La reconnaissance du timbre des sons

Daniel Pressnitzer, Trevor R. Agus
CNRS & Département d'études cognitives
Ecole normale supérieure
29 rue d'Ulm
75230 Paris CEDEX 05
E-mail : Daniel.Pressnitzer@ens.fr et trevor.agus@ens.fr

Clara Suied
Département Action et Cognition en Situation Opérationnelle,
Institut de recherche biomédicale des armées
91220 Brétigny sur Orge
E-mail : clarasuied@ens.fr

Résumé
La reconnaissance du timbre des sons semble fondamentale pour l'audition humaine : cette capacité nous permet entre autres de reconnaître une voix parmi un brouhaha, ou un instrument de musique au sein de l'orchestre.
Pourtant, il existe encore de nombreuses questions à résoudre pour mieux comprendre l'impressionnante efficacité des auditeurs pour reconnaître les timbres, alors que les systèmes artificiels actuels semblent encore largement perfectibles. Le but de cet article est de donner un bref tour d'horizon des recherches sur le sujet, allant de la perception aux neurosciences. Nous proposons une articulation des controverses autour de deux approches possibles : l'une recherchant un petit nombre de dimensions acoustiques génériques permettant de caractériser le timbre, l'autre faisant l'hypothèse de multiples traits spécifiques sous-jacents à l'identification d'une source.

Le timbre est ce qui permet à un auditeur de distinguer deux sons qui seraient perçus par ailleurs comme possédant les mêmes hauteur, intensité, position spatiale, et durée. Par exemple, lorsque les musiciens d'un orchestre s'accordent avant un concert, ils jouent tous la même note, et pourtant nous pouvons entendre une différence entre les divers instruments. Ceci est possible en grande partie grâce au timbre.

La définition classique du timbre que nous venons de donner souffre de plusieurs limitations. Tout d'abord, elle spécifie ce que le timbre n'est pas, mais sans dire pour autant ce qu'il est. Ensuite, elle se réfère à la comparaison entre deux réalisations sonores particulières, alors qu'une fonction qui semble plus utile pour l'audition est de pouvoir associer un timbre à une source sonore - et donc à tous les sons que cette source peut produire (le timbre du piano, ou le timbre particulier de la voix d'un ami). Peut-être à cause de cette définition imparfaite, il existe encore un débat animé sur les indices acoustiques, les représentations mentales, et les mécanismes neuronaux qui pourraient sous-tendre la perception du timbre.

Nous allons donner un bref aperçu de controverses actuelles, en se focalisant sur le problème pratique de la reconnaissance d'une source sonore grâce au timbre.

Nous esquisserons d'abord les principes élémentaires qui font du timbre un indice potentiel très puissant pour la reconnaissance de source. Nous évoquerons l'hypothèse qu'il existe deux approches possibles et clairement distinctes pour l'étude expérimentale du timbre. Ensuite, nous suivrons ces deux approches dans les domaines de l'acoustique, de la perception, de l'étude des mécanismes neuronaux, et des applications pratiques.

Pourquoi différentes sources sonores produisent-elles différents timbres ?

Les sources sonores sont des objets physiques qui peuvent prendre toutes les formes possibles et imaginables. Un son est produit quand un objet est mis en vibration par un apport d'énergie. Des vibrations s'établissent alors dans l'objet, se propagent dans l'air environnant, et atteignent les oreilles de l'auditeur sous la forme d'une onde de pression (Figure 1). La physique élémentaire montre que l'onde qui arrive à l'oreille peut contenir de précieuses informations sur ce qui s'est passé du côté de la source [1]. Par exemple, si l'apport d'énergie était bref, comme pour un «toc» sur une porte, il y a toutes les chances que l'onde de pression elle-même soit brève avec la plus grande partie de son énergie concentrée autour du moment du toc. Après le toc, la façon dont la porte va continuer à vibrer est étroitement liée à sa géométrie : certains modes de vibration sont compatibles avec certaines géométries, mais pas avec d'autres. Une de ces règles de base est que les basses fréquences et donc les grandes longueurs d'ondes ne sont pas stables pour des objets de petites dimensions. Donc, la proportion des différentes composantes fréquentielles du son sera contrainte par la géométrie de la porte. D'autres règles plus complexes s'appliquent, prenant en compte la forme précise de l'objet, la nature des matériaux impliqués, et ainsi de suite.

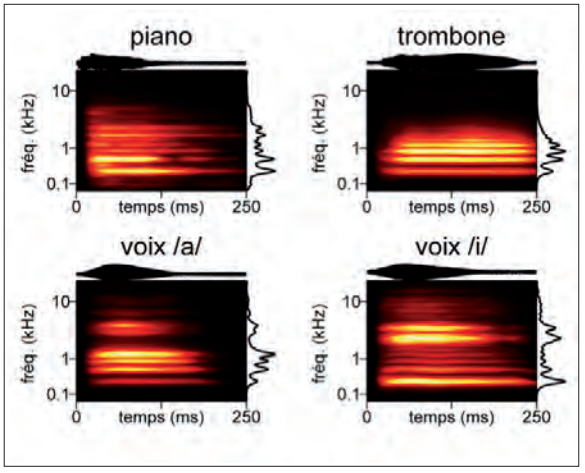

Fig. 1 : Représentations visuelles de quatre sons qui ont la même durée, intensité subjective, et hauteur : ils diffèrent donc par leur timbre. Chaque panneau représente une analyse temps-fréquence dérivée d'un modèle auditif (voir [8] pour les détails). En bref, la couleur indique la distribution d'énergie à l'intérieur de canaux fréquentiels (axe des y) et son évolution au cours du temps (axe des x). Le tracé au-dessus des panneaux est l'onde temporelle correspondante. Le tracé à droite est la moyenne de l'énergie au cours du temps, similaire (mais non identique) à une densité spectrale de puissance. Les deux instruments de musique ont été choisis pour illustrer les dimensions classiques du timbre : en fonction de la source et de son mode d'excitation, le temps d'attaque peut être rapide (piano) ou lent (trombone) ; le centre de gravité du spectre peut être haut et le timbre « brillant » (piano) ou bas et le timbre peu «brillant» (trombone). Les deux voyelles illustrent l'idée qu'il peut exister d'autres traits, potentiellement plus complexes, qui distinguent par exemple les voyelles des instruments, ou même les voyelles entre elles.

Être capable de décoder le lien complexe entre patron de vibrations et source sonore est extrêmement utile. Cela permet à l'audition d'être un sens d'alerte, permettant par exemple d'identifier des objets produisant du son avant même qu'ils ne rentrent dans le champ de vision. De plus, le timbre est aussi la base du langage parlé : les voyelles et consonnes sont produites par des modulations de la forme de l'appareil vocal, qui se traduisent par des changements de timbre.

«Dimensions» versus «Traits»

Il n'existe pas de consensus sur ce qui rend possible la reconnaissance de timbre chez les auditeurs humains. Pour esquisser les controverses actuelles, il semble utile de considérer deux points de vue opposés (Figure 2). Un premier point de vue est que le timbre est composé d'un nombre relativement faible de dimensions perceptives, qui sont chacune des descriptions subjectives du son au même titre que les dimensions classiques comme grave/aigü (hauteur) ou fort/faible (intensité subjective). Ces dimensions doivent être métamériques, au sens que de nombreux sons avec des timbres très différents doivent pouvoir se projeter sur un même point pour une dimension donnée.
Un deuxième point de vue est que la reconnaissance du timbre se base sur des traits distinctifs, spécifiques à une source sonore.

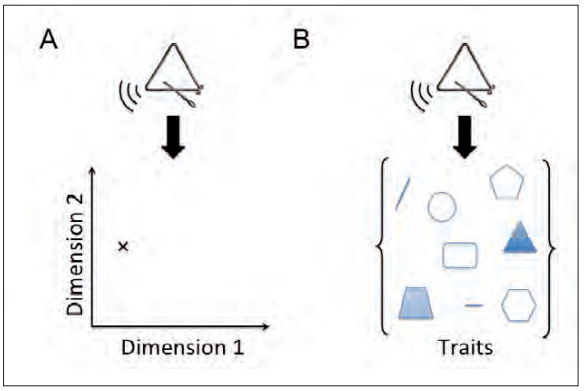

Fig. 2 : Illustration schématique des approches «Dimension» versus «Traits» pour la reconnaissance du timbre. A) Pour l'approche Dimension tous les timbres possibles peuvent être projeté dans un espace de faible dimensionnalité, le long de dimensions continues. Reconnaître un timbre serait alors reconnaître sa position dans l'espace. B) Pour l'approche Traits, chaque timbre est défini par un ensemble de traits distinctifs, choisis parmi un ensemble très vaste et non ordonné de traits complexes. Reconnaître un timbre est alors identifier une collection particulière de (peu) de traits.

Ces traits sont appris par l'expérience et sont choisis au sein d'un espace de traits potentiels qui peut être extrêmement vaste. Par exemple, le grain tout particulier de la voix d'un ami est peut-être unique, et c'est ce qui nous permet de le reconnaître instantanément.

Les traits ainsi définis sont conceptuellement très différents des dimensions évoquées au paragraphe précédent : un trait ne s'applique pas nécessairement à toutes les sources sonores ; en fait, c'est précisément parce le trait est spécifique à peu de sources (ou même à une seule source) qu'il pourrait être une base efficace pour la reconnaissance.

Il est fort probable qu'une compréhension globale du timbre se trouvera entre ces deux points de vue, grandement simplifiés pour l'exposé. Néanmoins, nous allons maintenant contraster ces deux approches dans divers domaines de la recherche sur le timbre.

Les représentations du timbre

Pour étudier le timbre, entre autres, il est utile de représenter visuellement les sons. De façon tout à fait classique, ceci est fait avec des outils comme la représentation temporelle de l'onde de pression ; l'analyse spectrale des composantes fréquentielles de cette onde (avec par exemple la transformée de Fourier) ; ou des transformations temps-fréquence comme la transformée de Fourier à court terme ou les ondelettes. Plus récemment, des modèles computationnels visant à simuler les traitements auditifs périphériques ou centraux ont aussi été proposés pour obtenir de telles représentations (e.g. [2]).

Dans l'approche «Dimensions», des statistiques descriptives sont calculées sur les représentations pour pouvoir réduire un son à quelques valeurs. C'est ce que l'on appelle des «descripteurs» du timbre.

Par exemple, l'évaluation du centre de gravité des composantes fréquentielles d'un son produit une valeur numérique qui est corrélée avec la «brillance» perçue du timbre [3]. Dans l'approche «Traits», la tendance sera plutôt de maximiser la richesse des représentations utilisées. Ceci peut se faire en incluant par exemples des transformées spectro-temporelles complexes [2]. Les représentations basées sur des traits ne sont pas nécessairement bien ordonnées.
Elles peuvent être largement redondantes, avec plusieurs milliers de traits différents mais non totalement indépendants les uns des autres. Au contraire, des représentations éparses peuvent aussi être utilisées [4], au sens qu'un son donné ne provoquerait l'activation que d'un très faible nombre de traits parmi un très grand choix possible [5].

Données perceptives

L'objectif affiché de l'approche «Dimensions» est de découvrir le nombre et la nature des dimensions de la perception du timbre. Pour ce faire, des techniques basées sur l'analyse multidimensionnelle sont utilisées : une paire de sons est présentée à l'auditeur, qui doit évaluer à quel point les sons lui semblent similaires ou non.
Cette évaluation est répétée pour toutes les paires possibles dans un ensemble donné de sons. Ensuite, les jugements de similarité sont traités comme des distances perceptives, et utilisés pour dériver la dimensionnalité et la géométrie de la représentation mentale sous-jacentes. Pour les instruments de musique, les études classiques suggèrent l'existence de deux à trois dimensions principales : l'une reliée au temps de l'attaque du son, une autre reliée au centre de gravité du spectre, et une dernière dont la nature varie selon les études [6,3]. Des données plus récentes, combinant l'analyse multidimensionnelle et des descriptions verbales, suggèrent cinq dimensions avec des interprétations plus complexes [7].

Dans l'approche «Traits», l'emphase n'est pas sur la perception de similarité entre sons mais plutôt sur l'identification d'une source bien particulière. De nouveau en utilisant des sons d'instruments de musique, nous avons par exemple mesuré le temps de réaction nécessaire à un auditeur pour reconnaître une source. Une grande variété de sons était présentée, avec des hauteurs qui variaient aléatoirement à chaque essai. Pourtant, un temps de reconnaissance remarquablement court a été observé [8]. Nous avons ensuite sévèrement appauvris les sons en les raccourcissant, jusqu'à des durées de quelques millisecondes seulement – la reconnaissance était largement préservée même dans les cas de dégradation les plus extrêmes [9]. Enfin, en utilisant un algorithme de traitement de signal permettant de dégrader les sons dans le domaine temps-fréquence, nous avons montré que la reconnaissance de stimuli dits émotionnels (rires, pleurs, etc.) était possible avec un très faible nombre de «traits» [10].

Un aspect commun qui est ressorti de ces études est que la reconnaissance est plus rapide et plus robuste pour des sources sonores familières, comme la voix humaine. Cette observation ne peut pas être expliquée en termes de dimensions acoustiques simples [8].

Ceci peut par contre être interprété comme suggérant l'existence de traits spécifiques, appris par les auditeurs à travers leur expérience avec les sources familières, qui permettent de les reconnaître de façon efficace et robuste.

Bases neuronales

Les corrélats neuronaux des dimensions génériques du timbre sont étudiés depuis quelques années, avec par exemple l'imagerie cérébrale fonctionnelle chez l'Homme. Avec un paradigme d'électroencéphalographie (EEG) permettant de sonder la mémoire sensorielle, il a été par exemple trouvé que des dimensions du timbre comme la brillance ou le temps d'attaque pouvaient chacune être représentée de façon séparée dans le cortex auditif [11].

Les enregistrements de l'activité de neurones isolés chez l'animal ont démontré une riche variété de réponses, souvent sans principe directeur évident (à part l'organisation par bandes de fréquence). Avec des techniques d'analyse des systèmes linéaires comme la corrélation inverse, des neurones spécialement sensibles à des propriétés spectro-temporelles ont été identifiés [12]. Si on rajoute une composante non-linéaire à l'analyse, un nouveau niveau de complexité apparaît [13]. On retrouverait donc là la richesse (et le désordre apparent) du type de représentation requis par l'approche «Traits».

Une question supplémentaire pour les neurosciences est de savoir si l'identité d'une source sonore est encodée par l'activité d'un vaste réseau cérébral, activé par de nombreuses sources différentes, ou alors par l'activité d'un petit réseau spécialisé pour un type de source donnée. Il existe des preuves expérimentales pour les deux hypothèses. En utilisant l'imagerie par résonnance magnétique fonctionnelle (IRMf), il est possible d'inférer l'identité d'une source à partir de l'activité distribuée dans de nombreuses régions cérébrales [14]. Mais il existe aussi des indications claires sur l'existence de régions spécialisées dans le traitement des sources familières comme la voix [15]. De façon surprenante, il a aussi récemment été démontré que les accordeurs de piano avaient de meilleures performances qu'un groupe contrôle pour juger d'un aspect bien particulier du timbre (les battements), et que suite à leur long apprentissage cette compétence spécifique se reflétait dans l'anatomie même de leur cerveau [16].

La reconnaissance du timbre par les machines

Il y a de nombreuses applications pratiques pour la reconnaissance du timbre : par exemple l'identification automatique des locuteurs, ou l'analyse de signaux musicaux (*music information retrieval*). Même si les techniques d'ingénierie développées changent rapidement et qu'il n'est pas possible de les décrire en quelconque détail dans ce paragraphe, il est intéressant de noter que le contraste «Dimension vs Traits» peut se retrouver aussi dans les architectures des systèmes computationnels utilisés.

La reconnaissance automatique de la parole peut être vue, jusqu'à un certain point, comme un exercice de décodage de timbre. Ce domaine a une longue tradition pour associer des techniques performantes de classification avec un petit nombre de coefficients génériques (par exemple les coefficients cepstraux et leurs variantes dans [17]).

Pour les instruments de musique, l'approche basée sur des descripteurs est directement inspirée des dimensions perceptives de l'analyse multidimensionnelle, avec un nombre relativement faible de descripteurs dont l'interprétation est explicite (18]. Néanmoins, il existe aussi d'autres systèmes qui sont basés sur la génération de traits au sein d'un énorme dictionnaire de traits potentiels, suivie par la sélection *ad hoc* des traits utiles pour une tâche de classification donnée [19,20]. Le codage parcimonieux est aussi utilisé dans cette perspective [4].

Pour la classification automatique d'instruments de musique, nous avons de plus proposé d'appliquer des algorithmes d'apprentissage-machine à des modèles auditifs de grande dimensionnalité [2].

Enfin, il peut être intéressant de remarquer que dans un domaine d'application distinct, la célèbre application «Shazam» permet l'identification d'un morceau de musique grâce à un ensemble de repères (positions relatives de pics dans une représentation temps-fréquences) qui servent à générer la signature unique d'un morceau donné. Ces repères peuvent être mis en parallèle des traits que nous avons évoqués jusqu'ici, à l'échelle d'un morceau de musique en entier.

Perspectives

Les questions en suspens concernant le timbre restent nombreuses, mais un aspect en particulier semble utile à considérer : les diverses stratégies disponibles pour l'auditeur lorsque l'on manipule le timbre. Par exemple, si un expérimentateur demande un jugement de similarité entre deux sons, alors une chose raisonnable à faire pour l'auditeur consiste à abstraire les dimensions communes à tous les sons à juger, puis à se focaliser sur ces dimensions. À l'inverse, si l'auditeur doit reconnaître une source aussi rapidement que possible, la simple détection d'un trait diagnostic (parmi d'autres) peut suffire. Il semble aussi intuitif que les dimensions ou traits utiles pourraient dépendre fortement de la tâche à réaliser : par exemple, pour un même ensemble de mots prononcés par divers locuteurs, différentes régions cérébrales sont recrutées si les auditeurs doivent identifier le locuteur ou s'ils doivent retranscrire le mot [21].

Enfin, les représentations neuronales elles-mêmes pourraient s'ajuster de façon dynamique au contexte acoustique, grâce à ce qui a été appelée la plasticité rapide [22]. Ceci suggère une raison supplémentaire expliquant pourquoi le timbre semble si difficile à capturer avec nos méthodes expérimentales actuelles : la reconnaissance de timbre pourrait être, en réalité, un mécanisme profondément adaptatif et changeant, capable de créer diverses stratégies efficaces en fonction des sons considérés et de la tâche à réaliser.

Note

Cet article est en partie adapté de l'entrée «*Acoustic Timbre Recognition*», par les mêmes auteurs, à paraître en 2013 dans Jaeger D., Jung R. (Ed.) Encyclopedia of Computational Neuroscience: SpringerReference (www.springerreference.com). Springer-Verlag Berlin Heidelberg, 2013.

Références bibliographiques

[1] Helmholtz, H. «On the sensations of tone» (Dover, New York), 1877.

[2] Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. «Music in our ears: the biological bases of musical timbre perception,» PLoS computational biology 8, e1002759, 2012.

[3] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. «Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes,» Psychological research 58, pp. 177-192, 1995.

[4] Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R., and Davies, M. E. «Sparse Representations in Audio and Music: From Coding to Source Separation,» P Ieee 98, pp. 995-1005, 2010.

[5] Hromadka, T., and Zador, A. M. «Representations in auditory cortex,» Current opinion in Neurobiology 19, pp. 430-433, 2009.

[6] Grey, J. M. «Multidimensional perceptual scaling of musical timbres,» J. Acoust Soc Am 61, pp. 1270-1277, 1977.

[7] Elliott, T. M., Hamilton, L. S., and Theunissen, F. E. «Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones,» J. Acoust Soc Am 133, pp. 389-404, 2013.

[8] Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. «Fast recognition of musical sounds based on timbre,» J Acoust Soc Am 131, pp. 4124-4133, 2012.

[9] Suied, C., Agus, T. R., Thorpe, S., and Pressnitzer, D. «Processing of short auditory stimuli: The Rapid Audio Sequential Presentation paradigm (RASP).» in Basic Aspects of Hearing: Physiology and Perception, edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel (Springer, New York), pp. 443-452, 2013.

[10] Suied, C., Drémeau, A., Pressnitzer, D., and Daudet, L. «Auditory sketches: Sparse représentations of sounds based on perceptual models,» Lecture Notes in Computer Science, In press, 2013.

[11] Caclin, A., Brattico, E., Tervaniemi, M., Naatanen, R., Morlet, D., Giard, M. H., and McAdams, S. «Separate neural processing of timbre dimensions in auditory sensory memory,» Journal of cognitive neuroscience 18, pp. 1959-1972, 2006.

[12] Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. «Spectrotemporal response field characterization with dynamic ripples in ferret primary auditory cortex,» Journal of neurophysiology 85, pp. 1220-1234, 2001.

[13] Machens, C. K., Wehr, M. S., and Zador, A. M. «Linearity of cortical receptive fields measured with natural sounds,» The Journal of neuroscience : the official journal of the Society for Neuroscience 24, pp. 1089-1100, 2004.

[14] Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. «Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex,» Current Biology 19, pp. 498-502, 2009.

[15] Belin, P. «Voice processing in human and non-human primates,» Philosophical transactions of the Royal Society of London. Series B, Biological sciences 361, pp. 2091-2107, 2006.

[16] Teki, S., Kumar, S., von Kriegstein, K., Stewart, L., Lyness, C. R., Moore, B. C., Capleton, B., and Griffiths, T. D. «Navigating the auditory scene: an expert role for the hippocampus,» The Journal of neuroscience : the official journal of the Society for Neuroscience 32, pp. 12251-12257, 2012.

[17] Hermansky, H. «Perceptual linear predictive (PLP) analysis of speech,» J. Acoust Soc Am 87, pp. 1738-1752, 1990.

[18] Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. «The Timbre Toolbox: extracting audio descriptors from musical signals,» J Acoust Soc Am 130, pp. 2902-2916, 2011.

[19] Coath, M., and Denham, S. L.. «Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience,» Biological cybernetics 93, pp. 22-30, 2005.

[20] Pachet, F., and Roy, P. «Analytical features: a knowledge-based approach to audio feature generation,» EURASIP Journal on Audio, Speech, and Music, Processing 2009.

[21] Formisano, E., De Martino, F., Bonte, M., and Goebel, R. «»Who» is saying «what»? Brain-based decoding of human voice and speech,» Science 322, pp. 970-973, 2008.

[22] Fritz, J., Shamma, S., Elhilali, M., and Klein, D. «Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex,» Nature neuroscience 6, pp. 1216-1223, 2003.