# Le son dans les applications interactives

Stéphane Conversy et Michel Beaudouin-Lafon

Laboratoire de Recherche en Informatique
U.R.A. 410 du C.N.R.S.
Bâtiment 490 - Université de Paris-Sud
91 405 ORSAY Cedex

novembre 1995

## Introduction

Dans la vie de tous les jours, nous utilisons tous nos sens pour acquérir des informations sur le monde qui nous entoure. En particulier, le système visuel nous donne des données détaillées sur une partie de ce monde, et le système auditif nous fournit des données générales qui ne peuvent être perçues par la vue. La combinaison de ces deux sens nous donne la plupart des informations importantes sur notre environnement.

Il est donc intéressant d'adapter ces deux vecteurs d'informations que sont la vue et l'ouïe à l'interaction homme-machine. Ainsi, comme dans un environnement naturel, l'utilisation de ces deux sens permet d'avoir un supplément d'information qu'un seul médium ne peut nous fournir : pendant que nous concentrons notre attention visuelle, pour éditer un document par exemple, nous pouvons surveiller l'état d'autres processus en tâche de fond (arrivée de courrier électronique, état d'une impression, etc). Cependant, le développement des interfaces homme-machine a privilégié l'aspect graphique au détriment de l'aspect sonore. Couramment, presque toutes les informations sont affichées à l'écran, conduisant à une surcharge de l'écran. La présence des deux média peut nous aider à présenter les informations de la façon la plus naturelle et la plus efficace possible.

## Le son dans l'informatique

L'utilisation de sons sur les ordinateurs n'est pas un fait nouveau, bien qu'il soit parfois fortuit. Un exemple courant est celui du bruit du disque dur ou de l'imprimante. Les utilisateurs peuvent dire quand une opération de sauvegarde ou de copie a pris fin, grâce au bruit que le disque dur produit. Cela leur permet de faire autre chose pendant que l'opération se déroule. Certes, le son doit être conçu de manière plus « formelle » que cet exemple, mais il faut retenir qu'il est un excellent médium.

L'évolution de l'informatique permet aujourd'hui l'utilisation de sons dans les systèmes informatiques. En effet, la plupart des postes de travail récents comportent des circuits sonores spécialisés, comme les stations NeXT, les Apple Macintosh AV, les Silicon Graphics Indigo (SGI) ou même les PC. De plus, l'avènement du standard MIDI dans le monde de la musique électronique et des synthétiseurs permet un contrôle toujours plus facile de ces machines. Malheureusement, ces capacités ne sont utilisées essentiellement que par des musiciens professionnels ou dans les jeux vidéos.
Nous nous intéressons ici uniquement à la communication sonore dans le sens ordinateur –› utilisateur (on exclut ainsi la reconnaissance de la parole ou de motifs sonores), et uniquement aux sons sont **non parlés** (pas de synthèse vocale), ceux-ci étant traités dans un autre chapitre.

On peut diviser le monde des sons non-parlés en trois groupes :

- *les alarmes et les avertissements* sont des signaux qui ont priorité sur toute autre information. Ils ont pour but d'interrompre les tâches courantes et d'alerter un utilisateur que quelque chose requiert son attention immédiate.
- *les messages d'état et de contrôle* donnent des informations sur une tâche qui se déroule.
- *les messages codés* servent à présenter des données numériques sous forme de motifs sonores. Ces sons complexes et variables tranchent par rapport aux sons simples des alarmes ou ceux plus répétitifs des sons de contrôles.

Les deux prochaines sections sont une introduction aux concepts et aux termes se rapportant à l'étude du son. Nous expliquons ce qu'est un son, les différentes manières de le représenter, ses caractéristiques et la manière dont nous le percevons. Cette partie peut être sautée par les lecteurs familiers du domaine. Ensuite nous présentons, dans les deux dernières sections, le traitement informatique des sons non parlés : les différents types de synthèses sonores y sont présentées, puis divers systèmes de sonifications.

## Acoustique

### Représentation du son

Les sons sont des variations de pression qui se propagent dans l'air. Une des manières les plus répandues de représenter le son est de représenter l'amplitude en fonction du temps. On appelle cette représentation une forme d'onde.

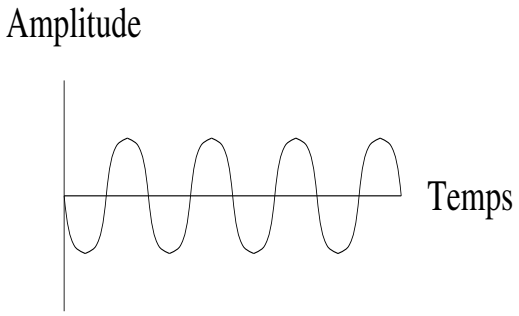



Fig 1 - *Une forme d'onde sinusoïdale. Ce type de graphe représente le son dans le domaine temporel.*

Le son montré en Figure 1 est une sinusoïde. Ce type de son est rarement rencontré dans la nature. Néanmoins, cette onde a des propriétés mathématiques extrêmement pratiques. En effet, grâce aux travaux de Fourrier [DLH88], on montre que les sons complexes, sous certaines conditions (il faut que leur forme d'onde soit périodique), peuvent être exprimés comme la somme d'un certain nombre de sinusoïdes d'amplitudes, de fréquences et de phases différentes. Grâce à la transformée de Fourrier, on peut retrouver ces sinusoïdes avec leurs caractéristiques respectives à partir d'une forme d'onde. On peut donc aussi représenter un son grâce à son spectre (amplitude en fonction de la fréquence). Ainsi, la forme d'onde de la Figure 1 peut être représentée par le spectre de la Figure 2.

Fig 2 - *Le spectre d'une sinusoïde. Ce type de graphe représente le son dans le domaine fréquentiel.*

Il est important de noter que la transformée de Fourrier permet de passer du domaine temporel au domaine fréquentiel et vice-versa (grâce à la transformée de Fourrier inverse). De plus, il n'y a pas de perte d'information lors de ces deux opérations. On peut remarquer que la fréquence est en quelque sorte l' « inverse » du temps : une forme d'onde est une fonction du temps (en *s*) et le spectre est une fonction des fréquences (en Hz ou $s^{-1}$. D'une manière plus générale, la transformée de Fourrier permet de passer d'une représentation en fonction d'une variable à une représentation en fonction de l'inverse de cette variable.

Avec le théorème de Fourrier et la transformée de Fourrier, on peut représenter une forme d'onde complexe de façon beaucoup plus explicite (cf. Figure 3), car on peut voir effectivement la décomposition du son en sinusoïdes.

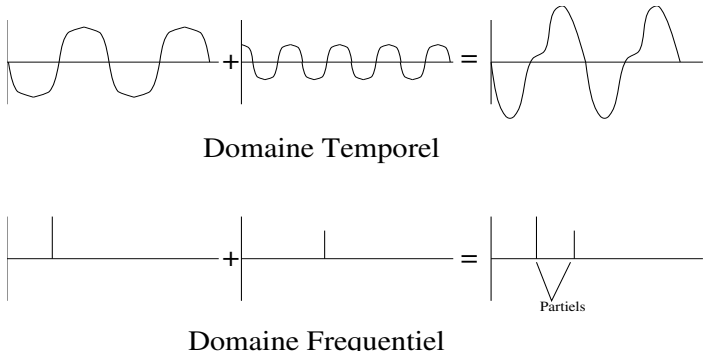



Fig 3 - *Un forme d'onde complexe. Le domaine des fréquences est plus facile à manipuler.*

Les composantes fréquentielles d'un son sont appelées des*partiels*. Très souvent, il n'existe aucune relation entre ces partiels, ce qui produit des sons *inharmoniques*. Cependant, certains sons ont des partiels qui sont des multiples entiers du partiel de plus basse fréquence (la fréquence *fondamentale*) : ces sons sont dits *harmoniques*. Les instruments traditionnels sont généralement conçus de manière à ce qu'il produisent ce type de sons, et c'est ce qui les rend agréables à écouter.

Si l'on regarde à nouveau la Figure 3, on remarque que les deux composantes sinusoïdales n'ont ni la même amplitude, ni la même phase. Si l'on changeait la phase de la deuxième composante par exemple, la forme d'onde ne serait plus du tout la même. Pourtant, le son entendu serait identique au précédent, c'est-à-dire que la perception d'un son ne dépend pas des phases de ses composantes. Le spectre (amplitude = f(fréquence)) est donc plus important que la forme d'onde.

## Enveloppe d'amplitude

Cependant, cette représentation ne tient pas compte de l'évolution du son au cours du temps. En effet, les sons naturels ne « stagnent » pas, ne serait-ce que parce qu'ils ont un début et une fin. Un moyen de caractériser l'évolution de l'amplitude au cours du temps est de tracer l'*enveloppe* d'amplitude, c'est-à-dire le contour de la forme d'onde (cf. Figure 4). La représentation la plus répandue est celle dite ADSR (Attack, Decay, Sustain, Release) ; cependant, on n'est pas obligé de se limiter à quatre segments, ni même à des segments rectilignes.

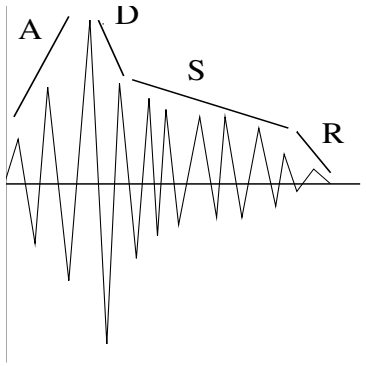



Fig 4 - *Une enveloppe ADSR*

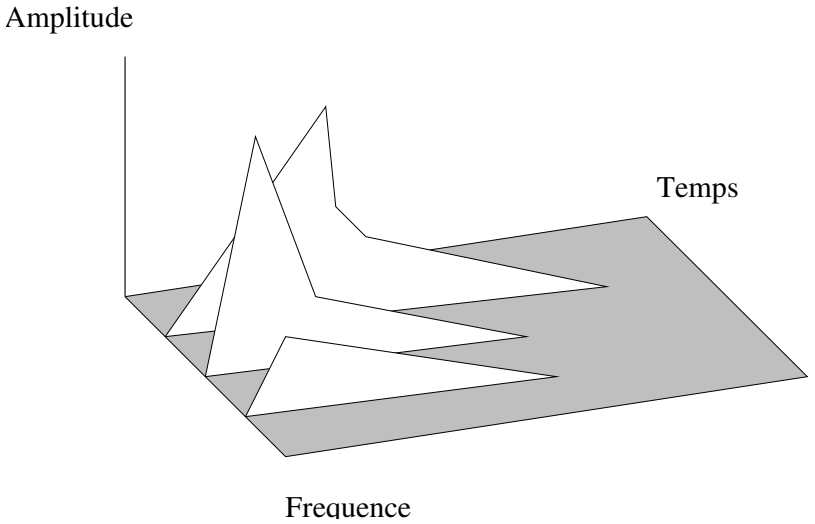



Fig 5 - *Un spectre variable au cours du temps*

Ce type de représentation peut être adapté au spectre, en définissant l'enveloppe de chaque partiel (Figure 5). Les évolutions relatives de l'amplitude de chaque partiel permettent de caractériser la « couleur » d'un son, autrement dit son *timbre*.

## Psycho-Acoustique

La psycho-acoustique est l'étude des phénomènes régissant la perception des caractéristiques d'un son par l'homme. La prise en compte des effets psycho-acoustiques est nécessaire lors de la conception d'une interface sonore. En effet, l'ignorance de ces effets peut conduire au fait qu'un utilisateur ne peut différencier deux sons, ou ne peut les entendre. On considère généralement quatre caractéristiques psycho-acoustiques : la hauteur, le volume, le timbre et la localisation [BGB90].

### La hauteur

La hauteur correspond à la fréquence fondamentale, mais sa perception n'est pas linéaire par rapport à la fréquence. Ainsi, tripler une fréquence entre 1 kHz et 3 kHz ne fait que doubler l'impression subjective de hauteur. De plus, la hauteur d'un son est influencée par son intensité : à moins de 2 kHz, une augmentation d'intensité diminue la hauteur perçue. Enfin, les timbres « brillants » , c'est-à-dire ceux des sons dont les hautes fréquences ont une grande énergie, paraissent plus haut que les sons sourds.

En général, notre oreille n'est pas « absolue » , c'est-à-dire qu'elle ne peut trouver la fréquence des sons qu'elle entend. En revanche, beaucoup de gens ont une oreille « relative » : ils peuvent différencier deux sons de hauteurs différentes (bien que certaines soient incapables de dire lequel des deux est le plus haut). Ces deux derniers faits indiquent qu'il ne faut pas essayer de communiquer des valeurs absolues avec des hauteurs ; l'affichage de l'évolution d'une fonction selon une dimension (que l'on associera au temps) est plus approprié à ce type de médium [Bly85].

### Le volume

Le volume correspond à l'amplitude d'un son selon la relation

$$L = kI^{0.3}$$

avec $L$, volume en Sones (une unité subjective), $I$, intensité en dB et $k \approx 0.3306$. Cette relation dit qu'une augmentation d'intensité de 10 dB double le volume.

Les autres facteurs qui influent sur la perception du volume sont la fréquence et la largeur de bande (phénomène de masquage). Le volume dépend de la fréquence : ainsi, pour une intensité donnée, un son à 1000 Hz n'aura pas le même volume qu'un autre à 100 Hz. La relation entre le volume et la fréquence est très compliquée, il suffit pour s'en convaincre de se référer à [BGB90]. Quant au phénomène de masquage, il se traduit par l'atténuation d'un son par un autre. Il dépend des fréquences mises en jeu par ces sons : pour les sons dont la fréquence tombe dans une bande critique, l'énergie est additionnée, alors qu'à l'extérieur de cette bande, c'est le volume qui est additionné. De plus, les sons avec une large bande de fréquence sonnent plus fort que ceux avec une bande étroite. Donc si une information doit être présentée avec le volume, le son correspondant doit être assez fort pour éviter le masquage par des bruits ambiants.

En général, les êtres humains ne sont pas capables de juger absolument le volume d'un son. De plus, l'échelle des jugements relatifs est limitée à trois niveaux. Ces pauvres capacités et les phénomènes vus plus haut obligent les concepteurs à être très prudents quant à l'utilisation du volume. De plus, il ne faut pas oublier que les utilisateurs ont souvent un contrôle du volume général d'une machine, et qu'ils peuvent de ce fait perdre des informatisons si le niveau est trop bas.

### Le timbre

Le timbre occupe une place particulière au sein des termes descriptifs d'un son, car il n'existe ni définition précise, ni relation mathématique qui puisse le caractériser formellement. La définition qu'en donne l'ANSI [ano73] est d'ailleurs plutôt vague : c'est « l'attribut auditif qui permet de distinguer deux sons de même hauteur et de même fréquence » . Pourtant, le timbre est une des caractéristiques d'un son les plus immédiates et les plus facilement reconnaissables.

Le timbre reste donc en grande partie inconnu.On peut dire qu'il dépend au moins du spectre et des enveloppes de chaque partiels. Actuellement, les chercheurs dirigent leur travaux sur la source du son plutôt que le son lui-même, afin de mieux contrôler ce dernier : c'est une approche par modèle physique, que nous verrons plus loin.

### La localisation

La localisation est la capacité à déterminer la position d'une source sonore dans l'espace. La localisation droite-gauche se fait grâce à deux phénomènes : la Différence Interaurale d'Intensité (DII), et la Différence Interaurale Temporelle (DIT). Comme l'homme dispose de deux oreilles espacées de quelques centimètres, les sons ne sont pas perçus de la même façon par chaque oreille.

La DIT provient du fait que le son ne met pas le même temps pour parvenir à chaque oreille : il existe une différence dont notre cerveau se sert pour construire la perception spatiale. Cette différence est d'autant plus importante que la source est située sur les côtés de la tête (c'est-à-dire que la source est proche d'une plan passant par les oreilles et les pieds) et que la source est proche de nous. En effet, plus la source est éloignée, moins la différence temporelle est importante. La DIT est donc négligeable pour les sources éloignées mais utile pour ceux qui sont proches.

La DII quant à elle, est la différence perçue entre les intensités. Elle est due à l'ombre sonore que crée la tête d'une personne. Si l'onde a une longueur d'onde plus grande que la largeur de la tête, la différence ne sera pas perçue. En revanche, si elle est plus petite, une ombre sonore se forme, différenciant ainsi les sons reçus par chaque oreille.

L'impression de distance est due à l'affaiblissement de l'intensité du signal et à l'appauvrissement du spectre. En effet, les fréquences basses voyagent mieux que les fréquences plus hautes, dont les pertes sont facilement constatables. Il résulte de ce fait que les sons sourds sont perçus plus éloignés que les sont brillants.

D'autres apects interviennent de manière importante dans la localisation, en particulier la façon dont les sons sont filtrés par l'oreille externe, la tête et le haut du corps. Des systèmes comme le *convoltron* [S.H91] modèlisent ces effets à l'aide de filtres, mais ils demandent beaucoup de puissance de calcul.

## Les Synthèses Sonores

Comme on l'a vu plus haut, les sons sont des variations de pression dans un milieu élastique. Le signal que représente l'amplitude en fonction du temps est donc un signal analogique. La numérisation de signaux analogiques permet leur traitement informatique. On peut donc représenter le processus de synthèse sonore comme un système ingérant une suite de nombres représentant un signal, et produisant une sortie correspondant à une transformation mathématique de l'entrée.

Cette section décrit les diverses classes de synthèses utilisées par les informaticiens et les musiciens. Pour des informations plus détaillées, voir [Moo90].

### Synthèse additive

Le principe de la synthèse additive est dérivé du théorème de Fourrier. Celui-ci dit que toute fonction périodique peut s'écrire comme la somme de fonctions sinusoïdales de fréquence, d'amplitude, et de phase différentes. Chaque composant est très simple à synthétiser, mais leur addition permet de produire des timbres très complexes, très riches et, en théorie du moins, identiques aux instruments simulés.

En effet, il suffit d'analyser un son pour déterminer chaque composante avec ses caractéristiques respectives, puis de les reproduire à l'identique pour obtenir un son très proche de l'original. C'est le principe qu'adopte le *phase vocoder*, ou *vocoder de phase* : on découpe la bande passante en plusieurs bandes étroites, qui sont analysées indépendamment (ou presque). On peut reconstruire le signal original en utilisant une sinusoïde par bandes, chose plus facile que d'imiter le son en entier. Le vocoder de phase est aussi utilisé pour créer des effets spéciaux. En effet, plutôt que de rejouer le son original, on peut appliquer un traitement sur certaines de ses composantes seulement, ce qui peut donner par exemple des voix robotisées.

Cette technique est donc très puissante mais elle a un défaut de taille : elle est trop gourmande en calcul. Ainsi, pour un son de piano de qualité CD (44,1 kHz), il faut 400 sinusoïdes pour synthétiser le La le plus grave. Si l'on veut une polyphonie de 32 voies, le nombre de partiels à traiter simultanément reste hors de portée des techniques actuelles. Cependant, de nouvelles techniques basées sur des composants VLSI voient le jour, et semblent assez prometteuses [SC92].

### Synthèse soustractive

La synthèse additive produit un son à partir de sinusoïdes primitives. La synthèse soustractive est basée sur l'idée complémentaire de passer un signal complexe dans un filtre afin de produire la forme d'onde voulue. C'est une technique de sculpture, le matériau étant le son en entrée, le filtre jouant le rôle du marteau et du ciseau. Ce filtre est généralement variable dans le temps.

On associe souvent cette technique au principe de *modèle physique*, car plusieurs instruments fonctionnent de cette façon. Une corde de violon est une excitation qui est filtrée par le corps du violon, la voix humaine est un bruit modulé par les cordes vocales, etc.

Tout l'art du programmeur est de concevoir des filtres qui imitent le comportement de tel ou tel instrument. Un des exemples les plus réussis est celui de l'algorithme de Karplus et Strong [Kev89], qui synthétise des cordes pincées ou frappées à partir d'un bruit blanc (suite aléatoire d'amplitudes). Cet algorithme est très simple et peu coûteux, et pourtant le son produit est réaliste, « acoustique » ; en particulier, son principe est tel que deux notes identiques jouées à la suite avec un tel algorithme ne sont pas exactement identiques.

### Synthèse non linéaire

Les synthèses vues précédemment sont dites linéaires. Elles peuvent être décrites comme un système recevant une entrée et produisant une sortie. Si pour un tel système, une entrée $x_1$ produit une sortie $y_1$ et une entrée $x_2$ produit une sortie $y_2$, alors une entrée ax1 + bx2 produira une sortie $ay_1 + by_2$. Les systèmes qui ne respectent pas cette règle sont dits *non-linéaires*.

Les synthèses non-linéaires sont généralement des générations de formes d'ondes complexes à partir de formes simples grâce à des opérations non-linéaires. La synthèse

non-linéaire la plus connue est sans doute la synthèse *FM* (*Frequency Modulation* ou *Modulation de Fréquence*) de Chowning [Cho73] , popularisée par Yamaha en 1983 avec son DX7. Elle consiste en une fréquence de base $f_c$ (*la porteuse*) modulée par une autre fréquence $f_m$ (*la modulation*) selon la formule :

$$f(n) = f_c + \partial f \cos( 2\pi n f_m / R)$$

avec $f(n)$ la fréquence instantanée à l'échantillon $n$, $\partial f$ la déviation maximum par rapport à la porteuse, et $R$ le taux d'échantillonnage.

Ce type de synthèse produit des spectres complexes, mais les formes d'onde obtenues ne correspondent pas exactement à un type de son réel précis. Néanmoins, les méthodes non-linéaires donnent un moyen efficace de contrôle de paramètres acoustiques, comme la largeur de bande d'un spectre harmonique.

### Synthèse par modèle physique

Les moyens informatiques toujours plus puissants et moins coûteux ont permis à un nouvelle forme de synthèse d'émerger ces dernières années. Certains chercheurs ont en effet décrit la source sonore plutôt que le son lui-même. La création de sons acoustiques y a beaucoup gagné car la technique utilisée imite le comportement d'un instrument. En effet, une description complète de l'instrument est introduite dans l'ordinateur (un *modèle physique*), et une excitation « virtuelle » lui fait produire un son [Roa94].

Les modèles mathématiques des instruments utilisent généralement des systèmes d'équations différentielles. Cependant, leurs résolutions demandent des techniques d'approximations successives très gourmandes en temps de calcul. Par exemple, les cordes sont modélisées par le modèle des masses et des ressorts. De ce fait, la moindre simulation de pincement ou de frottement demande un travail énorme des calculs complexes. Une technique plus rapide semble cependant émerger et pourrait avoir des retombées industrielles d'ici peu. Il s'agit de la technologie *Waveguide* ou *Guide d'Onde*, dont le principe résoudrait en grande partie les problèmes de résolutions d'équation différentielles [Smi92].

Malgré cette complexité, cette nouvelle approche est très utile, et son concept très intéressant pour la conception d'interfaces sonores.

## Exemples d'utilisation du son non-parlé

Il existe plusieurs exemples d'utilisation des sons dans l'interface homme-ma-chine. On peut rassembler ces cas en deux ensembles : les messages codés et les messages d'état et de contrôle.

### Les messages codés

Les messages sonores codés sont utilisés comme une alternative au graphique pour transmettre des données.

Ainsi, Sara Bly estime que certaines fonctions sont difficiles à représenter graphiquement : les fonctions multidimensionnelles, celles variant en fonction du temps, ou les fonctions logarithmiques [Bly85].

Pour les fonctions multidimensionnelles, Bly utilise sept caractéristiques sonores : hauteur, volume, durée, forme d'onde, enveloppe d'attaque, et addition d'une 5e et d'une 9e harmonique. Lors d'un test, des données à six dimensions sont représentées de

manière graphique ou sonore, ou les deux à la fois. L'expérience montre que l'utilisation du son combinée au graphisme augmente sensiblement la capacité d'une personne à dire si une variable appartient à un ensemble ou à un autre.

Bly utilise une technique similaire pour les données temporelles. Cependant, elle ne fait varier que la hauteur et le volume pour une fonction donnée tout en attribuant à chaque fonction un timbre qui permet de la différencier des autres. Ainsi, il est plus facile de comparer l'évolution relative de deux fonctions au cours du temps.

Enfin, pour les données logarithmiques, elle ne fait varier que la hauteur. En effet, la hauteur est elle-même une échelle logarithmique si on la représente en fonction de la fréquence, elle s'applique donc bien à ce type de fonction.

Mezrich, quant à lui, utilise la technique de Bly pour les données dépendantes du temps. Les données sont présentées par des mélodies, dont les notes correspondent à leurs valeurs [Bly85]. Comme pour Bly, les tests se sont révélés concluants : l'utilisation du son permet aux utilisateurs une meilleure distinction des fonctions que si elles sont affichées avec le graphisme seulement.

Ces techniques de présentation permettent par exemple à des étudiants en chimie non-voyants de reconnaître des spectres infra-rouges de diverses substances [Pet85] [Bly85] Les fréquences des notes entendues correspondent aux fréquences du spectre infra-rouge, leur durée à l'intensité des pics. Les notes sont jouées dans l'ordre décroissant des fréquences pour la première mélodie, dans l'ordre décroissant des intensités pour la deuxième, et enfin toutes ensembles (formant un accord presque toujours dissonant). Ces différentes auditions permettent de retrouver les substances correspondantes. Ainsi, selon Peterson, «  L'éthanol sonne comme une danse extra-terrestre. Le polystyrène part d'une fugue sombre, pour arriver à un tintement de cristal... » .  Les indicateurs d'état et de contrôle}

Les exemples vus précédemment concernaient la présentation de données par un moyen autre que le graphisme. Cette section aborde l'utilisation du canal auditif pour étendre et enrichir l'interaction homme-machine.

### SoundTrack

L'une des applications des interfaces sonores est l'accès aux ordinateurs par les aveugles et mal-voyants. L'étude de tels systèmes est non seulement utile pour les non-voyants, mais aussi pour les voyants, car elle montre les chemins à suivre quand l'écran est surchargé et que nous devenons nous aussi d'une certaine manière partiellement mal voyants.

Ainsi, Edwards propose SoundTrack [Edw89], une interface sonore pour un traitement de texte. SoundTrack utilise à la fois des sons parlés et des sons musicaux. L'écran est divisé en zones correspondant à des menus. Quand le curseur de la souris passe au-dessus d'une de ces zones, un son caractéristique de cette zone est émis : c'est une information brève de l'endroit où se trouve le curseur. Les sons utilisés sont des *signaux carrés*  (sons simples à synthétiser) dont la hauteur varie en fonction de la position de la souris de gauche à droite et de haut en bas. Si l'utilisateur veut plus d'information, il clique sur la zone survolée et son titre est dicté par le synthétiseur vocal. Pour aller plus loin, l'utilisateur double clique sur le menu et se retrouve dans un autre menu.  Lors des tests, il est apparu que la hauteur des signaux n'était pas utilisée. Les utilisateurs comptent le nombre de changements de tons (c'est-à-dire le nombre de zones traversées), pour savoir où se trouve le curseur.

### Les  Earcons

Dans [BSG89], Blattner, Sumikawa et Greenberg définissent l'équivalent sonore des icones : les *earcons*. Ils servent à donner des informations sur des objets (par exemple des fichiers), des opérations (par exemple éditer) ou des interactions (par exemple détruire un fichier).

Les auteurs considèrent trois classes d'icones picturales (cf. Figure 6) :

- *représentatives* : ce sont de simples images familières d'objets ou d'opérations. Ils sont facilement reconnaissables, mais malheureusement, tous les objets ou toutes les opérations n'ont pas de représentation picturale évidente. De plus, ces images peuvent apporter des informations parasites, car elles peuvent être trop détaillées.
- *abstraites* : ce sont des combinaisons de formes géométriques décrivant de manière conventionnelle un objet ou une opération qui ne peut être représentée par une image.
- *semi-abstraites* : ce sont soit un mélange d'icones représentatifs et d'icones abstraits, soit un icone représentatif tellement simplifié que l'on peut le considérer comme abstrait.

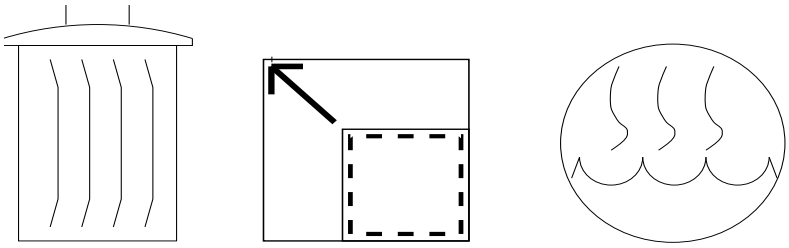


Fig 6 - *Les différents types d'icones (représentatif, abstrait, semi-abstrait)*

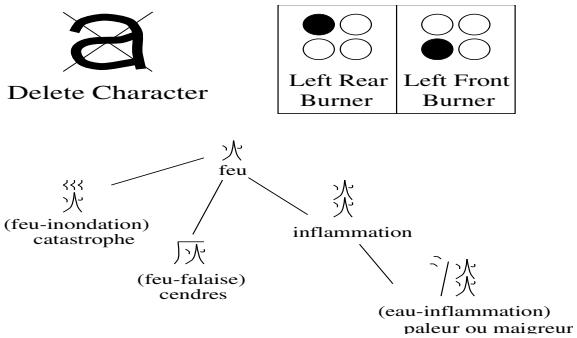



Fig 7 - *Les compositions d'icones (combinaison, transformation, héritage)*

Les auteurs admettent que l'échelle est continue (il est parfois difficile de classer un icone). Par ailleurs, ils considèrent que les icones sont soit des éléments, soit des compositions d'éléments. Ils présentent trois façons de créer des icones composées (cf. figure 7) :

- *la combinaison* : comme son nom l'indique, c'est un mélange d'icones ;

- *la transformation* : il s'agit d'une modification légère d'un icone connu. Les icones composés grâce à une transformation représentent souvent un objet avec un état associé (cf. l'exemple 2 Figure 7) ;
- *l'héritage* : c'est une transformation telle que l'ensemble des transformations se rapportant à un icone peut s'organiser en arbre hiérarchique. Plus on descend dans l'arbre, plus la signification d'un icone se spécialise. Le système d'écriture japonais, dont un exemple est donné Figure 7 fonctionne en partie grâce à l'héritage.

Blattner reprend ces principes pour les *earcons* : les éléments de base sont des *motifs* (cf. Figure 8) (petites mélodies de deux ou trois notes aisément reconnaissables par leur rythme, leur hauteur, et leur volume), qui peuvent être combinés, transformés ou hérités.

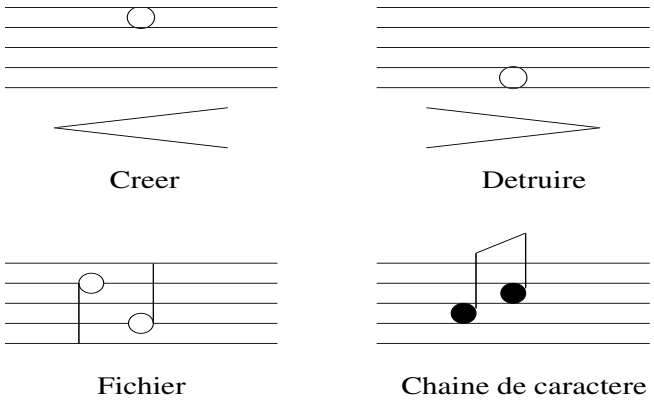
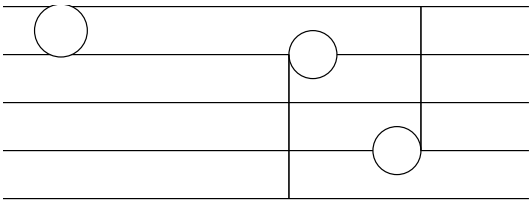



Fig 8 - *Des motifs*

Ainsi, si un mi *crescendo* (augmentation du volume) correspond à *créer* et si deux blanches, ré et sol correspond à *fichier*, ces deux notes jouées après un mi crescendo correspondent à *créer un fichier*. C'est un exemple de combinaison, comme pour les icones (cf. Figure 9)



Creer un Fichier

Fig 9 - *Une combinaison de motifs*

L'héritage est une peu plus complexe : chaque niveau correspond à un attribut (dans l'ordre hiérarchique : rythme, hauteur, timbre, dynamique et registre ou octave). Les familles de messages sont distinguées par le rythme (message d'erreurs, fenêtre...). Le deuxième niveau est une mélodie jouée au rythme de la famille à laquelle l'objet appartient. Cette mélodie est jouée avec une sinusoïde, car c'est le timbre le plus neutre. Le troisième niveau correspond au changement de timbre et à une petite transposition positive. Enfin, la dynamique et le registre interviennent dans le dernier niveau. Ainsi, quand on écoute un *earcon* on entend : un rythme, une mélodie, une mélodie avec un timbre différent jouée un peu plus haut, puis cette même mélodie avec un dynamique ou jouée dans un autre registre (Figure 10)

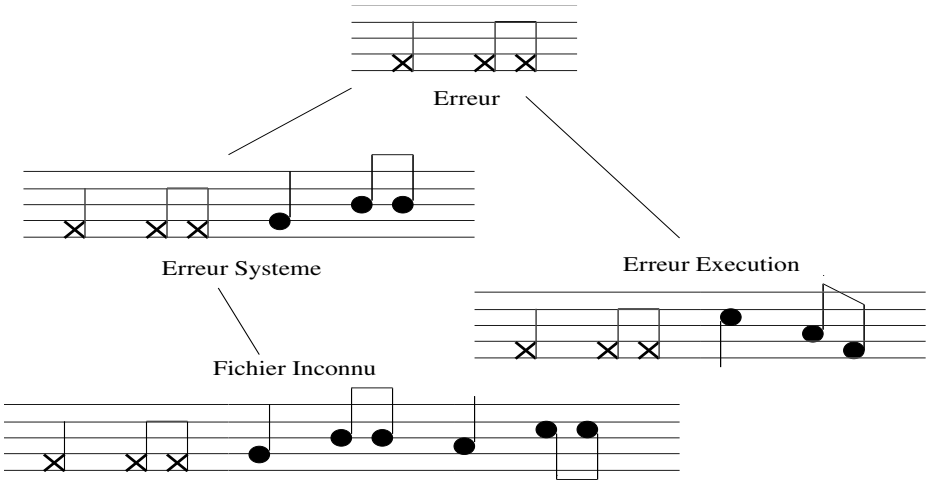



Fig 10 - *Un exemple d'héritage*

La transformation permet de «  sauter »  des étapes dans l'écoute d'un earcon. En effet, les utilisateurs experts connaissent au bout d'un certain temps le rythme caractéristique d'un message d'erreur par exemple, et ils n'ont pas besoin d'entendre ce rythme dans l'earcon correspondant. On supprime donc cette étape, l'utilisateur reconnaissant la famille de l'earcon au rythme de la mélodie.

L'avantage de cette structure hiérarchique réside dans le fait que l'ajout d'une fonctionnalité ne nécessite pas forcément l'ajout d'un earcon entièrement nouveau : ainsi, si une nouvelle erreur système doit être notifiée par un earcon, cette dernière héritera au moins du rythme et de la mélodie de la famille des erreurs système. L'apprentissage de nouvelles earcons n'est donc pas forcément difficile.

En revanche, il faut bien s'entraîner au moins une fois à reconnaître les types d'earcons, la famille à laquelle ils se rattachent etc. Les earcons hiérarchiques peuvent donc être classés dans la catégorie « abstraite » , nécessitant un effort de la part de l'utilisateur. Bien que l'apprentissage de nouveaux earcons ne soit pas obligatoirement difficile, il se peut que leur nature abstraite rende leur reconnaissance plus difficile si leur nombre est important.

### Auditory Icons

Pour présenter des informations, Bly utilisait les caractéristiques du son en lui-même comme la hauteur, le volume, le timbre etc. Ceci résultait en des sons qui n'avaient aucune correspondance avec notre environnement naturel. De la même façon, les earcons n'ont aucun analogue dans notre vie quotidienne. Ils nécessitent donc un apprentissage.

Pourtant, quand une porte se ferme, nous n'entendons pas le son en terme de hauteur, de volume, ou de timbre. Nous entendons une porte d'une certaine taille, d'une certaine matière qui se ferme avec une certaine force dans une pièce d'un certain volume. En fait, nous utilisons le son afin d'identifier les événements qui l'ont produit. Les caractéristiques de bas niveau d'un son nous intéressent peu : ce qui nous importe, ce sont les caractéristiques de la source sonore et de l'interaction sur cette source qui a donné naissance à ce son. Donc, si nous raisonnons en termes de sources et d'interactions, plutôt qu'en termes du son lui-même, nous pouvons esprérer communiquer des informations beaucoup plus naturellement.

En fait, un son fournit des informations sur des objets interagissant dans un environnement. Cette approche s'inspire de l'*approche écologique*  [Gav93b] et est

utilisée par Gaver pour ses *Auditory Icons* (ou*icones auditifs*) [Gav86]. Les icones auditifs sont des imitations de sons naturels comme des impacts ou des frottements. Le terme « imitation » est préferable au terme « reproduction » : il n'y a nullement besoin d'une reproduction parfaite. En effet, ces sons essaient de capturer les principales caractéristiques des sons naturels de telle façon qu'ils soient aisément identifiables. Ils correspondent en fait aux caricatures graphiques, ou aux dessins animés, où certains traits sont accentués.

Dans le système de classification de Blattner, il existe un ordre de facilité d'apprentissage. Ainsi, les icones représentatifs sont plus facilement reconnaissable que les icones semi-abstraits, eux-même plus facilement reconnaissables que les icones abstraits. Comme nous l'avons vu, la même hiérarchie peut être utilisée pour les sons, les earcons étant classés dans la troisième catégorie. Le fait que Gaver utilise des sons naturels, que nous avons appris à connaître et à reconnaître, permet une représentation beaucoup plus directe que les earcons. On peut donc classer les Auditory Icons dans la première et deuxième catégorie (représentative et semi-abstraite).

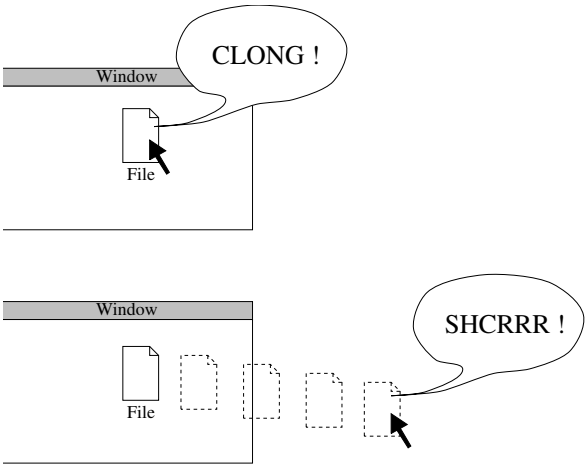



Fig 11 - *Sonic Finder : Sélection et déplacement d'un fichier*

En réalisant le SonicFinder [Gav89], Gaver a intégré les Auditory Icons à l'interface du Macintosh. Un exemple d'utilisation est celui de la destruction d'un fichier (cf. Figure 11). Pour détruire un fichier sur un Macintosh, on le sélectionne d'abord en cliquant dessus. Cette action est accompagnée d'un son d'impact. Le bruit du choc indique le type du fichier et sa taille. Par exemple, un fichier texte a un son mat, comme un bout de bois, alors qu'une application a un son brillant. De plus, plus le fichier est gros, plus on a l'impression que l'on a tapé sur un objet gros (le son est plus grave). Ainsi, le son est défini grâce à une source (fichier = objet) et une interaction (pointer = taper).

Le fait qu'à un type d'objet informatique (fichier texte, application) correspond un objet d'une certaine matière (bois, métal), classe cette représentation dans la catégorie semi-abstraite : le son est reconnu, mais la correspondance doit être apprise. En revanche, la correspondance entre la taille de l'objet informatique et la taille de la source sonore est totalement naturelle, donc à classer dans la catégorie représentative.

Il faut maintenant déplacer le fichier dans la poubelle, ce qui se réalise en faisant glisser le fichier sur le bureau. Ici, la source n'a pas changé (c'est toujours le même fichier). En revanche, l'interaction est passé du type « pointer » au type « glisser » , il faut donc faire évoluer le son en conséquence. Le son correspondant à l'action « glisser » est un frottement. Mais le changement d'interaction ne change pas la perception des caractéristiques de l'objet que l'on avait acquises lors de la première interaction. Si l'objet est gros, on entend un frottement d'un gros objet. S'il est en bois, on entend un frottement sur du bois. Il est important de distinguer les caractéristiques de l'objet de

celles de l'interaction. Ainsi, il doit être possible de reconnaître le même objet dans deux interactions différentes.

Gaver a aussi utilisé les Auditory Icons dans des tâches de collaboration : ARKola et EAR [Gav91]. La simulation ARKola il modélise une usine de soda, dont le but est de fabriquer du cola et de le mettre en bouteille. L'usine est conçue de telle manière qu'il faut au moins deux écrans pour la surveiller en entier. On fait donc collaborer deux personnes, qui doivent s'entendre afin de gérer au mieux l'usine. Généralement, une tâche de collaboration est rendue difficile par le fait que l'on veut permettre un travail individuel, et le fait que l'on doit toujours penser au but commun. L'apport d'un retour d'information sonore permet d'atténuer cette frontière. Ainsi, dans ARKola, le fait de travailler sur une partie de l'usine n'empêche pas de se tenir au courant de la partie invisible, grâce aux sons qu'elle produit. On se tient ainsi au courant de l'état global de l'usine. Par exemple, on peut entendre des sons de bouteilles qui se brisent, et le signaler à son collègue pour qu'il corrige le défaut.

Avec EAR, les sons sont utilisés pour signaler des événements à des personnes travaillant dans un même bâtiment. Ainsi, lorsqu'un groupe de personnes prend le thé, elles peuvent le signaler à leurs collègues en diffusant un son de thé versé dans une tasse. De la même façon, le début d'une réunion est signalée par le son de murmures augmentant en nombre et se terminant par un bruit de maillet. Ces sons n'interrompent pas le travail, et peuvent être facilement ignorés. De même, des événements relatifs à l'utilisation d'un Mediaspace peuvent être signalés par des signaux sonores. Ainsi, quand quelqu'un se connecte sur une caméra, on entend un grincement de porte qui s'ouvre, et quand il quitte la connexion, un son de claquement de porte se produit. Cet environnement sonore permet à chacun de se tenir au courant des activités de ses collègues

Enfin, le concept d' Auditory Icons a été adapté au monde Unix avec le serveur ENO [BLG94], que nous décrivons maintenant.

## ENO

Comme les Auditory Icons, ENO est basé sur la compréhension de l'origine des sons : les sons sont issus de *sources sonores* sur lesquelles se produisent des *interactions*. Les sources et les interactions sont caractérisées par un type et des attributs, comme la taille d'un objet, ou la force d'un impact. Un impact peut être considéré comme une interaction élémentaire. Des structures plus complexes peuvent être construites, en associant plusieurs sons de sources différentes (parallèlement ou séquentiellement), ou plusieurs sons d'une même source disposée temporellement d'un façon particulière (comme pour produire un effet de rebond).

Les sources sont placées dans un espace sonore (cf. Figure 12) qui permet non seulement de les distinguer quand plusieurs d'entre elles jouent en même temps, mais aussi de leur attribuer de nouvelles caractéristiques aisément identifiable. Par exemple, on peut émettre des messages urgents en premier plan alors que des sons de surveillance se situeront plutôt en arrière plan. Ou encore, les sons générés par une application viennent tous d'un même endroit. Beaucoup de systèmes d'écoute 3D ont été réalisés, mais ils demandent beaucoup de ressources matérielles et surtout nécessitent l'utilisation d'un casque. La solution retenue dans ENO est un espace sonore 2D (un plan horizontal), solution suffisante en regard de l'usage de la spatialisation.

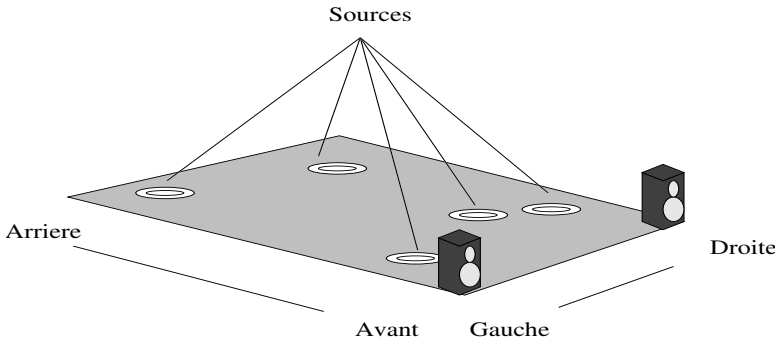



Fig 12 - *Espace sonore d'ENO*

### Les sources et l'arbre des sources

L'élément principal de l'architecture de ENO est la *source*. Ces sources produisent un son grâce à une excitation, ou *interactio*}. Les sources et les interactions ont un type et une ensemble d'attributs.

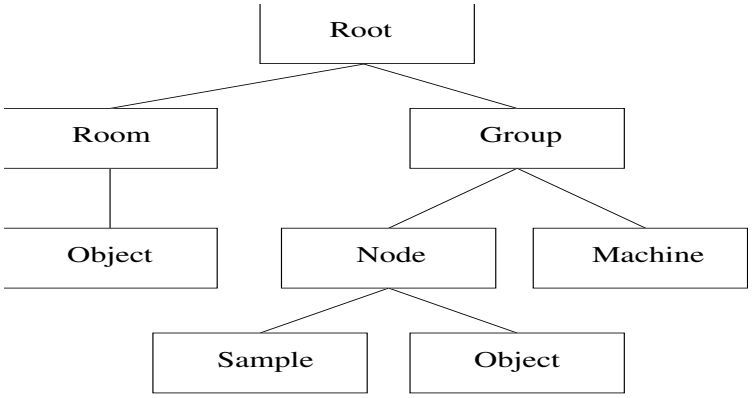



Fig 13 - *Structure d'arbre d'ENO. Les attributs sont hérités le long des branches.*

Les sources sont organisées en arbre, ce qui permet de partager des attributs par héritage, et de contrôler les *sources primitives* (les sous-sources) par les sources de *haut niveau* (cf. Figure 13)

Les attributs sont hérités par les sources inférieures : si un attribut n'est pas défini par une source, il prend la valeur de celui de sa source mère. Certains attributs sont relatifs, ce qui signifie que la valeur d'un tel attribut pour une source est une combinaison des attributs de ses parents. Ainsi, le gain final d'une source est la somme des gains de ses parents et du gain propre à cette source. Ainsi, on peut contrôler le gain d'un ensemble de source en modifiant le gain d'une source parente commune. Ce principe reprend celui des systèmes de fenêtrage comme X Window, dans lesquels les positions des fenêtres à l'écran sont déterminées par rapport aux positions des fenêtres parentes.

### Les sources primitives

Les sources primitives sont les feuilles de l'arbre. Elles sont pour l'instant de trois types (leurs attributs sont indiqués entre parenthèses) : les objets (taille, matériau), les machines (taille, vitesse, travail maxi, travail mini, rapport) et les échantillons (nom du fichier). Il y a deux interactions sur les objets, l'impact (force) et le frottement (force et vitesse), et une sur les machines et les échantillons, la mise en marche (travail maxi, vitesse) et la lecture (nombres de boucles). Les attributs de gain, de position en x et en z sont communs à toutes les sources.

L'une des difficultés lors de la conception d'un tel serveur, est de faire en sorte qu'il produise des échantillons à une vitesse supérieure à celle de la fréquence d'échantillonage. Sinon, des « clics» seront audibles lors de la reproduction. C'est pour cela que les algorithmes de synthèse se doivent d'être performant.

Gaver a conçu deux algorithmes [Gav93a] repris dans ENO :

- *les objets* : Gaver constate que le son produit par des objets ont le même spectre pour les deux interactions (impact et frottement). Il décrit donc les objets à l'aide d'une banque de filtres qui accentue les fréquences dont les caractéristiques dépendent de celles de l'objet (taille et matériau). Le signal traité par cette banque est soit une sinusoïde de courte durée (impact), soit un bruit blanc (frottement).

- *les machines* : Pour ce son, la synthèse FM est utilisée. En effet, une machine produit un bruit qui est assez répétitif et qui semble avoir un mouvement cyclique. En utilisant la modulation autour de la porteuse, on peut reproduire ces caractéristiques.

Les sons des objets sont d'un grand réalisme, alors que ceux des machines le sont moins. Ceci est dû aux approches choisies par Gaver les concevoir ces deux types de sons. Dans le premier cas, il a étudié le son de plusieurs objets de taille, de forme et de matériau différents, ce qui lui a permis d'en définir les principales propriétés. Dans le deuxième cas, il n'a pas étudié de vraie machine (pour la bonne raison qu'il y en a une multitude et qu'aucune n'est vraiment familière à tout le monde). Il s'est donc interessé aux caractéristiques qu'il voulait inclure et a créé un son qui les reprenait, sans qu'il existe vraiment. En fait, l'important n'est pas que les sons soient d'un grand réalisme : il suffit qu'on les reconnaisse et que leurs caractéristiques influent réellement sur le son produit.

Par ailleurs, les algorithmes de synthèse des sons d'objets et de machines permettent un contrôle des paramètres des sources et des interactions, même pendant que le son est joué. Cette particularité donne aux applications un moyen de modifier un son en temps réel, et de répercuter ainsi un changement d'état du système.

### Les sources de haut-niveau

Deux types de sources de haut-niveau ont été implémentés : les*nœuds* et les *groupes*. Les *nœuds* sont de simples regroupement de sources sous un ancêtre commun, ce qui permet de manipuler le gain et la position de l'ensemble facilement. Les groupes sont des *nœuds* qui permettent de jouer des sources de deux façons : séquentiellement et parallèlement. Il est prévu un troisième type de séquencement, le rebond.

### L'espace sonore

Toutes les sources sont positionnées dans l'espace sonore. Comme nous l'avons vu, cet espace est en deux dimensions : la position verticale (élévation) n'est pas incluse dans ENO, car elle nécessiterait des calculs trop importants et le port d'un casque (ou l'utilisation de quatre hauts-parleurs), contrainte que l'on souhaite éviter.

Le positionnement droite-gauche est obtenu en faisant varier les amplitudes des canaux stéréo : plus un son se situe d'un côté, plus l'écart de volume entre les haut-parleur droit et gauche est important.

L'impression d'éloignement fait appel à deux techniques. La première est l'atténuation du volume en fonction de la distance. Plus un son est faible, plus il nous paraît lointain. La deuxième part de la constatation que les fréquences graves voyagent mieux dans l'aire que les fréquences aigues, qui sont atténuées : un son lointain semble étouffé. Il faut donc filtrer le son pour lui enlever une proportion variable de ses fréquences aigues en fonction de la distance.

Ces techniques ont l'avantage d'être peu gourmandes en calcul, pour des résultats satisfaisants.

## Conclusion

Le domaine des sons non parlés a été encore peu étudié dans les interfaces. Les machines d'aujourd'hui permettent de développer des applications utilisant le son de manière sophistiquée, ce qui n'était pas le cas jusqu'à présent, car la synthèse de sons en temps réel demande des capacités de calcul importantes. En conséquence, on a surtout vu jusqu'à présent des systèmes dédiés à des applications spécifiques, souvent développés sous forme de prototypes. On peut espérer qu'à relativement court terme apparaitront des outils du type de ENO pour faciliter l'intégration du son non parlé dans les interfaces.

## Références


[ano73] anonymous. American national psychoacoustical terminology. Technical Report S3.20, American National Standards Institute, New York, 1973.

[BGB90] William Buxton and William W. Gaver and Sara Bly. The Use of Non-Speech Audio at the Interface - Section 2: Acoustics and Psychoacoustics. In Tutorial notes of CHI'90 Conference on Human Factors, volume 4, pages 2.1-2.15. ACM, 1990.

[BLG94] Michel Beaudouin-Lafon and William W. Gaver. ENO: Synthesizing Structered Sound Spaces. In *Seventh Annual Symposium on User Interface Software and Technology (UIST'94)*, pages 49-57. ACM Press, November 2-4 1994.

[Bly85] Sara Bly. Communicating With Sound. In *Proceedings of CHI'85 Conference on Human Factors in Computing Systems*, pages 115-119. ACM, 1985.

[BSG89] Meera M. Blattner and Denise A. Sumikarwa and Robert M. Greenberg. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, 4(1), 1989.

[Cho73] John Chowning. The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. *Journal of the Audio Engineering Society*, 21:526-534, 1973.

[DLH88] David J. DeFatta and Joseph G. Lucas and William S. Hodgkiss. "Digital Signal Processing: A System Design Approach". Wiley, 1988.

[Edw89] A. Edwards. SoundTrack. *Human-Computer Interaction*, 4(1), 1989.

[Gav86] William W. Gaver. Auditory Icons~: Using Sound in Computer Interfaces. In *Proceedings of CHI'86 Conference on Human Factors in Computing Systems,* volume 2, pages 167-177. ACM, 1986.

[Gav89] William W. Gaver. The SonicFinder: An Interface That Uses Auditory Icons. In *Proceedings of CHI'89 Conference on Human Factors in Computing Systems*, volume 4, pages 67-94. ACM, 1989.

[Gav91] William W. Gaver. Sound Support For Collaboration. In Second European Conference on Computer-Supported Collaborative Work (E-CSCW'91), pages 293-308, September 24-27 1991.

[Gav93a] William W. Gaver. Synthetizing Auditory Icons. *In Human Factors in Computing Systems, INTERCHI'93*, pages 228-325. ACM, April 24-29 1993.

[Gav93b] William W. Gaver. What In The World Do We Hear ? An Ecological Approach to Auditory Event Perception. *Journal of Ecological Psychology*, 5(1), 1993.

[Kev89] Kevin Karplus and Alex Strong. Digital Synthesis of Plucked-String and Drum Timbres. In Curtis Roads, editor, *The Music Machine*, pages 467-479. MIT Press, 1989.

[Moo90] F. Richard Moore. *"Elements of computer music"*. Prentice Hall, 1990.

[Pet85] Ivars Peterson. Some Labs Are Alive With ... The Sound of Data. *Science News,* 127:348-350, 1985.

[Roa94] Curtis Roads. Initiation à la Synthèse par Modèles Physiques. *Les cahiers de L'IRCAM,* pages 145-169, 1994.

[SC92] Julius O. Smith and Perry R. Cook. The Second-Order Digital Waveguide Oscillator. In *Proceedings of the International Computer Music Conference* , 1992.

[S.H91] S.H. Foster and E.M. Wenzel and R.M. Taylor. Real Time Synthesis of Complex Acoustic Environment. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, October 20-23 1991.

[Smi92] Julius O. Smith. Physical Modelling Using Digital Waveguides. *Computer Music Journal*, 16:74-91, 1992. Special issue on physical modelling.