Machines à chanter

Philippe Depalle, Christophe d'Alessandro, Xavier Rodet

Machines à chanter

Christophe d'Alessandro, Philippe Depalle, Xavier Rodet

Résonance nº 8, mars 1995

Appareil de Koenig, machine de von Kempelen, orgue parlant de Faber... Avec les moyens de leurs temps, les savants d'hier ont souvent cherché à comprendre et reproduire les mécanismes compliqués de la production vocale. L'avènement de l'ordinateur a fourni aux chercheurs contemporains les moyens de pousser le mystère de la voix dans ses derniers retranchements. Les machines apprenant aujourd'hui à chanter, c'est tout naturellement que les musiciens se retrouvent aux aguets.

Le langage, faculté d'exprimer la pensée par un ensemble de signes vocaux, entretient des rapports intimes avec la musique : avatars directs du langage, la voix, la parole et la langue se reflètent tout naturellement dans l'activité musicale.

Alors que la voix désigne à la fois le matériau sonore produit par l'être humain et l'instrument producteur de son, la langue représente le système de signes utilisés pour communiquer des idées. Support sonore du langage, la parole en est la manifestation par le double intermédiaire d'un ensemble de symboles et de règles (la langue) et d'un instrument (la voix). De même, le chant est la manifestation de la musique par le double intermédiaire de la forme musicale et de la voix.

Depuis son origine, l'humanité tend vers la démultiplication de ses facultés physiques et mentales. Cette extériorisation de facultés toujours plus élevées passe par l'invention d'outils de plus en plus raffinés. Après la distanciation entre langage et parole (écriture), la description analytique des sons de la langue (alphabet), la mécanisation de l'écriture (imprimerie), la fixation et le transport instantané de la parole (phonographe et téléphone), nous assistons depuis un demi-siècle au développement du traitement automatique du langage, tant sous sa forme écrite que parlée. L'invention de l'ordinateur a en effet rendu possible l'automatisation de la génération et de la compréhension du langage. Machine à traiter l'information, l'ordinateur a introduit le traitement numérique du signal et l'intelligence artificielle à la fois dans l'étude de la langue (linguistique), dans celle de la parole (phonétique) et dans celle de la voix (acoustique).

Grâce aux développements dont il a bénéficié, l'outil informatique est désormais capable de comprendre une requête simple, exprimée en langue courante, et peut aussi comprendre des métaphores ou des analogies. Il se trouve également en mesure d'élaborer un texte à partir d'un ensemble de concepts, comme de lire un texte quelconque de façon intelligible. Il permet en outre à l'utilisateur de transformer la parole en en changeant tour à tour la hauteur, le timbre ou la durée.

Un instrument unique

La voix est un instrument à vent d'un type unique. Comme tous les instruments à vent, elle peut cependant se décomposer en deux entités fonctionnelles : une source (qui produit le son) et un corps sonore.

La source sonore de la voix est multiple : une vibration périodique des cordes vocales (la hauteur du son est alors déterminée, comme pour les voyelles), un bruit (par exemple celui du souffle), ou encore une courte impulsion acoustique (claquement de langue, ouverture rapide des lèvres après accumulation de pression dans la bouche, etc.). Cela peut être enfin un mélange de ces différentes sources.

Comparé aux autres instruments à vent, le corps sonore de la voix, appelé conduit vocal, est court (17 cm

en moyenne chez l'homme adulte). Il se compose des cavités situées au-dessus du larynx : pharynx, cavité orale, éventuellement fosses nasales. Contrairement aux autres instruments à vent, le corps sonore de la voix est trop court pour donner la fréquence fondamentale du son. Il est en revanche extrêmement plastique : par l'action des organes articulateurs (mâchoires, lèvres, langue), il peut être raccourci, allongé, ouvert, fermé, divisé en plusieurs cavités, permettant ainsi de colorer le son. Cette coloration se manifeste par l'amplification de régions spectrales bien définies : ce sont les formants vocaux, liés aux fréquences de résonance, donc à la géométrie du conduit vocal.

Pour simplifier, on peut dire que la source donne au son sa hauteur, sa force et un timbre initial, tandis que le conduit vocal colore le son, jouant ainsi le rôle de filtre acoustique : il s'agit donc d'un modèle source/filtre. De leur côté, les mouvements du conduit vocal permettent d'articuler les sons de la parole.

En fonction des possibilités sonores offertes par l'appareil vocal, chaque langue a adopté un ensemble particulier de sons distinctifs ou phonèmes. Le français en compte au total une trentaine. Premier type de phonèmes, les voyelles résultent de la vibration des cordes vocales. Pour chaque voyelle, le conduit vocal se trouve dans une position stable, caractérisée par un ensemble de formants. Lorsque le nez est également utilisé, les voyelles sont nasales (on, an, un). Second type de phonèmes, les consonnes se subdivisent en plusieurs groupes. Les consonnes fricatives sont produites lorsque le son résulte d'un bruit de friction coloré par le conduit vocal (/f, s, ch, v, j, z/). Si la bouche est fermée, le son est alors rayonné par les narines : ce sont les consonnes nasales (/m, n, gn/). Par les mouvements des articulateurs, la forme du conduit vocal subit des changements rapides qui altèrent ces formes stables et donnent naissance à d'autres consonnes : les plosives, lorsqu'il y a occlusion du conduit vocal et un relâchement brusque (/p, t, k, b, d, g/) ; les liquides, lorsque l'occlusion est incomplète (/l, r/) ; les semi-consonnes, lorsqu'il y a une transition rapide de sons vocaliques (ié, oi).

Machines parlantes

Le but de la synthèse de la parole est de calculer automatiquement un signal vocal à partir d'un énoncé écrit. Un système de synthèse comporte donc plusieurs étapes. Il s'agira tout d'abord de calculer la prononciation du texte écrit en le transcrivant sous forme de phonèmes : c'est la phonétisation. En effet, dans une même langue, une lettre donnée peut parfois se prononcer de différente manières : ainsi du "s" en français. Cette première étape s'avérant insuffisante, il faudra ensuite calculer la prosodie, c'est-à-dire le rythme de la prononciation, les pauses aérant l'énoncé, les courbes de l'intonation et les variations de forces dans la prononciation des syllabes. En définitive, cela revient à maîtriser la structure et le sens de l'énoncé, ainsi que le rapport du locuteur et de l'auditeur, ce qui pose aujourd'hui encore un problème majeur.

Les phonèmes et les variations prosodiques étant désormais connus et les sons définis de manière abstraite, il faudra encore, pour passer à la synthèse sonore à proprement parler, construire l'instrument, le synthétiseur vocal, et le jouer.

Trois familles de synthétiseurs vocaux existent actuellement. La première, qui est aussi la plus ancienne, est issue du modèle source/filtre. A partir de la description des sons à synthétiser, des règles de synthèse permettent d'inférer les paramètres acoustiques du signal de source (fréquence fondamentale, force, durée, etc.) et les paramètres du conduit vocal (formants). Le son est ensuite calculé par un programme qui simule un générateur de source et un filtre évoluant dans le temps. Ce type de synthèse imite les propriétés acoustiques de la voix. Puissants, ces synthétiseurs sont aussi les plus économiques et donc les plus répandus. Reste que beaucoup de soins doivent être apportés à l'analyse des paramètres pour obtenir une qualité sonore acceptable.

Les synthétiseurs articulatoires forment la seconde famille de synthétiseurs vocaux. Ici, les sons ne sont plus décrits par leurs propriétés acoustiques, mais par les propriétés physiques, géométriques et dynamiques de l'appareil vocal. C'est donc le fonctionnement de l'instrument qui est imité et non plus les propriétés du son qu'il produit. Les paramètres de commande du synthétiseur sont alors la position des

articulateurs, ou bien la géométrie du conduit vocal (par exemple les surfaces de coupes transversales le long du conduit vocal). Ces synthétiseurs sont aujourd'hui en plein essor, malgré les difficiles problèmes théoriques qu'ils posent et la puissance de calcul qu'ils exigent.

La troisième famille est celle des synthétiseurs par échantillonnage et concaténation. Le principe est analogue à celui utilisé dans les pianos numériques : un catalogue des différents sons de parole et de leurs enchaînements est établi ; un locuteur prononce ensuite un texte contenant ces sons élémentaires, qui sont alors échantillonnés et découpés. La synthèse consiste à recoller (concaténer) le signal échantillonné correspondant aux sons élémentaires. Il est en outre nécessaire de modifier certaines qualités de ces sons, afin d'appliquer au signal de synthèse les schémas prosodiques calculés préalablement. Ce type de synthétiseurs, particulièrement économiques, donne des résultats sonores d'excellente qualité. Les recherches portent donc d'une part sur l'inventaire optimal des sons à enregistrer et, d'autre part, sur les techniques de modification du signal enregistré. Il s'agit en effet de transformer la voix en préservant l'illusion du naturel.

La voix chantée

Très proche de la voix parlée, la voix chantée s'en distingue cependant par quatre qualités qui lui sont spécifiques, à commencer par la prosodie. Directement issue du texte musical, la prosodie de la voix chantée se différencie en effet radicalement de celle de la voix parlée en ce qu'elle est rigoureusement codée par le texte musical écrit, alors qu'elle n'obéit qu'à des règles approximatives dans le discours parlé.

De même, les voyelles, qui dans la voix chantée portent l'information de hauteur de la note émise, acquièrent ici une importance particulière. Souvent plus longues que dans la parole, elles contribuent aussi davantage à qualifier le timbre de la voix.

Le registre ajoute encore aux différences distinguant la voix chantée de la voix parlée. La tradition occidentale a défini les différents types vocaux en fonction de l'ambitus et du caractère timbral des voix (basse, baryton et ténor pour les hommes, alto, mezzo-soprano et soprano pour les femmes). L'ambitus de la voix chantée est en moyenne plus étendu que celui de la voix parlée (deux octaves et demie contre une et demie). Notons ici que l'appartenance au registre est due aux caractéristiques naturelles de la voix du chanteur, mais aussi à la formation musicale que celui-ci a reçue.

Caractéristique intrinsèque de la voix chantée, le vibrato, enfin, consiste à faire fluctuer régulièrement le son autour d'une hauteur théoriquement fixe. Alors que ces fluctuations peuvent dépasser parfois le demi-ton, le vibrato est globalement perçue par l'oreille comme une caractéristique du timbre de la voix.

Analyser et traiter

Les applications d'analyse de la voix chantée cherchent moins à identifier le locuteur ou le message prononcé, qu'à estimer avec précision l'évolution des paramètres acoustiques de la voix. Ces applications se répartissent en trois types : le suivi de partition, l'aide à la composition et le contrôle de la synthèse ou du traitement.

Le suivi de partition permet au chanteur de piloter l'ordinateur à partir de sa propre voix. L'ordinateur analyse et identifie les événements vocaux émis par le chanteur et les compare à une partition stockée en mémoire ; selon les indications que celle-ci lui fournira, il pourra alors générer des processus de synthèse ou de traitement, ou déclencher toutes sortes d'événements sonores voulus par le compositeur.

En matière d'aide à la composition, certaines caractéristiques de la voix chantée peuvent être récupérées et travaillées par le compositeur pour la construction de certaines structures musicales. C'est ainsi que certaines oeuvres de musique dite « spectrale » recréent des timbres de voyelles par un agencement des instruments qui en reproduit les formants caractéristiques. D'autres compositeurs, tel François-Bernard

Mâche, élaborent des structures musicales dérivées de modèles linguistiques.

Le contrôle du timbre des sons produits ou traités par ordinateur à partir de la voix d'un chanteur reste possible en théorie. Cependant, cette technique rencontre de difficiles problèmes de mise en oeuvre et demeure donc encore peu développée.

Enfin, le traitement de la voix chantée suscite depuis longtemps un grand intérêt de la part des musiciens utilisant l'ordinateur. C'est spécialement vrai des techniques de traitement, tel que le vocodeur de phase, qui autorisent des filtrages très fins et des changements d'échelle temporelle d'excellente qualité.

Machines à chanter

De son côté, la synthèse de la voix chantée se heurte encore à des problèmes délicats, ce qui explique sans doute la grande prudence dont les compositeurs témoignent à son encontre. Sans doute l'extrême complexité et la variabilité de l'instrument vocal lui-même expliquent-elles cette carence. L'oreille humaine montre en outre une exceptionnelle aptitude à reconnaître et apprécier la voix humaine, ce qui ajoute encore à la difficulté.

Plusieurs réalisations de type source-filtre ont cependant prouvé leur aptitude à générer des voix de synthèse suffisamment satisfaisantes pour intéresser les musiciens. Dans chacun de ces systèmes, l'utilisateur contrôle la hauteur et les formants du son, et peut accéder à un ensemble de règles de contrôle de paramètres, notamment pour la synthèse des consonnes.

Le terme « contrôle » désigne l'ensemble des méthodes de génération des paramètres qui vont être recueillis par le synthétiseur. En effet, pour faire « chanter» au mieux le synthétiseur, il est nécessaire de lui communiquer un ensemble de règles définissant le chant (règles de gestion des types de voix, d'évolution du timbre en fonction de la hauteur, de la respiration musicale, etc.). Le contrôle de la synthèse de la voix chantée répond donc à des objectifs fondamentalement différents de ceux à atteindre avec la voix parlée : il s'agit non seulement d'aboutir à une voix aussi « naturelle » que possible, mais encore d'offrir au musicien la possibilité de modifier les sons de la manière la plus souple et la plus générale possible.

Parmi ces réalisations, signalons le programme Musse, développé par Johan Sundberg et son équipe de l'Institut Royal de Technologie (KTH) de Stockholm. Musse, qui était à l'origine un système de synthèse analogique contrôlé par ordinateur, est depuis une dizaine d'années entièrement numérique. La spécificité de ce programme réside dans la possibilité qu'il offre à l'utilisateur d'user d'un très grand nombre de règles déduites de l'étude de la voix chantée. C'est donc un outil remarquable pour l'étude du chant, tant sous ses aspects strictement vocaux que sous ceux de l'interprétation.

Contrairement à Musse, le programme Chant, développé à l'Ircam par l'équipe de Xavier Rodet, synthétise directement des formes d'ondes élémentaires simulant la sortie d'un filtre excité par des impulsions de la glotte. Chaque forme d'onde correspond ainsi à un formant. Chant offre moins de règles d'interaction de paramètres que le système suédois, mais permet de produire un éventail beaucoup plus large de signaux sonores (tels que des sons percussifs, de flûtes, etc.). Il constitue dès lors un outil très recherché par les compositeurs. Gérard Grisey, Jonathan Harvey et Kaija Saariaho l'ont, entre autres, utilisé.

En produisant en 1982 une surprenante version « synthétique » du fameux air de la Reine de la Nuit de la *Flûte enchantée* de Mozart, Yves Potard a démontré qu'un synthétiseur enrichi par des règles de contrôle était capable de produire des sons de synthèse d'une qualité impressionnante.

La parole visible

Inventé dans les années 40 aux États-Unis, le spectrographe acoustique donne un diagramme qui

représente l'image d'un son dans les dimensions temps/fréquence/amplitude. Outil de base pour l'analyse de la structure acoustique de la parole, le spectrographe est aujourd'hui calculé par ordinateur.

A partir d'un spectrogramme, le phonéticien parvient à lire ce qui a été prononcé. Le spectrographe peut être interprété comme la visualisation de l'amplitude spectrale d'une analyse de Fourier à court terme. C'est donc la partie visible du vocodeur de phase. Cela signifie qu'il est possible de modifier le diagramme spectrographique et de resynthétiser le signal ainsi modifié.

Sur le spectrogramme présenté ici (figure ci-contre), on reconnaît :

(A) Trois battements des cordes vocales. La fréquence de ces battements donne la hauteur de la voix.

(B) Le silence de tenue d'une plosive (p).

(C) L'explosion du relâchement d'une plosive (t).

(D) Un son fricatif : la source sonore est un bruit dû à la turbulence du jet d'air à une constriction du conduit vocal.

(E, F, G) Les trois premiers formants pour la voyelle de « lait ».

Les vocodeurs

Les techniques numériques ont considérablement accru la puissance des dispositifs de transmission et de stockage de la voix. Le téléphone ou le magnétophone transmettent et enregistrent directement un signal électrique analogue à la variation de pression acoustique, au son. A l'inverse, les codeurs vocaux (ou vocodeurs) sont des dispositifs d'analyse/synthèse.

Le signal vocal est d'abord analysé et décomposé en paramètres d'un modèle, tel que, par exemple, le modèle source/filtre. Ces paramètres sont ensuite stockés ou transmis. Le signal vocal est alors reconstitué par synthèse, c'est-à-dire calculé à partir des paramètres d'analyse et d'un modèle de synthèse.

Outre les applications au codage de la parole, les vocodeurs permettent également de modifier certains aspects de la parole codée. En tant que dispositifs d'analyse, de modification et de synthèse, les vocodeurs sont de plus en plus employés en synthèse vocale, dans le cadre des synthétiseurs par échantillonnage et concaténation.