

d'un changement de sexe ou d'âge. La reproduction des spécificités vocales des personnes âgées, telles que le chevrottement de la voix, nécessite d'affiner les traitements numériques réalisés. L'âge est généralement marqué par des dysfonctionnements dans la mécanique de l'appareil vocal : les plis vocaux se distendent, ce qui a pour effet d'augmenter le souffle et de faire apparaître un déraillement dans la voix, le chevrottement.

Plus surprenant, il est possible de créer des voix non humaines, voire fantastiques, par exemple faire parler une personne avec une voix de lion. Cette transformation est

réalisée à partir d'un enregistrement de voix humaine et d'un rugissement de lion, en substituant l'empreinte sonore des plis vocaux d'un lion à celle des plis vocaux de la voix humaine, tout en conservant le corps de résonance humain. Le résultat est stupéfiant : on entend une voix humaine dont on comprend clairement le sens, mais avec le souffle et les rugissements d'un lion !

Les ingénieurs de production dans le monde du cinéma et des films d'animation utilisent couramment ces traitements numériques pour modifier la voix d'un acteur ou créer des personnages imaginaires.

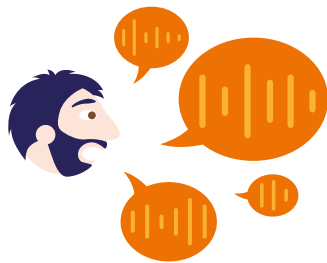
L'étape suivante, aujourd'hui devenue réalité, consiste à reproduire la voix d'acteurs ou de personnalités historiques. Ainsi, à l'Ircam, nous sommes aujourd'hui régulièrement sollicités pour ressusciter des voix à partir d'archives sonores.

Parler avec la voix d'un autre

La prosodie et le timbre font l'identité d'une voix. Mais quelle est la contribution de ces deux caractéristiques sonores lors d'une imitation ? Lorsqu'une personne A (Alice)

TRANSFORMATION DE L'IDENTITÉ DE LA VOIX

Les techniques actuelles permettent de transformer la voix d'une personne en celle d'une personne célèbre, d'une personne plus jeune, plus âgée ou même en celle d'un monstre. Dans l'exemple ci-dessous, la voix d'Alice est transformée en celle de Bernard. Les barres symbolisent la prosodie et le fond coloré des bulles, le timbre.



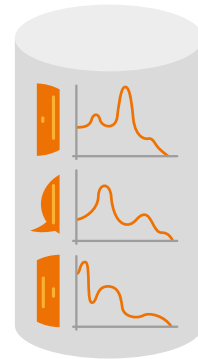
1. ENREGISTREMENT DE BERNARD

Cette étape consiste à récupérer plusieurs heures de paroles prononcées par Bernard.



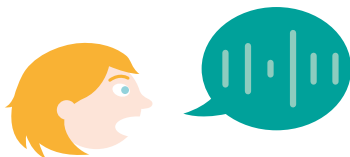
2. DÉCOUPAGE EN PHONÈMES

Un logiciel analyse les paroles de Bernard, isole les phonèmes et détermine le filtre vocal associé au timbre de chaque phonème.



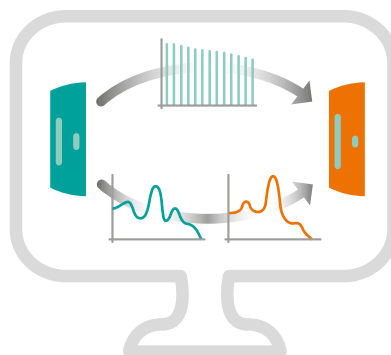
3. CONSTITUTION D'UNE BASE DE DONNÉES

Les phonèmes sont classés avec le filtre vocal associé.



4. ENREGISTREMENT D'ALICE

Alice dit le texte à transformer. Si l'objectif est d'obtenir la voix de Bernard, elle doit imiter sa prosodie. En revanche, elle ne peut imiter parfaitement son timbre.



5. ANALYSE ET TRANSFORMATION

Chaque phonème d'Alice est isolé et analysé. La prosodie est conservée tandis que le filtre vocal associé au timbre d'Alice est remplacé par celui de Bernard.



6. ASSEMBLAGE DES PHONÈMES

La voix obtenue conserve la prosodie d'Alice, mais avec le timbre de Bernard.

cherche à imiter la voix d'une personne B (Bernard), elle joue principalement sur sa prosodie pour reproduire celle de B et ses spécificités. En revanche, il lui est plus difficile de contrefaire son timbre : les dimensions du conduit vocal étant déterminées par la physiologie, il est propre à chaque individu. Ainsi, même les imitateurs les plus doués ne sont pas en mesure de reproduire le timbre d'une voix avec précision et se limitent généralement à une imitation de la prosodie.

Pour imiter vocalement et de façon réaliste une personne, il est essentiel de

reproduire aussi son timbre. Pour ce faire, la conversion d'identité de la voix consiste à créer une collection de « masques vocaux » de Bernard (la personne à imiter) à partir d'enregistrements de sa voix, puis d'appliquer ces masques vocaux sur la voix d'Alice. Nous devons ainsi collecter des enregistrements sonores de la personne à imiter pour créer une base de données de la voix à reproduire. Ces données sont ensuite découpées en phonèmes et analysées pour créer une signature acoustique de la voix à reproduire. Plus précisément, pour chaque phonème, on détermine un « masque vocal »,

c'est-à-dire le filtre vocal qui spécifie la répartition des amplitudes selon les fréquences du son (voir la figure page 57). La conversion d'identité est alors un jeu de copier-coller des masques vocaux créés. Pour chaque phonème émis par Alice, il faut sélectionner le filtre du phonème correspondant dans la base de données de la voix de Bernard, puis le substituer au filtre d'Alice.

Un algorithme sélectionne dans la base de données de la voix d'Alice la séquence de filtres vocaux qui garantira la meilleure conversion d'identité de la voix. La voix a alors le timbre de la personne à imiter...

CRÉATION D'UNE VOIX DE SYNTHÈSE

Les voix de synthèse existaient déjà, mais elles étaient mécaniques et peu naturelles. Grâce aux progrès récents, elles deviennent naturelles et expressives.



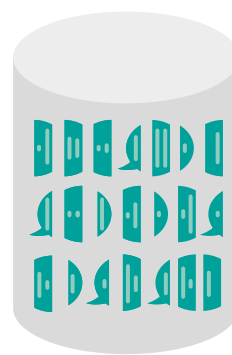
1. ENREGISTREMENT D'ALICE

La voix d'Alice est enregistrée pendant quelques heures.



2. DÉCOUPAGE EN UNITÉS

Un logiciel analyse les paroles d'Alice et les découpe en phonèmes, syllabes, mots, etc., et analyse la prosodie et le timbre.



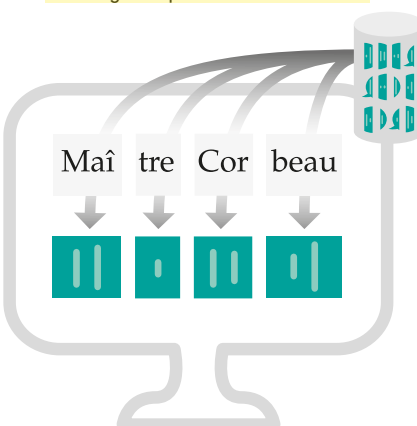
3. CONSTITUTION D'UNE BASE DE DONNÉES

Plus la base est importante, plus la voix créée sera naturelle et expressive.

Maître Corbeau,
sur un arbre perché,
Tenait en son bec
un fromage.
Maître Renard,
par l'odeur alléché,
Lui tint à peu près
ce langage :

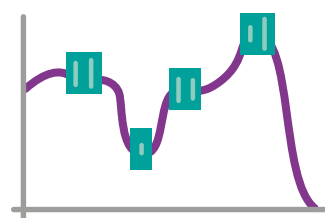
4. TEXTE À VOCALISER

Le texte est analysé et converti en une séquence de phonèmes, syllabes, etc.



5. SÉLECTION DES UNITÉS

S'il existe plusieurs versions d'une unité dans la base, on sélectionne celui dont le timbre et la prosodie s'adaptent le mieux au contexte dans la phrase.



6. CONCATÉNATION

Un algorithme retouche localement le signal afin de lisser la transition entre les unités et gomme les imperfections de la voix de synthèse.

Cette technique a été utilisée en première mondiale pour reconstruire les voix de Marilyn Monroe (dans le film *Marilyn*, de Philippe Parreno, 2012), du maréchal Pétain (dans le documentaire *Juger Pétain*, de Philippe Saada, 2014), et de Louis de Funès (dans le film *Pourquoi j'ai pas mangé mon père*, de Jamel Debbouze, sorti en 2015). Un comédien dont la mission était d'imiter la prosodie et le jeu de Louis de Funès sur le texte du personnage dans le film a été enregistré en plusieurs séances. Parallèlement, une base de données d'une dizaine de minutes de la voix de Louis de Funès a été constituée à partir d'une collection d'enregistrements historiques.

Les chercheurs ont alors pu convertir le timbre du comédien à partir de la signature vocale de Louis de Funès. L'effet est saisissant : Louis de Funès semble avoir été enregistré aujourd'hui, mais avec sa voix des films des années 1970.

Un puzzle vocal

Les chercheurs de l'Ircam repoussent encore les limites de la manipulation vocale grâce à la synthèse de la parole. Ici, il n'est plus question de transformer une voix à partir d'un enregistrement, mais bien de créer une voix de synthèse capable d'exprimer n'importe quel texte (voir l'encadré page 59).

Les premières tentatives de synthèse de la parole remontent au XVIII^e siècle, le siècle des automates et des machines parlantes. Avec l'avènement de l'électricité, au XX^e siècle, les machines se modernisent et se perfectionnent (tel le « Voder », pour *Voice Operation Demonstrator*, fabriqué en 1939 par les laboratoires Bell), puis s'automatisent avec l'arrivée des ordinateurs. Ainsi, le synthétiseur de l'IBM 704, conçu par les laboratoires Bell en 1962, est connu pour avoir inspiré la voix de l'intelligence artificielle HAL dans le film 2001, *l'Odyssée de l'espace* de Stanley Kubrick.

Aujourd'hui, les systèmes de synthèse de la parole sont totalement automatiques et offrent la possibilité de personnaliser les voix numériques. La synthèse de la parole repose essentiellement sur l'utilisation de bases de données de voix, et plusieurs heures d'enregistrement sont nécessaires pour créer une voix de synthèse. Car contrairement à la conversion de voix, la synthèse ne se limite pas à modifier les caractéristiques d'une voix existante. Il faut avoir à disposition toutes les briques élémentaires pour reconstituer intégralement le langage et prononcer n'importe quel texte. La plupart des synthétiseurs actuels fonctionnent par « sélection d'unités » dont le principe est similaire à un jeu de puzzle.

Chaque pièce est un morceau de voix (phonème, syllabe, mot, etc.) dont la prosodie et le timbre lui sont spécifiques. La base de données sert alors de réservoir de pièces de puzzle qui doit couvrir au mieux la variabilité acoustique de la voix à créer. Pour synthétiser un texte, il faut alors chercher et réassembler les pièces destinées à reconstituer le texte. La construction du puzzle relève cependant du casse-tête : des algorithmes trient une immense masse de pièces de puzzle, chacune ayant une signature sonore spécifique qui n'autorise pas à les accoler simplement les unes aux autres. La réalisation du puzzle textuel nécessite de trouver les pièces qui, mises bout à bout, se correspondent au mieux.

Afin qu'une voix de synthèse paraisse naturelle, il est indispensable de s'assurer de l'intelligibilité du texte en associant parfaitement les phonèmes. Il faut également s'assurer de la forme d'ensemble des pièces du puzzle, la prosodie. La qualité de la restitution dépend de la taille des bases de données : le grand nombre de pièces sonores disponibles pour chaque fragment de voix garantit la bonne connexion des

Un interprète dans votre poche

De voyage au Japon, vous souhaitez converser avec un habitant. Vous dégainez votre téléphone, parlez en français dans le microphone et votre voix ressort du haut-parleur quasi instantanément en japonais ! Votre interlocuteur fait de même en vous répondant dans sa propre langue traduite en français par votre téléphone.

Ce scénario est d'ores et déjà possible grâce aux logiciels d'interprétation qui ont vu le jour en 2009 au sein du projet européen EMIME (www.emime.org) et développés parallèlement par IBM, Google ou Mobile Technologies avec l'application Jibbigo.

Leur fonctionnement repose sur une chaîne d'actions. Le message vocal capté par le micro est tout d'abord retranscrit sous forme de texte par un module de transcription. Le texte est alors traduit vers la langue cible par un module de traduction avant d'être prononcé par un module de synthèse et finalement émis par le haut-parleur du téléphone. Le module de traduction réalise la traduction à partir de règles, de la même manière que le ferait un interprète humain. Cependant, ces règles sont difficiles à formuler

explicitement. Ainsi, plutôt que de fournir au module un ensemble limité de règles à appliquer, celles-ci sont apprises par un ordinateur à partir d'un ensemble composé d'une multitude d'exemples de textes écrits dans les deux langues. De la même façon que Champollion eut appris à décrypter les hiéroglyphes, en partie grâce à la pierre de Rosette sur laquelle était gravé un même texte écrit en trois langues, l'ordinateur va par lui-même déduire les méthodes pour passer d'une langue à l'autre. Contrairement à la pierre de Rosette, la quantité d'exemples disponibles pour l'apprentissage de ces règles, collectés entre autres sur Internet, est quasi illimitée et grandit de jour en jour.

De même, les modules de transcription et de synthèse sont en mesure de retranscrire la parole

en texte et de le prononcer au moyen de méthodes apprises à partir de livres audio par exemple, mais ce n'est pas tout. Lorsque nous rencontrons une personne pour la première fois, si elle a un fort accent et si l'environnement est bruyant, la compréhension est difficile dans les premières minutes de discussion, mais s'améliore rapidement. Ainsi, à partir de quelques secondes de parole, le module de transcription s'ajuste à la voix de l'utilisateur, à sa manière de parler, mais également à l'environnement acoustique (lieu calme ou bruyant) afin d'améliorer ses performances. Cette capacité d'adaptation est également partagée par le module de synthèse qui produit de la parole à partir d'un texte avec une voix ayant des caractéristiques acoustiques semblables à celle de l'utilisateur. Nous sommes dorénavant doués pour parler le japonais (ou toute autre langue) sans l'avoir jamais appris...

– Pierre Lanchantin

Machine
Intelligence Laboratory,
université de Cambridge