Les mécanismes cognitifs de l'analyse séquentielle des scènes auditives

Nicolas Grimault, Aymeric Devergie
UMR CNRS 5292
Centre de Recherche en Neurosciences de Lyon
Équipe Cognition Auditive et Psychoacoustique
Université Lyon 1
50, avenue Tony Garnier
69366 Lyon CEDEX 07
E-mail : nicolas.grimault@olfac.univ_lyon1.fr

Résumé
Dans notre quotidien auditif, nous sommes généralement exposés à des situations concurrentielles d'écoute. Notre système auditif est ainsi rapidement confronté à la nécessité d'analyser la scène auditive afin d'organiser le paysage sonore en objets acoustiques auxquels il pourra consécutivement associer des attributs perceptifs et cognitifs. La situation quotidienne la plus évidente où il devient absolument impératif d'organiser notre espace sonore en objets auditifs est la situation de paroles concurrentes. En effet, pour accéder au message linguistique, il est nécessaire de parvenir à un état d'organisation suffisant de la scène auditive qui permette d'attribuer un flux de parole distinct et signifiant à chaque locuteur ou tout du moins à notre interlocuteur. Toute différence acoustique existant entre deux flux de parole est potentiellement utile à cette analyse. Cependant, l'analyse d'une scène auditive se fait généralement en présence d'un contexte multisensoriel. Ce contexte active des processus cognitifs mettant en œuvre les connaissances qui induisent des attentes perceptives. Nous proposons ici de discuter la contribution du contexte multisensoriel et des attentes perceptive pour l'analyse des scènes auditives.

L'analyse des scènes auditives (ASA)

Dans notre quotidien auditif, nous sommes généralement confrontés à des situations concurrentielles d'écoute. Notre système auditif est ainsi rapidement confronté à la nécessité d'analyser la scène auditive afin d'organiser le paysage sonore en objets acoustiques auxquels il pourra consécutivement associer des attributs perceptifs et cognitifs. Hiérarchiser les processus neuronaux est un exercice difficile et superfétatoire mais quelques éléments de réflexion permettent de suggérer la nature précoce et complexe de ces mécanismes. Ainsi, par exemple, lorsque plusieurs sources sonores concurrentes sont actives, calculer la sonie globale au moyen d'un modèle périphérique (principalement cochléaire) de sonie tels que ceux de [1] ou [2] semble peu pertinent. Il faudra au contraire calculer la sonie de chaque source avant de s'attacher à comprendre comment ces sonies peuvent interagir entre elles.

Pour se convaincre de cela, une petite expérience simple [3] consiste à mettre en évidence la capacité d'un auditeur à égaliser manuellement la sonie d'un son pur, soit avec la sonie d'un son complexe à deux tons, lorsqu'ils sont intégrés dans un même percept, soit avec l'une des composantes tonales de ce son complexe lorsque les deux composantes sont perçues séparément. Le simple fait que cette expérience soit réalisable montre que la sonie est bien une propriété perceptive complexe et cognitive associée par le système auditif à chaque objet sonore. Cela montre également que en contexte multi-sources, l'analyse de la scène auditive survient en amont du processus d'émergence de la sonie.

L'analyse de la scène auditive semble donc être un préalable à la plupart des situations naturelles d'écoute mais, si elle intervient dès les premiers niveaux de traitement, elle est également sous l'influence potentielle de traitements cognitifs de haut niveau faisant intervenir les connaissances (langagières ou musicales, par exemple) ou l'hédonicité des signaux perçus.

Étudier les mécanismes de l'analyse des scènes auditives peut donc permettre de tester certaines hypothèses sur l'existence de processus cognitifs descendants exerçant un contrôle de la périphérie auditive par les centres.

Les processus cognitifs et multisensoriels de l'analyse des scènes auditives

La perception en général et l'analyse des scènes auditives en particulier se fait en contexte. Il existe ainsi très probablement une sorte de bouclage où la perception évoque un contexte qui influence la mise en flux qui *a posteriori* vient alimenter la perception auditive. Par exemple la sonie contribuerait à la mise en flux auditif [5,6] puis, une fois cette mise en flux réalisée, la sonie deviendrait un attribut associé à chaque objet sonore [3].

La perception d'un contexte fait appel à des processus cognitifs mettant en œuvre les connaissances. Des attentes perceptives induites par ces connaissances vont ainsi venir moduler des processus attentionnels. Il semble raisonnable également de proposer que ces processus cognitifs, impliqués dans la perception des contextes, puissent être multisensoriels et en particulier audiovisuels.

Il est généralement admis que les processus cognitifs basés sur les connaissances sont impliqués dans l'analyse des scènes auditives. Ainsi, Bregman [6] propose que des connaissances (des schémas), stockées dans la mémoire à long terme puissent contribuer à l'extraction d'une cible dans une mixture. Pourtant, peu d'études se sont réellement penchées sur cette question. L'étude la plus pertinente alimentant cette hypothèse est celle de Dowling, et al. [7] qui ont montré que des sujets pouvaient extraire une mélodie familière intercalée avec une mélodie non familière en absence de différence de hauteur.
Hafter [8] propose de regrouper sous la terminologie d'attention perceptive les mécanismes qui permettent d'extraire d'un environnement complexe une information utile.
Mises en commun, ces propositions de Bregman et Hafter suggèrent donc que l'activation de connaissances puisse moduler des processus d'attention perceptive qui permettent d'analyser une scène auditive.

Les processus attentionnels de la ségrégation volontaire

L'étude de Devergie et al. [9] a permis à la fois d'apporter de nouvelles données en faveur de l'hypothèse des schémas de Bregman mais également d'étudier les processus attentionnels mis en œuvre. Dans cette étude, la tâche consistait à extraire une mélodie hautement familière d'une mélodie cible intercalée. Comme dans Dowling et al [7], aucune différence de timbre ou de hauteur ne permettait de réaliser la tâche. Par contre, contrairement à Dowling et al [7], la tâche était rendue plus difficile en introduisant une irrégularité de rythme non seulement dans la mélodie cible mais également dans la mélodie distractrice dans l'une des deux conditions expérimentales.

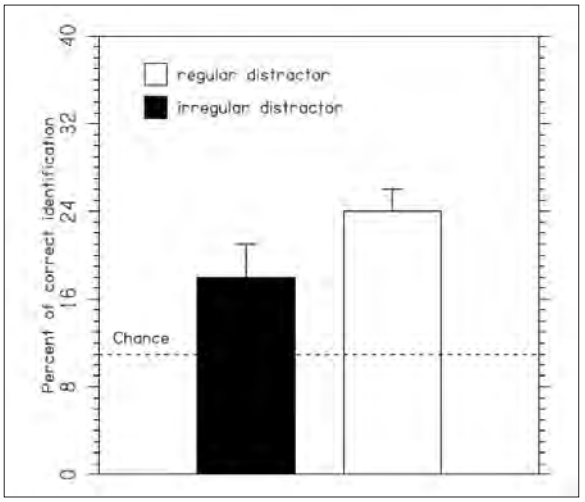

Fig. 1 : Performances d'identification d'une mélodie cible intercalée avec une mélodie distractrice. Dans la condition régulière, le tempo de la mélodie distractrice est régulier. Dans la condition irrégulière, ce tempo est irrégulier. Le tempo de la mélodie cible est toujours irrégulier dans les deux conditions. Figure reproduite de [9].

Les résultats présentés sur la figure 1 montrent que la tâche reste possible même en absence de régularité rythmique. Cela indique donc un fort effet des connaissances sur la ségrégation volontaire.

Plus intéressant, ces résultats montrent également que rendre le rythme de la mélodie distractrice isochrone renforce la ségrégation. Nous avons donc ainsi pu monter que l'attention rythmique (e.g. [10]) était probablement impliquées dans ces processus d'extraction d'une mélodie cible d'une mélodie distractrice.

Effet multimodal sur la ségrégation auditive

L'organisation perceptive de notre environnement en «objet» est par nature multimodale. Ainsi dans la plupart des situations, un objet auditif pourra être mis en correspondance avec un objet visuel, un objet olfactif...

Une mise en flux peut donc potentiellement s'effectuer dans chaque modalité et ces mises en flux ont également la possibilité d'interagir entre elles. De nombreuses études sur la multistabilité (mise en flux instable ou liage instable) se sont penchées sur la description de ces phénomènes dans les différentes modalités sensoriels et sur ces interactions (pour une revue, voir le numéro spécial Phil. Trans. B préfacé par Schwartz, et al. [11]).

Un cas particulier mais particulièrement pertinent dans notre quotidien concerne la parole. En effet, dans le cas de la parole, sans nécessairement postuler que Speech Is Special [12,13], le flux auditif est généralement en correspondance étroite avec un flux visuel consistant en un visage pourvu de lèvres animées. L'effet ventriloque [14,15] par exemple, peut être interprété comme une illusion perceptive où un flux auditif est attribué à tord à un flux visuel localisé en un autre lieu que la source sonore (par exemple une marionnette animée). L'effet Mc Gurck [16] consistant à biaiser la perception d'un /ba/ auditif vers un /da/ en présentant un /ga/ visuel (lèvres parlantes) est une autre illustration de ces interactions multimodales. Enfin, depuis les années 50, de nombreuses études ont montré que la lecture labiale pouvait améliorer jusqu'à 40% les performances d'intelligibilité dans le bruit (e.g. [17]). Il n'est toutefois pas clair si ce gain est dû à une intelligibilité renforcée *per se* ou à une facilitation de l'extraction de la voix cible du bruit de fond.

Supposant ainsi que l'intégration audiovisuelle dans le cas de lèvres parlantes devait être particulièrement robuste, nous avons ainsi souhaité tester l'hypothèse selon laquelle des mouvements de lèvres pouvaient venir moduler de façon irrépressible l'état de ségrégation d'un flux de voyelles auditives. Le paradigme expérimental de ségrégation de voyelles (cf. Figure 2) a donc été adapté afin d'ajouter une stimulation visuelle, comme des mouvements de lèvres, plus ou moins corrélés aux stimuli auditifs. Cette étude a permis d'apporter des premiers éléments pouvant laisser penser que des interactions audiovisuelles existent dès les premiers niveaux de traitement liés à la ségrégation auditive irrépressible. En effet, lorsqu'une stimulation visuelle congruente à l'un des deux flux auditifs était introduite, la ségrégation auditive irrépressible entre les deux flux auditifs était significativement augmentée.

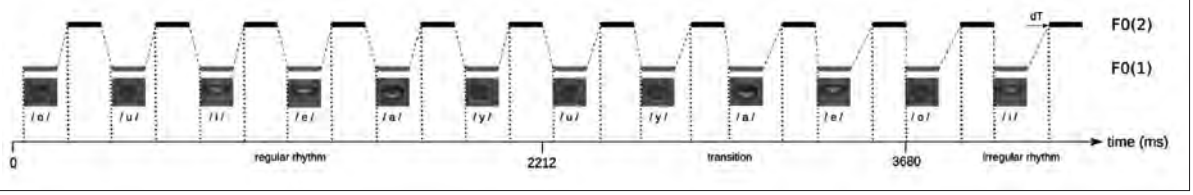

Fig. 2 : Séquence audiovisuelle schématique utilisée dans [18]. Chaque trait horizontal correspond à la présentation d'une voyelle. Certaines voyelles (traits noirs) sont purement audio et sont prononcées à la hauteur F0(2). Les autres (traits gris) sont audiovisuelles et sont prononcées à la hauteur F0(1). Un décalage rythmique entre les voyelles (dT) est progressivement introduit au cours de la séquence. La tâche du sujet consistait à détecter cette irrégularité rythmique.

Conclusions et perspectives

Ces deux études exploratoires [9,18] posent de très nombreuses questions à ce jour non résolues. Par exemple, les études de Dowling et al [7] et Devergie et al [9] confrontées aux résultats de Bey [19] semblent suggérer qu'il est nécessaire, pour permettre l'extraction d'une mélodie cible d'une mélodie distractrice, que cette mélodie cible soit fortement encodée dans la mémoire à long terme. En effet, un simple encodage dans la mémoire de travail semble insuffisant. Si ce résultat était confirmé par des études à venir, cela suggérerait un effet *topdown* impliquant des hauts niveaux de traitement. En général, étudier l'implication des mécanismes mnésiques pour la ségrégation auditive semble nécessaire.

De la même façon, l'étude audiovisuelle de Devergie et al [18] devrait être reproduite avec un panel de stimuli plus important proposant une gradation de la corrélation physique et temporelle entre les stimuli audio et visuels afin de confirmer l'hypothèse proposée et selon laquelle cette corrélation est responsable de l'effet audiovisuel observé en créant une forte association (ou un fort liage) entre l'objet visuel et l'objet audio. Par ailleurs, outre la cohérence physique entre les signaux audio et visuels, il serait intéressant de faire varier une cohérence contextuelle plus cognitive en introduisant par exemple des conflits de genre entre le signal audio et visuel [20] ou des conflits de type émotionnel [21]. Par exemple, cette étude pourrait être reproduite dans une condition où le visage exprimerait la peur alors même que le son exprimerait la sérénité.

Enfin, il me semblerait intéressant de mettre en commun les travaux de Devergie et al. [9,18] afin de tester l'hypothèse selon laquelle les cycles attentionnels puissent être renforcés par la présentation d'images synchrones. Pour cela, une idée serait de reproduire la tâche attentionnelle de mélodies intercalées de Devergie et al [9] en ajoutant un indice visuel de mouvement congruent afin d'observer un éventuel renforcement de la ségrégation auditive. Une telle étude pourrait permettre de mettre en évidence pour la première fois un *effet métronome* ou un *effet chef d'orchestre* sur la ségrégation auditive de séquences musicales.

Références bibliographiques

[1] Fastl H., Zwicker E. «Psychoacoustics: facts and models». Springer-Verlag New York Inc, Vol. 22, 2007.

[2] Moore B.C.J., Glasberg B.R., Baer T. «A model for the prediction of thresholds, loudness, and partial loudness». Journal of the Audio Engineering Society 45, 224-40, 1997.

[3] Grimault N., Mcadams S., Allen J.B. «Auditory Scene Analysis: A Prerequisite for Loudness Perception». In B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, et J. Verhey (Éd.), Hearing – From Sensory Processing to Perception (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 295-302, 2007.

[4] van Noorden L. «Temporal coherence in the perception of tone sequences». Unpublished PhD dissertation, 1975.

[5] Stainsby T.H., Moore B.C.J., Medland P.J., Glasberg B.R. «Sequential streaming and effective level differences due to phase-spectrum manipulations». J. Acoust. Soc. Am. 115, 1665-1673, 2004.

[6] Bregman A. S. «Auditory Scene Analysis: The Perceptual Organization of Sounds». (Cambridge MA), The MIT Press, 1990.

[7] Dowling W.J., Lung K.M., Herrbold S. Aiming attention in pitch and time in the perception of interleaved melodies. Percept Psychophys 41, pp. 642-656, 1987.

[8] Hafter E.R., Sarampalis A., Loui P. «Auditory attention and filters». Auditory perception of sound sources, pp. 115-142, 2007.

[9] Devergie A., Grimault N., Tillmann B., Berthommier F. «Effect of rythmic attention on the segregation of interleaved melodies». J. Acoust. Soc. Am. 128, EL1-7, 2010.

[10] Jones M.R. «Time, our lost dimension: Toward a new theory of perception, attention, and memory». Psychological review 83, 323, 1976.

[11] Schwartz J.-L., Grimault N., Hupe J.-M., Moore B.C.J., Pressnitzer D. «Multistability in perception: binding sensory modalities, an overview. Philos. Trans. R. Soc. B-Biol. Sci. 367, pp. 896-905, 2012.

[12] Remez R.E., Rubin P.E., Berns S.M., Pardo J.S., Lang J.M. «On the perceptual organization of speech». Psychological review 101, 129, 1994.

[13] Liberman A.M., Mattingly I.G. «The motor theory of speech perception revised». Cognition 21, pp. 1-36, 1985.

[14] Radeau M., Bertelson P. Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Perception & Psychophysics 22, pp. 137-146, 1977.

[15] Bertelson P., Aschersleben G. «Automatic visual bias of perceived auditory location». Psychonomic Bulletin & Review 5, pp. 482-489, 1998.

[16] McGurk H., MacDonald J. «Hearing lips and seeing voices». Nature 264, pp. 746-748, 1976.

[17] Sumby W.H. «Visual Contribution to Speech Intelligibility in Noise». J. Acoust. Soc. Am. 26, 212, 1954.

[18] Devergie A., Grimault N., Gaudrain E., Healy E.W., Berthommier F. «The effect of lipreading on primary stream segregation». J. Acoust. Soc. Am. 130, pp. 283-291, 2011.

[19] Bey C. «Recognition of interleaved melodies and formation of auditory streams: Functional analysis and neuropsychological exploration». Unpublished, PhD dissertation, 1999.

[20] Green K.P., Kuhl P.K., Meltzoff A.N., Stevens E.B. «Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. Percept Psychophys 50, pp. 524-536, 1991.

[21] de Gelder B. «Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures». Proceedings of the National Academy of Sciences 99, pp. 4121-4126, 2002.