signaler le début d'un son.

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.




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 Brown et al., 1993).

### 2.1.5 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 (1991) 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 (1992), les "wefts" (trames) de Ellis (1996). Le principe de synchronicité d'attaque est utilisé par Brown pour former les "objets auditifs".

### 2.1.6 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 (Ellis, 1996 ; Nakatani et al., 1997 ; Godsmark et al., 1997 ; Kashino et al., 1997). 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 in-

dispensables pour gérer l'ensemble des sources d'informations et hypothèses qui interviennent dans l'organisation d'une scène auditive.

### 2.1.7 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. À signaler la proposition d'Ellis (1997) d'utiliser un système de reconnaissance de la parole pour guider l'analyse de scène auditive. 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.

### 2.1.8 Le problème des composantes partagées

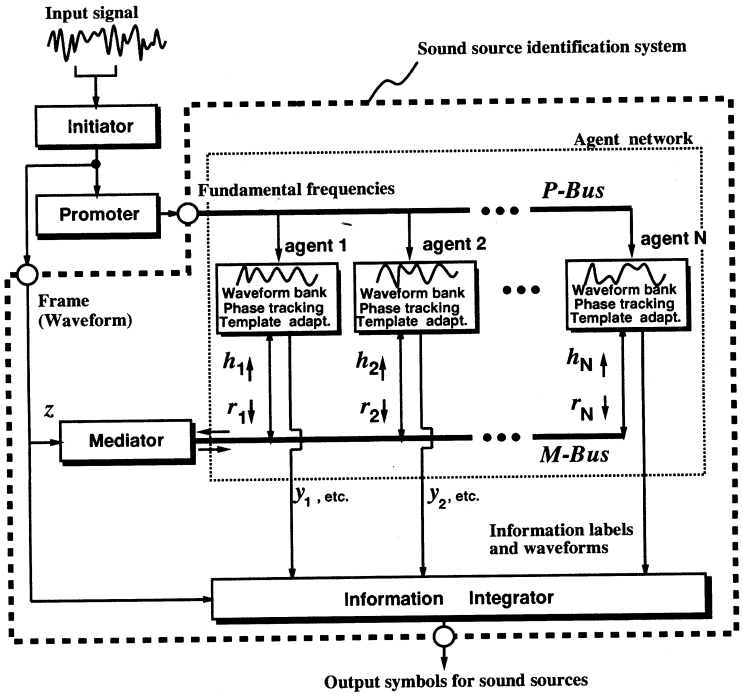
Quelle que soit la richesse et la dimensionnalité de la représentation de base, il arrive que l'appartenance d'un "élément" soit ambiguë. Les stratégies divergent selon qu'on décide alors de l'attribuer à une seule des sources (principe d'allocation exclusive), aux deux (attribution "duplex") ou à aucune. On peut aussi essayer de *scinder* l'élément, par exemple selon des critères de continuité fréquentielle ou temporelle (Weintraub, 1985). 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.

### 2.1.9 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 (Ellis, 1996 ; Masuda-Katsuse et al., 1997).
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 de reconnaissance de formes (Cooke et al., 1997 ; Morris et al. 1998 ; Lippmann et al., 1998).

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




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 Kashino et al., 1996).

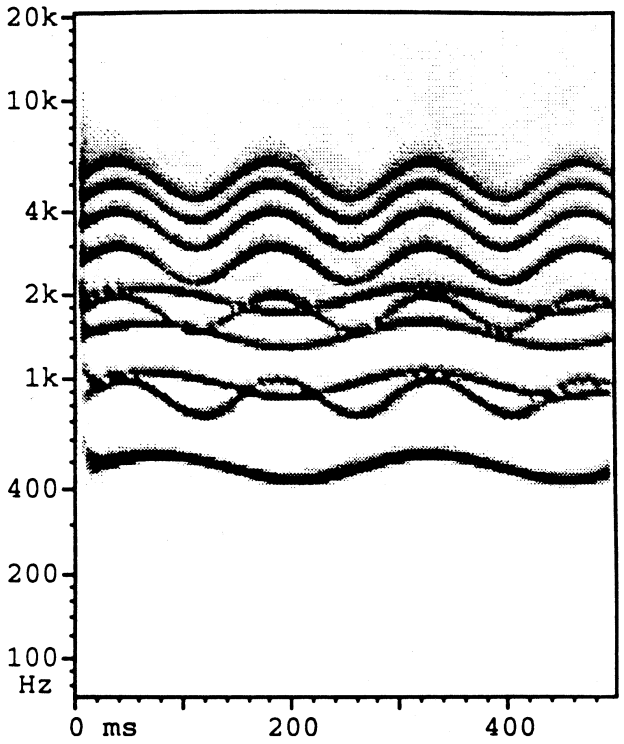
# 3 Critique de l'approche CASA

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

## 3.1 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




L'incohérence de modulation de fréquence entre deux sources concurrentes est exploitée par le système d'analyse musicale de Mellinger (1991). 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.

*fond* (masqueur) facilite la ségrégation, mais celle de la cible n'a guère d'effet (Summerfield et al., 1992 ; Lea, 1992 ; de Cheveigné et al., 1995, 1997). 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, moindre que celle de l'interférence (de Cheveigné, 1993b ; de Cheveigné et al., 1994).

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 (McAdams, 1989 ; Demany et al., 1990 ; Summerfield et al., 1992 ; Carlyon, 1994 ; Darwin et al., 1995 ; Marin et al., 1997).

Encore un exemple, la qualité de la corrélation binaurale d'une cible détermine la précision de sa localisation. On pourrait penser que cela faciliterait du coup 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 (Durlach, 1963 ; Colburn, 1995). Chose curieuse, il n'est pas nécessaire que cette corrélation soit cohérente entre les différents canaux fréquentiels (Culling et al., 1995).

## 3.2 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 (canaux, etc.) (Parsons, 1976 ; Weintraub, 1985 ; Cooke, 1991 ; Ellis, 1996). Si la séparation est possible de cette façon-là, on peut alors se demander si l'étape de "représentation séparable" est nécessaire.

Par exemple, de Cheveigné et al. (1997) ont montré que le modèle de Meddis et al. (1992) 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 (de Cheveigné, 1997). Ce modèle 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. (de Cheveigné et al. 1993a, 1998).

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

## 3.3 Ni modèle, ni méthode ?

L'approche CASA offre un riche champ de liberté pour l'expérimentation d'idées, modèles et 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.

# 4 Perspectives intéressantes

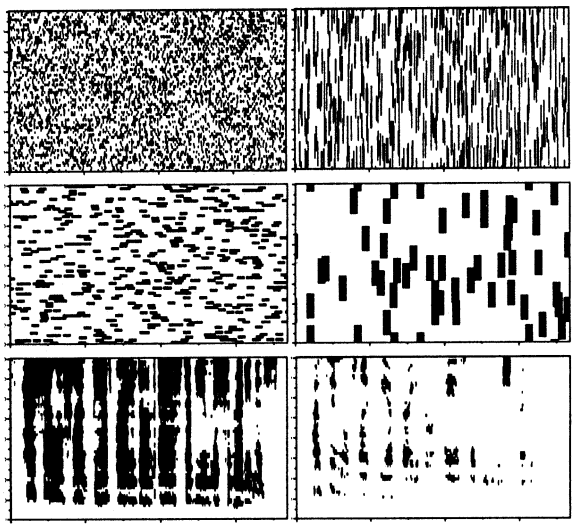
Malgré ces faiblesses, l'approche CASA continue à contribuer à la compréhension des mécanismes perceptifs, et à l'élaboration d'idées nouvelles en traitement du signal. Quatre évolutions récentes sont intéressantes.

## 4.1 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 (Ahmad et al., 1993 ; Cooke et al., 1994, 1996, 1997 ; Green et al., 1996 ; Morris et al. 1998 ; Lippmann 1997, de Cheveigné, 1993b).

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 (Morris et al., 1996).

La maîtrise des techniques "données manquantes" est sans doute une clé de l'utilisation efficace 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



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 Cooke et al., 1996).

bouche n'est plus visible. Au contraire il faut attribuer un poids faible au son lorsque la parole est masquée par un bruit.

## 4.2 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 (de Cheveigné, 1993a, b). Le système auditif exploite la périodicité du fond plutôt que celle de la cible (Summerfield et al., 1992 ; Lea, 1992 ; de Cheveigné et al., 1995, 1997). Le critère d'annulation est proche de ce-

lui employé par les techniques de "séparation aveugle". L'analyse de scènes par annulations successives est une caractéristique du système de Nakatani (1995a,b, 1997).

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.

### 4.3 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 (Okuno et al., 1999). Là encore, les techniques de "données manquantes" promettent d'être utiles pour l'intégration de données modales de fiabilité variable.

### 4.4 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 confus 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 multimédia (de Cheveigné et Smith, 1999).

## 5 Pour en savoir plus

La "bible" de l'Analyse de Scène Auditive est l'ouvrage de Bregman (1990). Y figurent tous les concepts de l'ASA, appuyés par de nombreux résultats expérimentaux. Darwin et Carlyon (1995) donnent une perspective plus récente. L'ouvrage de Rosenthal et Okuno (1997) et le numéro spécial de Speech Communication (Vol. 27, numéros 3-4, Avril 1999) donnent chacun un échantillonnage des développements récents de l'ASA computationnelle. Beaucoup de résultats sont disponibles sur WWW sous forme de thèses ou rapports techniques. La page

`http://www.ircam.fr/equipes/pcm/cheveign/sh/casa.html`

recense quelques pointeurs intéressants.

# Bibliographie


Ahmad, S., et Tresp, V. (1993). "Some solutions to the missing feature problem in vision," in "Advances in Neural Information Processing Systems 5," Edité par S. J. Hanson, J. D. Cowan et C. L. Giles, San Mateo, Morgan Kaufmann, 393-400.

Assmann, P. F., et Summerfield, Q. (1990). "Modeling the perception of concurrent vowels : Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, 680-697.

Berthommier, F., et Meyer, G. (1995). "Source separation by a functional model of amplitude demodulation.", Proc. ESCA Eurospeech, 135-138.

Bodden, M., et Rateikshek, K. (1996). "Noise-robust speech recognition based on a binaural auditory model.", Proc. Workshop on the auditory basis of speech perception, Keele, 291-296.

Bregman, A. S. (1990). "Auditory scene analysis," Cambridge, Mass., MIT Press.

Brokx, J. P. L., et Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," Journal of Phonetics 10, 23-36.

Brown, G. J. (1992), "Computational auditory scene analysis : a representational approach," Sheffield, Department of Computer Science unpublished doctoral dissertation.

Brown, G. J., et Cooke, M. P. (1992). "Computational auditory scene analysis : grouping sound sources using common pitch contours," Proc. Inst. of Acoust. 14, 439-446.

Brown, G. J., et Cooke, M. (1993). "Physiologically-motivated signal representations for computational auditory scene analysis," in "Visual representations of speech signals," Edité par M. Cooke, S. Beet et M. Crawford, Chichester, John Wiley and Sons, 181-188.

Helmholtz, H. v. (1877). "On the sensations of tone (English translation A.J. Ellis, 1954)," New York, Dover.

Carlyon, R. (1994). "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, 949-961.

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

Colburn, H. S. (1995). "Computational models of binaural processing," in "Auditory Computation," Edité par H. Hawkins, T. McMullin, A. N. Popper et R. R. Fay, New York, Springer-Verlag, 332-400.

Cooke, M. P. (1991), "Modeling auditory processing and organisation," Sheffield, Department of Computer Science, thèse non publiée.

Cooke, M. P., et Brown, G. J. (1993). "Computational auditory scene analysis : exploiting principles of perceived continuity," Speech Comm. 13, 391-399.

Cooke, M., Green, P., Anderson, C., et Abberley, D. (1994), "Recognition of occluded speech by hidden markov models," University of Sheffield Department of Computer Science technical report, TR-94-05-01.

Cooke, M., Morris, A., et Green, P. (1996). "Recognising occluded speech.", Proc. Workshop on the Auditory basis of Speech Perception, Keele, 297-300.

Cooke, M., Morris, A., et Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.

Cooke, M., and Ellis, D. P. W. (1999). "The auditory organization of speech in listeners and machines," Speech Communication (submitted)

Culling, J. F., et Summerfield, Q. (1995). "Perceptual segregation of concurrent speech sounds : absence of across-frequency grouping by common interaural delay.," J. Acoust. Soc. Am. 98, 785-797.

Darwin, C. J., et Carlyon, R. P. (1995). "Auditory grouping," in "Handbook of perception and cognition : Hearing," Edité par B. C. J. Moore, New York, Academic Press, 387-424.

de Cheveigné, A. (1993a). "Separation of concurrent harmonic sounds : Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

de Cheveigné, A. (1993b), "Time-domain comb filtering for speech separation," ATR Human Information Processing Laboratories technical report, TR-H-016.

de Cheveigné, A., Kawahara, H., Aikawa, K., et Lea, A. (1994). "Speech separation for speech recognition," Journal de Physique IV 4, C5-545-C5-548.

de Cheveigné, A., McAdams, S., Laroche, J., et Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels : A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. 97, 3736-3748.

de Cheveigné, A. (1997). "Concurrent vowel identification III : A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857-2865.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., et Aikawa, K. (1997a). "Concurrent vowel identification I : Effects of relative level and F0 difference," J. Acoust. Soc. Am. 101, 2839-2847.

de Cheveigné, A., McAdams, S., et Marin, C. (1997b). "Concurrent vowel identification II : Effects of phase, harmonicity and task," J. Acoust.Soc. Am. 101, 2848-2856.

de Cheveigné, A. (1998). "Cancellation model of pitch perception," J. Acoust. Soc. Am. 103, 1261-1271.

de Cheveigné, A., et Kawahara, H. (1999). "Multiple period estimation and pitch perception model," Speech Communication 27, 175-185.

de Cheveigné, A., et Smith, B. (1999), "A 'sound transparency' descriptor" ISO/IEC JTC1/SC29/WG11, MPEG99/m5199.

Durlach, N. I. (1963). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. 35, 1206-1218.

Ellis, D. (1996), "Prediction-driven computational auditory scene analysis," MIT, thèse non publiée.

Ellis, D. P. W. (1997). "Computational auditory scene analysis exploiting speech-recognition knowledge.", Proc. IEEE Workshop on Apps. of Sig.

Proc. to Acous. and Audio, Mohonk.
Green, P. D., Cooke, M. P., et Crawford, M. D. (1995). "Auditory scene analysis and hidden markov model recognition of speech in noise.", Proc. IEEE-ICASSP, 401-404.
Greenberg (1997). "Understanding speech understanding : towards a unified theory of speech perception.", Proc. ESCA Workshop on the auditory basis of speech perception, Keele, 1-8.
Hartmann, W. M. (1996). "Pitch, periodicity, and auditory organization," J. Acoust. Soc. Am. 100, 3491-3502.
Hermansky, H., et Morgan, N. (1994). "RASTA processing of speech," IEEE trans Speech and Audio Process. 2, 578-589.
Holdsworth, J., Nimmo-Smith, I., Patterson, R. D., et Rice, P. (1988), "Implementing a GammaTone filter bank," MRC Applied Psychology Unit technical report, SVOS final report, annex C.
Holdsworth, J. (1990), "Two dimensional adaptive thresholding," APU AAM-HAP Report technical report, vol1, annex 4.
Holdsworth, J., Schwartz, J.-L., Berthommier, F., et Patterson, R. D. (1992). "A multi-representation model for auditory processing of sounds," in "Auditory physiology and perception," Edité par Y. Cazals, L. Demany and K. Horner, Oxford, Pergamon Press, 447-453.
Joris, P. X., et Yin, T. C. T. (1998). "Envelope coding in the lateral superior olive. III. Comparison with afferent pathways," J. Neurophysiol. 79, 253-269.
Kanadera, N., Hermansky, H., et Arai, T. (1998). "On properties of the modulation spectrum for robust automatic speech recognition.", Proc. IEEE-ICASSP, 613-616.
Lea, A. (1992), "Auditory models of vowel perception," Nottingham University, thèse non publiée.
Licklider, J. C. R. (1959). "Three auditory theories," in "Psychology, a study of a science," Edité par S. Koch, New York, McGraw-Hill, I, 41-144.
Lippmann, R. P., et Carlson, B. A. (1997). "Using missing feature theory to actively select features for robust speechj recognition with interruptions, filtering, and noise.", Proc. ESCA Eurospeech, KN-37-40.
Lyon, R. F. (1983-1988). "A computational model of binaural localization and separation," in "Natural computation," Edité par W. Richards, Cambridge, Mass, MIT Press, 319-327.
Lyon, R. (1984). "Computational models of neural auditory processing.", Proc. IEEE ICASSP, 36.1.(1-4).
Lyon, R. (1991). "Automatic gain control in cochlear mechanics," in "Mechanics and biophysics of hearing," Edited by P. Dallos, C.D.Geisler, J. W. Mathews, M. A. Ruggero et C.R.Steele, New York, Springer-Verlag.
Marin, C., et de Cheveigné, A. (1997). "Rôle de la modulation de fréquence dans la séparation de voyelles.", Proc. CFA.
Marr, D. (1982). "Representing and computing visual information," in "Artificial Intelligence : an MIT perspective," Edité par P. H. Winston et R. H. Brown, Cambridge, Mass, MIT Press, 2, 17-82.

McAdams, S. (1984), "Spectral fusion, spectral parsing, and the formation of auditory images," Stanford University, thèse non publiée.

McAdams, S. (1989). "Segregation of concurrent sounds. I : Effects of frequency modulation coherence," J. Acoust. Soc. Am. 86, 2148-2159.

Meddis, R. (1988). "Simulation of auditory-neural transduction : further studies," J. Acoust. Soc. Am. 83, 1056-1063.

Meddis, R., et Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91, 233-245.

Mellinger, D. K. (1991), "Event formation and separation in musical sound," Stanford Center for computer research in music and acoustics, thèse non publiée.

Meyer, G., et Berthommier, F. (1996). "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, 212-215.

Meyer, G. F., Plante, F., et Berthommier, F. (19976). "Segregation of concurrent speech with the reassigned spectrum", Proc. IEEE ICASSP, 1203-1206.

Morris, A. C., Cooke, M. P., et Green, P. D. (1998). "Some solutions to the missing feature problem in data classification, with application to noise robust ASR", Proc. ICASSP, 737-740.

Nadeu, C., Pachès-Leal, P., et Juang, B.-H. (1997). "Filtering the time sequences of spectral parameters for speech recognition," Speech Comm. 22, 315-332.

Nakatani, T., Okuno, H. G., et Kawabata, T. (1995a). "Residue-driven architecture for computational auditory scene analysis.", Proc. IJCAI, 165-172.

Nakatani, T., Goto, M., Ito, T., et Okuno, H. G. (1995b). "Multi-agent based binaural sound stream segregation.", Proc. IJCAI Workshop on Computational Auditory Scene Analysis, 84-91.

Nakatani, T., Goto, M., et Okuno, H. G. (1996). "Localization by harmonic structure and its application to harmonic stream segregation.", Proc. IEEE ICASSP, 653-656.

Nakatani, T., Kashino, K., et Okuno, J. G. (1997). "Integration of speech stream and music stream segregations based on a sound ontology.", Proc. IJCAI Workshop on computational auditory scene analysis, Nagoya, 25-32.

Okuno, H. G., Nakagawa, Y., and Kitano, H. (1999a). "Incorporating visual information into sound source separation.", Proc. International workshop on Computational Auditory Scene Analysis, .

Okuno, H. G., Ikeda, S., and Nakatani, T. (1999b). "Combining independant component analysis and sound stream segregation.", Proc. International workshop on computational auditory scene analysis, .

Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, 911-918.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., et Allerhand, M. (1992). "Complex sounds and auditory images," in "Auditory physiology and perception," Edité par Y. Cazals, K. Horner et L. Demany, Oxford, Pergamon Press, 429-446.

Patterson, R., Anderson, T. R., et Francis, K. (1996). "Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition.", Proc. Workshop on the auditory basis of speech perception, Keele, 245-252.

Rosenthal, D. F., and Okuno, H. G. (1997). "Computational auditory scene analysis," Lawrence Erlbaum.

Scheffers, M. T. M. (1983), "Sifting vowels," Gröningen University, th.

Seneff, S. (1985), "Pitch and spectral analysis of speech based on an auditory synchrony model," MIT, thèse non publiée (technical report 504).

Shamma, S. A. (1985). "Speech processing in the auditory system I : The representation of speech sounds in the responses of the auditory nerve," J. Acoust. Soc. Am. 78, 1612-1621.

Slaney, M. (1993), "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer technical report, 35.

Slaney, M. (1995). "A critique of pure audition.", Proc. Computational auditory scene analysis workshop, IJCAI, Montreal.

Summerfield, Q., Lea, A., et Marshall, D. (1990). "Modelling auditory scene analysis : strategies for source segregation using autocorrelograms," Proc. Institute of Acoustics 12, 507-514.

Summerfield, Q., et Culling, J. F. (1992). "Auditory segregation of competing voices : absence of effects of FM or AM coherence," Phil. Trans. R. Soc. Lond. B 336, 357-366.

Summerfield, Q. (1992). "Roles of harmonicity and coherent frequency modulation in auditory grouping," in "The auditory processing of speech : from sounds to words," Edité par M. E. H. Schouten, Berlin, Mouton de Gruyter, 157-166.

Summerfield, Q., et Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency.", Proc. 124th meeting of the ASA, 2317(A).

Wang, A. L.-C. (1995), "Instantaneous and frequency-warped signal processing techniques for auditory source separation," CCRMA (Stanford University), thèse non publiée.

Warren, R. M. (1970). "Perceptual restoration of missing speech sounds," Science 167, 392-393.

Warren, R. M., Obusek, C. J., et Ackroff, J. M. (1972). "Auditory induction : perceptual synthesis of absent sounds," Science 176, 1149-1151.

Weintraub, M. (1985), "A theory and computational model of auditory monaural sound separation," Stanford University, thèse non publiée.

Yost, W. A., Dye, R. H., et Sheft, S. (1996). "A simulated "cocktail party" with up to three sound sources," Perception and Psychophysics 58, 1026-1036.