Analyse de scène auditive et parole

Alain de Cheveigné

Ircam - CNRS
1 place Igor Stravinsky, 75004, France
Tél.: ++33 1 44 78 48 46 - Fax: ++33 1 44 78 15 40
Mél: cheveign@ircam.fr - http://www.ircam.fr/pcm/cheveign

ABSTRACT

Auditory Scene Analysis (ASA) is a more basic competence than speech perception, phylogenetically more ancient, but the two share important relations. Speech communication often occurs in presence of interfering noises and voices, and therefore depends on segregation mechanisms for reception. Speech has often been used as a stimulus to investigate ASA phenomena, but in some respects it appears to escape from basic principles. The most interesting potential application of computational auditory scene analysis (CASA) is in speech recognition systems.

1. INTRODUCTION

Pour guider son action et faciliter sa survie, l'organisme construit un modèle du monde qui l'entoure à l'aide d'informations fournies par ses sens. Lorsque l'environnement est complexe, cette information doit être triée et distribuée parmi les éléments du modèle. Si le tri se fait bien, le modèle est fidèle, l'action efficace, et les chances de survie bonnes. Dans le domaine de l'audition, cette opération s'appelle *analyse de scène auditive* (ASA). L'analyse de scène a été pratiquée par nos ancêtres avant qu'ils se mettent à parler, et on peut penser que ses mécanismes sont pour l'essentiel génériques et non spécifiques à la parole. ASA et parole entretiennent néanmoins des rapports privilégiés pour plusieurs raisons, qui tiennent à l'importance que revêt pour nous la communication parlée.

• Même si la parole n'était pas le stimulus le plus important pour nos lointains ancêtres, elle l'est devenue pour nous. La voix d'un locuteur est souvent l'objet des processus ASA, et celle d'un locuteur concurrent une source de bruit masqueur.

• La parole semble paradoxalement échapper à certains principes qui régissent l'ASA de sons plus simples. Par exemple on s'attendrait à ce que l'irrégularité des sons qui composent la parole s'oppose à sa fusion en un flux unique, mais il n'en est rien.

• La parole (plus ou moins stylisée) a souvent été utilisée comme matériau d'expérience pour explorer l'ASA. L'identification des voyelles concurrentes constitue notamment un paradigme puissant pour l'étude des processus d'organisation simultanée.

• Des applications importantes relèvent de la parole: reconnaissance de la parole en milieu bruité, prothèses auditives, implants cochléaires. C'est ainsi que s'est développée la discipline d'*analyse de scène auditive computationnelle* (CASA), qui cherche à reproduire les opérations de l'ASA par des moyens computationnels.

L'article est en trois parties. La première résume brièvement quelques principes de l'ASA. La deuxième discute du rôle du voisement, un indice de ségrégation parmi les plus importants. La troisième passe en revue quelques tentatives d'application de l'analyse de scène computationelle à la reconnaissance de la parole.

2. LES PRINCIPES DE L'ASA

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*. La phonétique décrit de même les propriétés acoustiques d'une voix isolée. En pratique nous percevons souvent les sons parmi une cacophonie de voix et bruits superposés. Chaque oreille reçoit des ondes provenant d'une multitude de sources, mais 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. Les modèles classiques, conçus pour traiter une source isolée, ne sont pas suffisants pour expliquer la perception dans ce cas.

Helmholtz [Hel87] déjà se demandait comment on pouvait percevoir les qualités individuelles des instruments de l'orchestre, mais il a fallu attendre le travail de Bregman [Bre90] pour que l'analyse de scène auditive devienne un sujet d'étude à part entière. Pour Bregman, le problème de l'émergence de sources subjectives (flux, ou "streams") est principal, et la détermination de leurs qualités secondaire, puisque logiquement le premier est un préalable au second. Pour élaborer sa théorie, Bregman s'est appuyé sur l'analogie avec l'analyse de scène en vision, et les principes de la psychologie Gestalt.

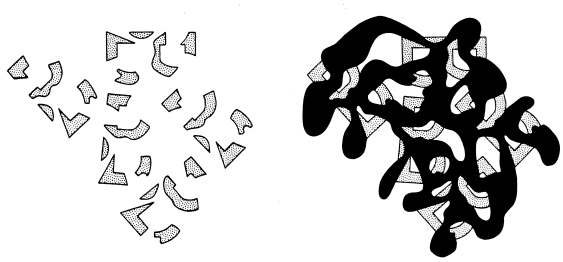

Figure 1. Analyse de scène visuelle. A gauche, les fragments paraissent inorganisés. A droite, la présence d'une forme masquante permet leur regroupement perceptif. L'ASA cherche des principes analogues pour l'organisation du monde sonore. (D'après [Bre90]).

Avec le développement de l'Informatique et de l'Intelligence Artificielle sont apparues des tentatives d'*Analyse de Scène Auditive Computationnelle* (CASA) [Lyo83, Wei85, Coo91, Mel91, Bro92, Wan92, Ell96]. 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 [Mar82], a joué un rôle déterminant dans l'élaboration de ces modèles.

Les mots de l'ASA

Il faut distinguer la *source acoustique* de l'entité perceptive qui lui correspond après analyse, qu'on désigne par *flux* ("stream") ou par les termes plus ambigus d'*objet* ou *événement* perceptif. On oppose la *fusion* (groupement) à la *scission* (ségrégation) selon que l'information acoustique évoque un ou plusieurs flux. On parle d'organisation *simultanée* ou *séquentielle* selon que les sources à analyser se manifestent de façon simultanée ou séquentielle. Enfin, on distingue les processus *primitifs* (montants, ou "bottom-up") des processus *à base de schémas* (descendants, ou "top-down").

Organisation simultanée

Il arrive que des sources distinctes se manifestent en même temps (ou avec un chevauchement temporel). Les corrélats acoustiques qui parviennent aux oreilles sont intimement mêlés, mais malgré cela nous entendons parfois plusieurs objets perceptifs, correspondant chacun à une source, plutôt que le flux unique qu'évoquerait le son d'une source unique. Quels aspects du signal acoustique font qu'il évoque la perception d'un objet plutôt que deux ?

Un premier facteur de cohésion est la *simultanéité d'attaque*, et plus généralement la communauté de variation des composantes fréquentielles du son. Lorsque toutes démarrent en même temps on tend à percevoir une source unique. Une asynchronie d'attaque évoque la perception d'objets multiples. C'est un exemple du principe plus général de *destin commun*.

Un deuxième facteur important est l'*harmonicité*. Lorsque les partiels suivent une série harmonique unique (sons de certains instruments, parole voisée) le son évoque une source unique. Dans le cas contraire (la "polypériodicité" de Marin [Mar91]), le stimulus paraît provenir de plusieurs sources.

Un troisième facteur est la *corrélation binaurale*. Si les composantes du son ont toutes la même relation binaurale, leur fusion est favorisée. Des disparités entre composantes peuvent évoquer des sources multiples, distinctes dans l'espace.

L'organisation auditive va au-delà d'une simple perception de multiplicité, puisque dans une certaine mesure une source parmi des sources concurrentes est perceptible comme si elle était isolée. Pour des sources qui se chevauchent dans le temps, la ségrégation perceptive est gouvernée par la règle *ancien plus nouveau* (old plus new). A l'apparition d'une nouvelle source, le spectre est analysé en défalquant la contribution estimée de l'ancienne (supposée toujours présente). Bien entendu, un tel mechanisme doit savoir faire la différence entre les variations de spectre dues à l'apparition d'une nouvelle source, et celles qui tiennent de la nature intrinsèquement variable d'une source unique.

La structure harmonique de sources concurrentes est aussi exploitée pour la ségrégation, comme on le verra en détail plus loin. Bregman [Bre90] donne d'autres exemples de traits qui gouvernent l'analyse simultanée. La *modulation de fréquence* a souvent été évoquée comme un exemple du principe de destin commun. Pour comprendre cette notion, il suffit d'imaginer 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. L'idée est attrayante, mais on verra qu'en fait cet indice joue un rôle mineur.

Organisation séquentielle

Il arrive que des sources se manifestent de façon répétée dans le temps, et évoquent la perception d'une suite d'entités cohérente (flux), distincte des autres sons de l'environnement. C'est le cas d'une voix, d'une ligne mélodique, d'une succession de crissements de pas dans la neige, etc. Quels aspects d'une succession de sons font qu'ils se groupent en un flux unique, plutôt qu'en plusieurs flux parallèles ?

Un premier facteur est la similitude. Des sons disparates tendent à former des flux multiples, alors que des sons proches forment un flux unique. Il peut s'agir d'une similitude de hauteur, de timbre, de position spatiale, de sonie, etc., mais le facteur le plus important est la similitude des activités évoquée dans le système auditif périphérique [Har91] et donc du contenu spectral.

Un deuxième facteur est la vitesse de présentation des sons. Un taux lent favorise la formation d'un flux unique, alors qu'un taux rapide favorise la scission en des flux multiples.

Bregman interprète ces facteurs en termes de principes Gestalt tels que celui de destin commun (ou origine commune): les sons d'une même source ont toutes les chances d'être semblables, ou de varier lentement, et il est naturel d'interpréter une variation importante et/ou rapide comme l'intervention d'une nouvelle source. Bregman note cependant le paradoxe de la parole, dont les variations (notamment les transitions consonantiques) sont grandes et rapides, sans que cela compromette la cohérence de la voix.

Fusion vs scission

La psychoacoustique classique considère des sources uniques et leur attribue l'ensemble des représentations de bas niveau (physiologiques) et de haut niveau (perceptives) qu'elles évoquent. L'ASA suppose des sources multiples et donc une *décomposition* des représentations. On peut imaginer que cette décomposition a lieu selon des dimensions tonotopiques, "périodotopiques", "spatiotopiques", etc., dans des cartes supposées exister dans le système auditif central. Si une telle décomposition (scission) est

possible, alors il faut expliquer pourquoi elle ne se fait pas systématiquement, c'est-à-dire qu'il faut expliquer la cohésion dans le cas d'une source unique. Fusion et scission sont les deux faces d'une même pièce.

Processus primitifs vs schémas

Bregman distingue les processus d'organisation primitifs, qui font intervenir des mécanismes ascendants ("bottom-up"), de ceux qui font intervenir des attentes produites par le contexte, ou des "schémas" présents chez l'auditeur, et qui pourraient faire intervenir des processus descendants ("top-down"). Les processus primitifs joueraient pour tout type de son, alors que les schémas seraient spécifiques de sons particuliers, par exemple la parole. Les schémas qui interviennent dans le décodage de la parole pourraient ainsi jouer un rôle dans l'organisation auditive de scènes comprenant une voix. Cependant il est difficile de démontrer que le schéma est intervenu dans le processus d'organisation lui-même, plutôt que dans une phase ultérieure d'interprétation des données organisées.

Un exemple souvent cité est celui de la "restauration phonémique" [War70]. Un segment phonémique de parole est excisé et remplacé par un bruit, mais l'auditeur croit entendre le segment manquant. Plus étrange, il est incapable de situer avec précision l'interruption au sein du mot. Lorsque l'interruption introduit une ambiguïté, la restauration peut dépendre du contexte qui la précède (ou même qui la suit). On voit parfois dans ce phénomène l'indice d'un mécanisme de bas niveau de partition de l'information acoustique, ou de synthèse de la partie manquante, à partir d'un schéma. On peut aussi l'attribuer plus prosaïquement au mécanisme d'*interprétation* des données présentes, exploitant le fait que la présence de la parole derrière le bruit est une hypothèse plus probable que celle de son absence.

Outre les processus à base de schémas, la parole devrait aussi bénéficier des processus primitifs génériques applicables à tout son, mais l'hypothèse a aussi été défendue que la parole est "spéciale" et leur échappe [Rem94].

3. VOISEMENT ET SEGREGATION

Les aspects de l'ASA qui relèvent de la parole (et inversement) sont nombreux. Plutôt que d'entreprendre une revue générale qu'on peut trouver ailleurs [Bre90, CoE00], je développerai un aspect de l'ASA que j'ai exploré plus en profondeur.

Une particularité de la voix humaine est qu'elle comprend des portions *voisées* à la structure approximativement périodique ou harmonique. Le voisement joue un rôle bien connu comme vecteur d'informations prosodiques liées à la fréquence de vibration des cordes vocales (F0). Cherry [Che53] a suggéré qu'une différence de F0 moyenne entre des voix d'homme et de femme pourrait aussi faciliter la compréhension lorsque ces voix sont superposées. La F0 participerait au fameux "effet cocktail" (cocktail party effect). Brokx et Nooteboom [Bro82] ont confirmé cette intuition, en montrant que l'intelligibilité est meilleure lorsque les F0 sont dans des plages différentes plutôt que superposées. On peut interpréter ce résultat de plusieurs façons.

Une première interprétation est que la F0 sert à "suivre" une voix au cours du temps. La F0 décrit une courbe relativement continue dans les parties voisées, et entre parties voisées les variations sont limitées et partiellement prévisibles. Dans une expérience de Darwin [Dar75] des auditeurs devaient répéter les mots d'une voix présentée à une oreille en ignorant une voix parasite dans l'autre. Lorsque la F0 fut brusquement intervertie entre les voix, les sujets répétèrent quelques mots de l'oreille opposée, en suivant donc la continuité de la F0. Dans une autre expérience, un changement brusque de F0 dans une transition entre voyelles évoquait la perception de voix multiples prononçant des consonnes [Dar77].

Une deuxième interprétation des résultats de Brokx et Nooteboom est qu'une différence de F0 (ΔF0) facilite la ségrégation dans les parties voisées qui se chevauchent. Pour tester cette interprétation, Scheffers [Sch83] a assemblé des mélanges de voyelles synthétiques concurrentes, et demandé à des sujets de les identifier. L'identification était meilleure lorsque les F0 des voyelles étaient différentes, ce qui confirme donc aussi cette deuxième interprétation. Par la suite, de nombreux auteurs ont utilisé le paradigme des "voyelles doubles" pour tenter de comprendre les mécanismes de ségrégation (Figure 2).

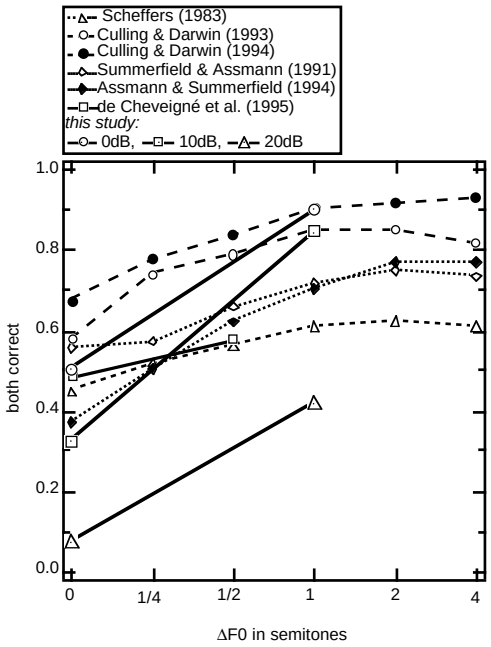

Figure 2. Taux d'identification de paires de voyelles en fonction de la différence entre leurs F0. L'identification profite de mécanismes de ségrégation qui exploitent la structure harmonique des voyelles [deC97a].

Une hypothèse est que la *structure harmonique* de la parole voisée est exploitée par un mécanisme de ségrégation. Les partiels des deux voix suivent des séries harmoniques distinctes, ce qui pourrait faciliter leur tri (sauf bien sûr à ΔF0=0). Par exemple Parsons [Par76] a proposé une méthode de séparation de voix fondée sur des spectres de Fourier calculés sur des fenêtres de 51.2 ms. Les pics du spectre étaient regroupés dans des séries harmoniques, et attribués à l'une ou l'autre voix. Scheffers [Sch83], puis Assmann et Summerfield [AsS90] s'en sont inspirés pour

élaborer un modèle de ségrégation fondé sur l'analyse fréquentielle de la cochlée, mais ils ont constaté une résolution trop faible. La sélectivité *fréquentielle* cochléaire étant insuffisante pour isoler les partiels de chaque série, les modèles plus récents s'orientent vers des mécanismes *temporels* (voir plus loin).

A regarder de plus près, pour extraire une voix on dispose de *deux* structures harmoniques: celle de la voix cible et celle de sa concurrente. Servent-elles toutes les deux, ou seulement une, et si oui, laquelle ? Deux mécanismes sont possibles: le *renforcement harmonique* par lequel une voix est favorisée par son harmonicité, ou l'*annulation harmonique*, par laquelle elle est favorisée par l'harmonicité de son concurrent. Le premier est attrayant car il s'applique quel que soit le bruit, périodique ou non, mais il n'est utile que pour les parties voisées de la cible. L'annulation harmonique, elle, marche pour les parties voisées et non voisées, mais seulement si le concurrent est harmonique. Zissmann et Weinstein [ZiW83] ont simulé ces deux stratégies avec de la parole mélangée, en supprimant la voix concurrente soit lorsque cette voix était voisée, soit lorsque la cible était voisée. L'intelligibilité était meilleure dans le premier cas, ce qui indique que la stratégie d'annulation est plus utile (ou le serait si on pouvait implémenter les deux stratégies de façon parfaite). Du fait des apériodicités de la parole naturelle l'implémentation de l'une et l'autre stratégie est forcément imparfaite, mais on peut montrer que cela affecte moins l'annulation que le renforcement [deC93].

On peut utiliser le paradigme expérimental de Scheffers pour connaître la stratégie utilisée par le système auditif, à condition de mesurer séparément l'identification de chaque voyelle d'une paire (le paradigme classique compte les réponses pour lesquelles les voyelles sont simultanément correctes). Pour des mélanges de voyelles voisées et chuchotées, Lea [Lea92] a constaté que la ségrégation bénéficiait à la seule composante chuchotée. D'autres auteurs ont confirmé ce résultat en montrant que le facteur qui détermine la ségrégation est l'*harmonicité de la voyelle concurrente* [SuC92a, deC95, deC97a, deC97b]. L'harmonicité propre de la cible ne lui est d'aucun secours, résultat d'autant plus étonnant que plusieurs algorithmes de séparation de parole et modèles CASA utilisent l'harmonicité de la voix cible.

Les modèles et méthodes peuvent se classer selon qu'ils adoptent la stratégie de renforcement ou celle d'annulation (que favorisent les arguments précédents) [deC93a, deC95]. Parmi les modèles d'annulation, celui de Meddis et Hewitt [MeH92] est le mieux connu. Ce modèle fait un tri parmi les canaux périphériques issus de la cochlée. Les trains d'impulsions du nerf auditif sont soumis à un processus d'autocoincidence neuronale selon le modèle de Licklider [Lic56] qui permet de mesurer leur périodicité. Les canaux dont la périodicité ne correspond *pas* à celle qui domine le stimulus sont regroupées et utilisées pour extraire la voix la plus faible. Alors que le modèle de Parsons/Scheffers nécessitait une résolution spectrale à l'échelle des *partiels*, celui de Meddis et Hewitt nécessite seulement une résolution à l'échelle des *formants*. Une différence des structures formantiques fait que différents canaux sont dominés par différentes voix, même lorsque les voix sont de même amplitude.

En revanche, si les amplitudes sont suffisamment différentes, la voix la plus forte risque de dominer *tous* les canaux périphériques, dans quel cas ce mécanisme de ségrégation ne marche pas et on ne prévoit pas d'effet de $\Delta F0$. Cela fournit un moyen de tester le modèle de Meddis et Hewitt. Les premières expériences "voyelles doubles" utilisaient des voyelles égalisées en amplitude, sonie ou "force d'excitation" (selon l'expérience), mais des expériences plus récentes ont introduit des différences de niveau entre voyelles [deC99b]. L'effet de $\Delta F0$ est plutôt renforcé pour la composante faible, et reste important jusqu'à -25 dB (Figure 3). Dans ces conditions, si on modélise le pattern de dominance des canaux périphériques, on s'aperçoit que *tous* sont dominés par la voyelle la plus forte. Le modèle de Meddis et Hewitt ne peux donc pas expliquer ces effets de $\Delta F0$.

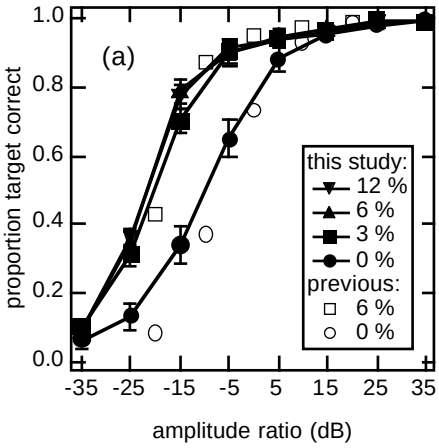

Figure 3. Taux d'identification en fonction de l'amplitude d'une voyelle relative à la sa concurrente, pour des F0 égaux (cercles) ou différents (autre symboles) [deC99b]

A côté de modèles spectraux [Sch83, Par76] et spectro-temporels [MeH92, AsS90], on peut imaginer des modèles purement *temporels* de ségrégation harmonique. Le *filtre d'annulation neuronal* est un circuit hypothétique comprenant un neurone "porte" muni de deux synapses, l'un excitateur alimenté directement, et l'autre inhibiteur alimenté via une ligne à retard. Toute impulsion arrivant par le chemin direct est transmise, sauf si une impulsion arrive simultanément par le chemin retardé. L'effet du filtre est de modifier la statistique des intervalles inter-impulsions, et il se trouve que cela suffit à supprimer la contribution d'une source dont la période serait égale au retard [deC93]. A l'aide de ce filtre on peut constituer un modèle de perception de voyelles concurrentes qui explique l'ensemble des données expérimentales [deC97, deC99d]. Il peut aussi servir de base à des modèles de perception de la hauteur [deC98, deC99a, deC99f].

Avant qu'on ne découvre le peu d'effet de l'harmonicité d'une cible, on pouvait imaginer que la structure harmonique pourrait grouper ensemble les formants d'une voix, en les étiquetant d'une même périodicité. Culling et Darwin [CuD93] ont présenté à des sujets des paires de voyelles synthétisées de telle façon que le formant F1 de l'une ait la même périodicité que les formants supérieurs de l'autre. Cette manipulation eut peu d'effets sur la ségrégation, ce qui suggère le manque d'efficacité d'une harmonicité commune pour grouper ensemble les formants

d'une voyelle. Ces résultats et d'autres [CuD94] suggéraient même que l'harmonicité pourrait ne pas avoir du rôle du tout, puisqu'ils révélaient une ségrégation entre des voyelles dont ni l'une ni l'autre n'est parfaitement harmonique. Cela amena ces auteurs à proposer que, pour des ΔF0 petits (<6%) la ségrégation pourrait être due à un simple effet de battements [CuD94, AsS94], sans aucun lien avec la structure harmonique. Cette hypothèse a l'attrait de ne pas nécessiter une résolution fréquentielle fine. Une hypothèse apparentée, proposée à la même époque (l'hypothèse PPA de [AsS94]), est que la structure temporelle particulière de la période de voisement (impulsion glottique suivie d'un intervalle d'énergie faible) fournirait une "fenêtre" qui faciliterait la perception de la deuxième voyelle.

Ces deux hypothèses (battements et PPA) impliquent, si elles sont vraies, une certaine dépendance sur la *phase* des voyelles. Par exemple une manipulation de phase qui chamboule la structure temporelle intra-période devrait mettre en échec le mécanisme PPA. Deux séries d'expériences [deC97b, deC99d] ont montré l'absence quasi-total d'effets du spectre de phase, ce qui permet d'écarter ces deux hypothèses, même pour des ΔF0 faibles. C'est seulement à des F0 très bas (50 Hz) qu'on constate des effets de l'alignement temporel entre périodes [AsS90].

Dans les expériences de voyelles doubles, l'identification croit en fonction de ΔF0 pour atteindre rapidement un plateau à partir d'environ 1/2 ton (6%). Elle décroît ensuite à l'octave, un phénomène qu'avait déjà noté Brokx et Nooteboom. En dessous de 6% l'effet de ΔF0 décroît, mais reste mesurable jusqu'à 0.4%, ou 1/16e de ton [deC99d]! Ce résultat indique la très grande finesse de résolution fréquentielle ou temporelle du mécanisme de ségrégation. Par exemple si on retient le modèle d'annulation neuronal, il doit pouvoir exploiter des disparités de l'ordre de 30 μs.

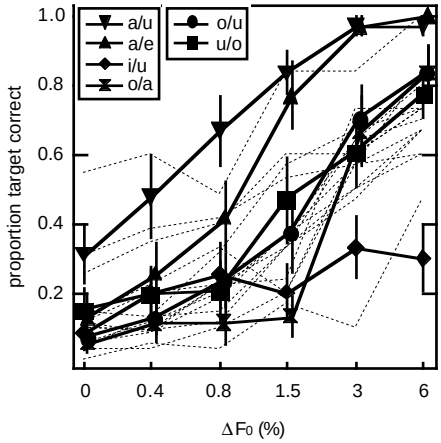

Figure 4. Taux d'identification en fonction de ΔF0 pour différentes paires de voyelles (20 combinaisons des 5 voyelles du japonais). Une ΔF0 de 0.4% suffit à produire un effet significatif [deC99d].

Des ΔF0 non-nuls facilitent la ségrégation, mais de nombreux auteurs ont noté avec étonnement que l'identification à ΔF0=0 dépasse de loin le hasard. Le spectre du stimulus est pourtant très différent de l'une ou l'autre voyelle. Pour mieux comprendre les facteurs qui déterminent l'identification, j'ai utilisé des paires de voyelles synthétisées avec le même spectre de phase (pour simplifier la sommation), et une large plage d'amplitudes relatives. L'analyse des résultats par paire a montré qu'une voyelle est identifiée dès lors que ses formants F1 et F2 sont saillants (les formants supérieurs semblent moins importants sans qu'on puisse écarter tout rôle). L'identification n'est pas perturbée par la saillance des formants du concurrent. Les deux voyelles peuvent d'ailleurs partager un formant: le principe d'allocation exclusive [Bre90] ne joue donc pas. Ces conclusions furent tirées à ΔF0=0. L'effet d'une ΔF0 non-nul est apparemment de renforcer encore la saillance des indices formantiques.

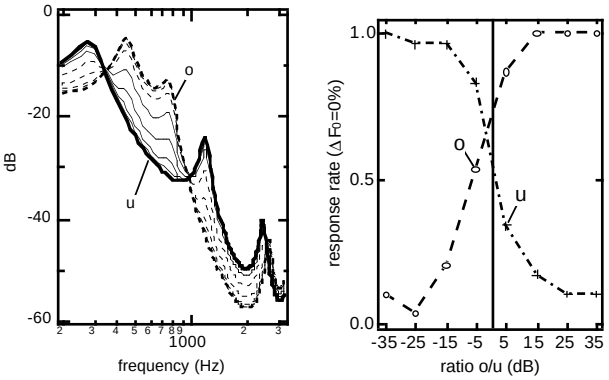

Figure 5. A gauche: enveloppe spectrale de /o/ (trait pointillé gros), de /u/ (trait continu gros) et de stimulis hybrides /o/+/u/ pour des rapports d'amplitudes allant de +35 à -35 dB par pas de 10 dB (traits fins). A droite: taux d'identification des voyelles /o/ et /u/ pour ces stimulis hybrides à ΔF0=0. L'analyse des données pour l'ensemble des paires de voyelles permet de comprendre quels indices supportent l'identification des voyelles [deC99a].

La *largeur de bande des formants* est connue pour avoir peu d'effet sur le timbre ou l'identité d'une voyelle. En revanche elle se révèle jouer un rôle important dans la ségrégation. Une voyelle à formants étroits est plus masquante, et plus résistante au masquage, qu'une voyelle à formants larges [deC99e], et cet effet s'ajoute aux effets éventuels de la ΔF0. Dans la mesure où les formants s'affinent lorsqu'on raccourcit la phase d'ouverture de la glotte, ce facteur pourrait varier sous contrôle musculaire et jouer ainsi un rôle actif dans les situations de compétition verbale (cette hypothèse reste à vérifier).

Un facteur qui s'apparente à la ΔF0 est la *modulation de fréquenc*e (FM), souvent citée comme exemple du principe Gestalt de destin commun [Bre90]. Une modulation commune des partiels d'un son permettrait (selon ce principe) leur regroupement, et leur ségrégation d'avec les partiels d'un fond statique ou modulé de façon différente. Des essais informels ont montré des effets spectaculaires, parfois repris dans des démos, mais dans chaque exemple la FM avait pour effet secondaire de créer une ΔF0 instantanée, suffisante pour expliquer à elle seule l'effet. Plusieurs études ont cherché à mettre en évidence un effet spécifique de la FM au delà des ΔF0 induits, sans succès [McA89, Dem90, SuC92b Car94, deC96, MaM97].

Pour résumer, l'harmonicité liée au voisement est l'un des facteurs les plus importants de l'analyse auditive de scènes

comprenant de la parole. Les effets sont mesurables sur une large plage d'amplitudes (pour des voyelles jusqu'à -25 dB par rapport à leur concurrent), ce qui argue en faveur de leur utilité écologique. Des différences de F0 de 6% ou plus sont pleinement exploitables, et des effets sont mesurables pour des $\Delta F0$ bien plus faibles. Cela confirme les intuitions des premiers chercheurs tels que Cherry. En revanche certains aspects sont étonnants et vont à l'encontre des intuitions et principes Gestalt. La structure harmonique de la voix cible ne joue guère de rôle dans sa ségrégation (sauf peut-être pour maintenir la continuité de la voix), et la modulation de fréquence a un effet faible ou nul.

4. CASA ET RECONNAISSANCE DE LA PAROLE

Notre système auditif exploite l'harmonicité liée au voisement pour améliorer l'identification d'une voix masquée par une voix concurrente. C'est l'un des mécanismes qui font que la perception d'un auditeur humain est plus robuste face au bruit que les systèmes de reconnaissance de la parole [Lip97]. Il est évident qu'il serait utile de reproduire ces mécanismes artificiellement. L'*analyse de scène auditive computationelle* (CASA) est une approche possible (pas la seule) pour y parvenir.

Weintraub [Wei85] le premier a tenté d'utiliser un modèle du système auditif pour améliorer la reconnaissance de la parole en présence d'une voix concurrente. Le modèle, proche de celui de Meddis et Hewitt [MeH92], s'appuyait sur une analyse par autocorrélation des canaux d'un banc de filtres. Weintraub travaillait à partir des idées de Lyon [Lyo83, Lyo84], qui lui-même s'inspirait de Licklider [Lic56]. Les canaux étaient assignés à une voix ou l'autre selon leur périodicité, puis la parole resynthétisée et présentée à un système de reconnaissance. Les taux de reconnaissance obtenus n'étaient pas excellents, mais le système de Weintraub a inspiré de nombreux efforts depuis.

Cooke [Coo91], Mellinger [Mel91], Brown [Bro92] et Ellis [Ell96] ont tous essayé de trouver des modèles physiologiquement plausibles capables d'exploiter le voisement, et d'autres indices, pour la ségrégation. Dans la plupart des cas, l'objectif fixé était la *resynthèse* de voix séparées. Cette objectif, qui correspond à l'idée naïve de "séparation" des voix, a l'avantage de permettre une évaluation immédiate par écoute ou mesure du rapport signal-sur-bruit [Bro92], mais on peut se demander s'il s'agit d'un objectif raisonnable. Une première remarque est qu'une resynthèse parfaite est impossible dans le cas général (du fait de l'indétermination du mélange). Un critère de qualité perceptive de la resynthèse risque donc d'être irréaliste. Une deuxième remarque est que la resynthèse n'a pas sa place dans un modèle du système auditif qui, lui, ne resynthétise pas. Une troisième remarque est que la resynthèse n'est ni nécessaire, ni forcément souhaitable pour une application de reconnaissance de la parole [Sla95, C0G00]. Certes, la resynthèse permet une architecture modulaire, mais une intégration plus étroite entre ségrégation et reconnaissance est souhaitable pour pleinement profiter de l'analyse de scène.

Cooke et ses collègues ont investi beaucoup d'effort dans cette question, en particulier en développant la *théorie des données manquantes* (TDM) pour gérer les données incomplètes fournies par un système CASA. Une ségrégation parfaite étant impossible, les données sont incomplètes, mais si les parties manquantes sont connues il est possible d'en tenir compte dans l'étape de reconnaissance [AhT93, CMG97, CGJ00, LiC97, deV99]. Deux méthodes sont proposées. L'une attribue un poids nul aux parties manquantes (méthode des "marginales"), l'autre les interpole à partir des données présentes à l'aide d'un modèle (méthode des "imputations"). La première est plus efficace, mais la deuxième a l'avantage de fournir des données "completes" (utiles par exemple pour une resynthèse éventuelle).

La TDM est sans doute une clé pour l'utilisation effective des modèles CASA pour la reconnaissance. Elle est d'une utilité plus large pour exploiter des données distordues ou de fiabilité inhomogène, dès lors que la distortion ou la fiabilité sont connues. C'est le cas par exemple de la distortion convolutive d'un réseau de microphones ou d'une analyse en composantes indépendantes. C'est le cas aussi pour la fusion de données multimodales [RoD98]. La TDM est proche par son esprit du modèle FLMP de Massaro [Mas90], et c'est un paradigme intéressant pour l'élaboration de modèles perceptifs [deC99c].

A noter que l'objectif de *reconnaissance* de la parole à la sortie d'un système CASA est plus réaliste que celui de *resynthèse* des voix séparées. Celle-ci est difficile du fait de l'indétermination notée plus haut, et aussi du fait de l'exigence des auditeurs. Une resynthèse de qualité n'est possible qu'à l'aide de modèles normatifs ou de production sophistiqués (dont les paramètres seraient contraints par les données incomplètes).

Pour résumer, la reconnaissance de la parole est l'application potentielle la plus intéressante des modèles d'analyse de scène. Il est certain que les systèmes profiteraient d'une robustesse semblable aux auditeurs humains. Inversement, les techniques de reconnaissance constituent l'un de nos meilleurs modèles pour la perception de la parole [Moo96], et les progrès qui seront faits pour lui conférer une résistance aux interférences seront aussi bien des progrès dans la compréhension de nos mécanismes perceptifs.

5. CONCLUSION

L'analyse de scène est indissociable de la perception en général, et de la parole en particulier. La parole a servi de stimulus privilégié dans la découverte de ses principes, même si elle paraît parfois leur échapper et obéir à des principes qui lui sont propres. Parmi les indices d'organisation importants figure l'harmonicité, lié au voisement, qui permet une amélioration importante de l'intelligibilité. Les systèmes de reconnaissance de la parole auraient bien besoin de profiter eux aussi d'indices de ce type, mais pour l'instant les tentatives de constituer des systèmes d'analyse computationelle (CASA) sont restées à l'état expérimental. Il faut espérer que des progrès tels que la théorie des données manquantes, liés à une meilleure compréhension des mécanismes physiologiques de l'ASA, permettront des avancées dans un avenir proche.

6. POUR EN SAVOIR PLUS

L'ouvrage très complet de Bregman [Bre90] résume la recherche en ASA jusqu'en 1990, et a inspiré l'essentiel de ce qui s'est fait depuis. Des revues plus récentes sont celles de Darwin et Carlyon [DaC95] et Cooke et Ellis [CoE00]. Les articles de Cherry [Che53], Brokx et Nooteboom [Bro82], Warren [War70], Cutting [Cut76], Darwin [Dar81] sont des classiques à lire. Le versant computationnel (CASA) est résumé par Cooke et Ellis [CoE00] (voir aussi [deC00]). La fraîcheur des pionniers est à retrouver dans les thèses de Scheffers [Sch83], Weintraub [Wei85], Cooke [Coo91], Brown [Bro92], Lea [Lea92], Mellinger [Mel91], Ellis [Ell96]. Pour les travaux récents on peut consulter les actes des workshop CASA satellites de l'IJCAI, dont certains ont été publiés [RoO97, SpC99]. Pour les applications de reconnaissance de la parole, et en particulier la théorie des données manquantes, lire Cooke et Green [CoG00], Cooke, Green, Josifovski et Vizinho [CGJ00]. D'autres articles intéressants sont ceux de Lippmann [Lip97, LiC97], Hermansky [Her98]. Pour quelques ressources en ligne, consultez la page: http://www.ircam.fr/pcm/cheveign/sh/casa.html

BIBLIOGRAPHIE

[AhT93] Ahmad, S., and Tresp, V. (1993). "Some solutions to the missing feature problem in vision," in "Advances in Neural Information Processing Systems 5," Edited by S. J. Hanson, J. D. Cowan and C. L. Giles, San Mateo, Morgan Kaufmann, 393-400.

[AsS90] Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, 680-697.

[AsS94] Assmann, P. F., and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. 95, 471-484.

[BeM95] Berthommier, F., and Meyer, G. (1995). "Source separation by a functional model of amplitude demodulation.", Proc. ESCA Eurospeech, 135-138.

[Bre90] Bregman, A. S. (1990). "Auditory scene analysis," Cambridge, Mass., MIT Press.

[Bro82] Brokx, J. P. L., and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," Journal of Phonetics 10, 23-36.

[Bro92] Brown, G. J. (1992), "Computational auditory scene analysis: a representational approach," Sheffield, Department of Computer Science unpublished doctoral dissertation.

[BrC93] Brown, G. J., and Cooke, M. P. (1992). "Computational auditory scene analysis: grouping sound sources using common pitch contours," Proc. Inst. of Acoust. 14, 439-446.

[Car94] 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.

[Che53] Cherry, E. C. (1953). "Some experiments on the recognition of speech with one, and with two ears," J. Acoust. Soc. Am. 25, 975-979.

[Coo91] Cooke, M. P. (1991), "Modeling auditory processing and organisation," Sheffield, Department of Computer Science unpublished doctoral dissertation.

[CoB93] Cooke, M. P., and Brown, G. J. (1993). "Computational auditory scene analysis: exploiting principles of perceived continuity," Speech Comm. 13, 391-399.

[CoH00 Cooke, M., and Ellis, D. P. W. (2000). "The auditory organization of speech and other sources in listeners and computational models," Speech Comm., in press.

[CoG00] Cooke, M., and Green, P. (2000). "Auditory organization and speech perception," in "Listening to speech: an auditory perspective," Edited by S. Greenberg and W. Ainsworth, Oxford, Oxford University Press, in press.

[CGJ00] Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2000). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Communication (submitted)

[CMG97] Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.

[CuD93] Culling, J. F., and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. 93, 3454-3467.

[CuD94] Culling, J. F., and Darwin, C. J. (1994). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. 95, 1559-1569.

[CuS95] Culling, J. F., and 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.

[Cut76] Cutting, J. E. (1976). "Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening," Psychol. Rev. 83, 114-140.

[Dar75] Darwin, C. J. (1975). "On the dynamic use of prosody in speech perception," in "Structure and process in speech perception," Edited by A. Cohen and S. G. Nooteboom.

[Dar81] Darwin, C. J. (1981). "Perceptual grouping of speech components differing in fundamental

frequency and onset-time," QJEP A 33, 185-207.

[Dar77] Darwin, C. J., and Bethel-Fox, C. E. (1977). "Pitch continuity and speech source attribution," J. Exp. Psychology: Human Perception and Performance 3, 665-672.

[DaC95] Darwin, C. J., and Carlyon, R. P. (1995). "Auditory grouping," in "Handbook of perception and cognition: Hearing," Edited by B. C. J. Moore, New York, Academic Press, 387-424.

[deC93] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

[deC94] de Cheveigné, A., Kawahara, H., Aikawa, K., and Lea, A. (1994). "Speech separation for speech recognition," Journal de Physique IV 4, C5-545-C5-548.

[deC95] de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. 97, 3736-3748.

[deC96] de Cheveigné, A., and Marin, C. (1996). "The segregation of frequency-modulated concurrent harmonic sounds," J. Acoust. Soc. Am. 100, 2718.

[deC97a] de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997). "Concurrent vowel identification I: Effects of relative level and F0 difference," J. Acoust. Soc. Am. 101, 2839-2847.

[deC97b] de Cheveigné, A., McAdams, S., and Marin, C. (1997b). "Concurrent vowel identification II: Effects of phase, harmonicity and task," J. Acoust.Soc. Am. 101, 2848-2856.

[deC97c] de Cheveigné, A. (1997). "Concurrent vowel identification III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857-2865.

[deC98] de Cheveigné, A. (1998). "Cancellation model of pitch perception," J. Acoust. Soc. Am. 103, 1261-1271.

[deC99a] de Cheveigné, A., and Kawahara, H. (1999). "Multiple period estimation and pitch perception model," Speech Communication 27, 175-185.

[deC99b] de Cheveigné, A. (1999). "Vowel-specific effects in concurrent vowel identification," J. Acoust. Soc. Am. 106, 327-340.

[deC99c] de Cheveigné, A., and Kawahara, H. (1999). "Missing data model of vowel perception," J. Acoust. Soc. Am. 105, 3497-3508.

[deC99d] de Cheveigné, A. (1999). "Waveform interactions and the segregation of concurrent vowels," J. Acoust. Soc. Am. 106, 2959-2972.

[deC99e] de Cheveigné, A. (1999). "Formant bandwidth affects the identification of competing vowels.", Proc. ICPhS, 2093-2096.

[deC99f] de Cheveigné, A. (1999). "Pitch shifts of mistuned partials: a time-domain model," J. Acoust. Soc. Am. 106, 887-897.

[deC00] de Cheveigné, A. (2000). "L'analyse de scènes auditives computationnelle," in "La parole, des modèles cognitifs aux machines communicantes - Développement," Edited by J. Mariani, Paris, Hermès, en préparation, .

[deV99] de Veth, J., Cranen, B., de Wet, F., and Boves, L. (1999). "Acoustic pre-processing for optimal effectivity of missing feature theory.", Proc. Eurospeech, 65-68.

[Dem90] Demany, L., and Semal, C. (1990). "The effect of vibrato on the recognition of masked vowels," Percept. & Psychophys. 48, 436-444.

[Ell96] Ellis, D. (1996), "Prediction-driven computational auditory scene analysis," MIT unpublished doctoral dissertation.

[Ell97] 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.

[Gre97] 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.

[Har96] Hartmann, W. M. (1996). "Pitch, periodicity, and auditory organization," J. Acoust. Soc. Am. 100, 3491-3502.

[Har91] Hartmann, W. M., and Johnson, D. (1991). "Stream segregation and peripheral channeling," Music Perception 9, 155-184.

[Hel77] Helmholtz, H. v. (1877). "On the sensations of tone (English translation A.J. Ellis, 1954)," New York, Dover.

[Her98] Hermansky, H. (1998). "Should recognizers have ears?," Speech Comm. 25, 3-27.

[Lea92] Lea, A. (1992), "Auditory models of vowel perception," Nottingham unpublished doctoral dissertation.

[Lic56] Licklider, J. C. R. (1959). "Three auditory theories," in "Psychology, a study of a science," Edited by S. Koch, New York, McGraw-Hill, I, 41-144.

[Lip97] Lippmann, R. P. (1997). "Speech recognition by machines and humans," Speech Comm. 22, 1-16.

[LiC97] Lippmann, R. P., and 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.

[Lyo94] Lyon, R. (1984). "Computational models of neural auditory processing," Proc. IEEE ICASSP, 36.1.(1-4).

[Lyo83] Lyon, R. F. (1983-1988). "A computational model of binaural localization and separation," in "Natural computation," Edited by W. Richards, Cambridge, Mass, MIT Press, 319-327.

[Mar91] Marin, C. (1991), "Processus de séparation perceptive des sources sonores simultanées," Paris III unpublished doctoral dissertation.

[MaM91] Marin, C., and McAdams, S. (1991). "Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width," J. Acoust. Soc. Am. 89, 341-351.

[Mar82] Marr, D. (1982). "Representing and computing visual information," in "Artificial Intelligence: an MIT perspective," Edited by P. H. Winston and R. H. Brown, Cambridge, Mass, MIT Press, 2, 17-82.

[Mas90] Massaro, D. W. (1990). "Models of integration given multiple sources of information," Psychological review 97, 225-252.

[McA84] McAdams, S. (1984), "Spectral fusion, spectral parsing, and the formation of auditory images," Stanford unpublished doctoral dissertation.

[McA89] McAdams, S. (1989). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am. 86, 2148-2159.

[MeH92] Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91, 233-245.

[Mel91] Mellinger, D. K. (1991), "Event formation and separation in musical sound," Stanford Center for computer research in music and acoustics unpublished doctoral dissertation.

[Moo96] Moore, R. (1996). "Critique: The potential role of speech production models in automatic speech recognition," J. Acoust. Soc. Am. 99, 1710-1713.

[NOK95] Nakatani, T., Okuno, H. G., and Kawabata, T. (1995). "Residue-driven architecture for computational auditory scene analysis.", Proc. IJCAI, 165-172.

[SpC99] numéro spécial Speech Communication v. 27 nos 3-4, 1999.

[Par76] Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, 911-918.

[Rem94] Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994). "On the perceptual organization of speech," Psychological Review 10, 129-156.

[RoO97] Rosenthal, D. F., and Okuno, H. G. (1997). "Computational auditory scene analysis," Lawrence Erlbaum.

[RoD98] Rozogan, A., and Deléglise, P. (1998). "Adaptive fusion of acoustic and visual sources for automatic speech recognition," Speech Comm. 26, 149-161.

[Sch83] Scheffers, M. T. M. (1983), "Sifting vowels," Gröningen unpublished doctoral dissertation.

[Sla95] Slaney, M. (1995). "A critique of pure audition.", Proc. Computational auditory scene analysis workshop, IJCAI, Montreal.

[Sum92] Summerfield, Q. (1992). "Roles of harmonicity and coherent frequency modulation in auditory grouping," in "The auditory processing of speech: from sounds to words," Edited by M. E. H. Schouten, Berlin, Mouton de Gruyter, 157-166.

[SuC92a] Summerfield, Q., and 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).

[SuC92b] Summerfield, Q., and 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.

[Wan95] Wang, A. L.-C. (1995), "Instantaneous and frequency-warped signal processing techniques for auditory source separation," unpublished doctoral dissertation, CCRMA (Stanford University).

[War70] Warren, R. M. (1970). "Perceptual restoration of missing speech sounds," Science 167, 392-393.

[Wei85] Weintraub, M. (1985), "A theory and computational model of auditory monaural sound separation," Stanford unpublished doctoral dissertation.

[YDS96] Yost, W. A., Dye, R. H., and Sheft, S. (1996). "A simulated "cocktail party" with up to three sound sources," Perception and Psychophysics 58, 1026-1036.

[ZiW83] Zissmann, M. A., and Weinstein, C. J. (1989). "Speech-state-adaptive simulation of co-channel talker interference suppression.", Proc. IEEE-ICASSP, 361-364.