Transformation et synthèse de la voix parlée et de la voix chantée

par Xavier Rodet[1]

La mission principale de l'Institut de recherche et coordination acoustique/musique (Ircam) est la création musicale et la création artistique en général, ce qui inclut notamment les arts du spectacle comme le théâtre ou le film. Cet institut possède une longue expérience dans l'analyse et la synthèse des sons, et en particulier de la parole. En effet, de nombreux compositeurs contemporains portent un vif intérêt à la voix, chantée mais aussi parlée. Ils considèrent la voix non seulement comme un matériau musical qui peut entrer, d'une façon ou d'une autre, dans leurs compositions, mais aussi pour sa structure, depuis les niveaux acoustiques et phonétiques jusqu'aux niveaux linguistiques les plus élevés.

Dans ce contexte, l'équipe « Analyse-synthèse » des sons de l'Ircam a développé depuis plusieurs années un savoir-faire, des études et des outils, en particulier informatiques, concernant l'analyse, le traitement et la synthèse de la voix et de la parole. Ces moyens sont d'abord utilisés pour la création musicale à l'Ircam. Ils ont été employés, par exemple, pour des pièces récentes de Jean-Baptiste Barrière, Joshua Fineberg, Stefano Gervasoni ou Jonathan Harvey. Mais ces moyens trouvent également des applications dans le multimédia en général. En effet, alors que les images de synthèse ont envahi de nombreux médias, dessins animés, jeux vidéo et films notamment, la voix reste aujourd'hui le parent pauvre en la matière : elle est, la plupart du temps, simplement enregistrée par des acteurs, souvent synchronisée de façon « manuelle » avec le mouvement des personnages et n'utilise presque aucune technique de synthèse, sauf à de rares

1. Avec Grégory Beller, Niels Bogaards, Gilles Degottex, Snorre Farner, Pierre Lanchantin, Nicolas Obin, Axel Roebel, Christophe Veaux et Fernando Villavicencio.

exceptions. Cependant, les méthodes et outils développés par l'Ircam ont permis de créer la voix du castrat dans le film *Farinelli* de Gérard Corbiaud, d'améliorer la prononciation anglaise de Gérard Depardieu pour le film *Vatel* de Roland Joffé, ou de transformer une voix de femme en voix d'homme pour le film *Tirésia* de Bertrand Bonello, et, inversement, une voix d'homme en voix de femme pour le film *Les Amours d'Astrée et de Céladon* d'Éric Rohmer (2007). Bien d'autres applications sont expérimentées pour les jeux vidéos, le dessin animé, les avatars, etc.

Les principaux sujets sur la voix traités à l'Ircam sont la constitution de corpus oraux, l'analyse de ces corpus, la synthèse à partir du texte, la transformation du type et de la nature de la voix, la conversion d'identité de la voix, l'étude de la transformation d'expressivité de la voix, la séparation de la source glottique de l'influence du conduit vocal, et la modélisation de la prosodie dans différents modes de discours. Ces divers travaux ont pour but premier de fournir de nouveaux moyens aux compositeurs et artistes travaillant à, ou avec, l'Ircam. Pour cela, l'institut collabore avec de nombreux centres de recherches et mène des projets de recherche dans les cadres institutionnels français (agence nationale de la recherche, CNRS), européens ou autres. Enfin l'Ircam valorise ses compétences, connaissances et résultats vers d'autres instituts et vers l'industrie.

Dans la première section de ce chapitre, j'exposerai la problématique et les moyens nécessaires à la gestion de corpus de parole enregistrée. En effet, les méthodes scientifiques et techniques d'étude de l'oral s'appuient de plus en plus sur l'analyse statistique de grands corpus enregistrés, qui requièrent donc un outil spécifique. Dans la deuxième section, j'exposerai cet outil : la plate-forme logicielle IrcamCorpusTools, développée par l'équipe « Analyse-synthèse » des sons de l'Ircam pour la gestion de corpus de parole enregistrée. Dans la troisième section seront présentés les principaux logiciels développés par l'équipe pour l'analyse, la synthèse et la transformation de voix (SuperVP et AudioSculpt). Enfin, je conclurai en décrivant quelques applications d'analyse, de synthèse et de transformation de voix, aussi bien dans le domaine de la recherche que dans la création musicale ou le multimédia.

Gestion de corpus de parole enregistrée

LES MÉTHODES À BASE DE CORPUS

Les méthodes à base de corpus sont désormais très largement répandues en traitement de la parole et en traitement du langage pour le développement de modèles théoriques et d'applications technologiques. Que ce soit pour vérifier des heuristiques, découvrir des tendances ou modéliser des données, l'introduction de traitements calculatoires et/ou statistiques fondés sur les données des corpus a multiplié les possibilités et permis des avancées considérables dans les technologies de la parole et du langage. La reconnaissance et la synthèse de parole en sont des exemples pour le traitement automatique de la parole. De même, l'utilisation de corpus annotés (annotations d'ordre phonétique, prosodique, des phénomènes paraverbaux et des disfluences, par exemple du corpus LeaP[2], mais aussi d'ordre syntaxique et discursif) intéresse la recherche en linguistique aussi bien qu'en traitement de la parole. Toutefois, cette complémentarité n'est possible que par la mise en commun des corpus. C'est pourquoi les questions de représentation et de gestion des données des corpus sont centrales. Les corpus oraux sont constitués de deux types principaux de ressources : les signaux temporels et les annotations. Les signaux temporels sont les enregistrements audio, vidéo et/ou physiologiques, ainsi que leurs descriptions (fréquence fondamentale, spectrogramme, etc.). Les annotations sont la transcription textuelle ainsi que toutes les notations ajoutées manuellement ou automatiquement qui permettent de caractériser d'un point de vue linguistique le signal acoustique (transcription phonétique, catégories grammaticales, structure du discours, etc.). Les différents niveaux d'annotations possèdent généralement des relations hiérarchiques et/ou séquentielles et sont synchronisés temporellement sur le signal acoustique. Les outils de gestion des corpus recouvrent tout un ensemble de fonctionnalités, allant de la création et de la synchronisation des ressources aux requêtes (pouvant porter autant sur les annotations que sur les signaux temporels), en passant par le stockage et l'accès aux données. La plupart des systèmes de gestion de corpus existants ont été développés pour des corpus spécifiques et sont difficilement adaptables et extensibles[3].

2. Learning Prosody Project : http://leap.lili.uni-bielefeld.de
3. Oostdijk (2000).

Des efforts ont été faits pour faciliter l'échange de données par la conversion de formats[4] ou pour dégager une représentation formelle pouvant servir d'interface commune entre les divers outils et les données[5]. Cette notion d'interface entre les méthodes et les données est à la base de la plate-forme IrcamCorpusTools présentée dans ce chapitre. Cette plate-forme utilise l'environnement de programmation Matlab afin d'être facilement extensible. Elle permet notamment la synchronisation d'informations provenant de différentes sources (vidéo, audio, symbolique, etc.) ainsi que la gestion de nombreux formats (XML, AVI, WAV, SDIF[6], etc.). Elle est munie d'un langage de requête prenant en compte les relations hiérarchiques multiples, les relations séquentielles et les contraintes acoustiques. Elle permet ainsi l'analyse contextuelle de variables acoustiques (prosodie, enveloppe spectrale) en fonction de variables linguistiques (mots, groupe de sens, syntaxe). Elle est employée pour la synthèse de la parole par sélection d'unités, les analyses prosodiques et phonétiques contextuelles, la modélisation de l'expressivité, et pour exploiter divers corpus de parole en français et en d'autres langues.

SYSTÈMES DE GESTION ET DE CRÉATION DE CORPUS DE PAROLE

Depuis l'essor de la linguistique de corpus[7], de nombreux corpus annotés ont été exploités par le traitement automatique des langues, dont des corpus oraux comme ceux qui sont recensés par LDC[8]. L'automatisation de ce traitement nécessite de traiter une grande quantité de métadonnées linguistiques. Aussi, de nombreux systèmes de gestion de larges corpus sont aujourd'hui disponibles pour cette communauté[9]. Dans le domaine du traitement automatique de la parole, le corpus TIMIT fut le premier corpus annoté à être largement diffusé. Une tendance actuelle est l'utilisation de corpus multimodaux avec l'intégration de données visuelles, ce qui accroît encore la diversité des formats à gérer. Permettre à une communauté de chercheurs de partager et d'exploiter de tels corpus ne pose pas simplement la question de la gestion des formats, mais aussi celles de la représentation des données, du partage des outils de génération, d'accès et d'exploitation, et du langage de requête associé.

4. Gut *et al.* (2004).
5. Bird *et al.* (2000).
6. http://sdif.sourceforge.net/
7. Chafe (1992).
8. *Linguistic Data Consortium* : http://www.ldc.upenn.edu/Catalog/
9. Cunningham *et al.* (2002).

MODÈLES DE REPRÉSENTATION DES DONNÉES

Un modèle de représentation des données doit pouvoir capturer les caractéristiques importantes de celles-ci et les rendre facilement accessibles aux méthodes utilisées pour leur traitement. Ce modèle constitue en fait une hypothèse sous-jacente sur la nature des données et sur leur structure. Il doit donc être aussi général que possible afin de pouvoir représenter différents types de structures phonologiques et de permettre une grande variété de requêtes sur ses structures. Les modèles principalement utilisés pour le traitement automatique du langage sont des structures hiérarchiques comme celles du Penn Treebank[10], qui peuvent être alignées temporellement dans le cas des corpus oraux. Certains systèmes, comme Festival[11] ou EMU[12], vont au-delà de ces modèles en arbre unique et supportent des hiérarchies multiples, c'est-à-dire qu'un élément peut avoir des parents dans deux hiérarchies distinctes sans que ces éléments parents soient reliés entre eux. Ces représentations sont particulièrement adaptées pour les requêtes multiniveaux sur les données du corpus. D'autres approches, telles que celle de Bird et Liberman[13] ou de Müller[14], se concentrent sur des représentations des données qui facilitent la manipulation et le partage des corpus multiniveaux. Il s'agit généralement de représentations « à plat » des données qui donnent uniquement la structure temporelle : les relations hiérarchiques y sont représentées implicitement par la relation d'inclusion entre les marques temporelles. Enfin, Gut expose une méthode et des spécifications minimales permettant de convertir entre elles les différentes représentations des données utilisées par les corpus[15].

PARTAGE DES DONNÉES

Afin de pouvoir partager les corpus, comme dans le cas du projet PFC[16], des efforts de standardisation ont été entrepris à différents niveaux. Un premier niveau de standardisation consiste à établir des conventions sur les formats de fichiers et les métadonnées décrivant leur contenu. Ainsi, le format XML[17] s'est de plus en plus imposé comme le

10. *Penn Treebank* : http://www.cis.upenn.edu/treebank/home.html/
11. Taylor *et al.* (2001).
12. Cassidy et Harrington (2001).
13. Bird et Liberman (2001).
14. Müller (2005).
15. Gut *et al.* (2004).
16. PFC : Phonologie du français contemporain : http://www.projet-pfc.net/ ; *cf.* Durand et Tarrier (2006).
17. XML : eXtensible Markup Language : http://www.w3.org/XML/

format d'échange des annotations. Cette solution permet la compréhension des données par tous les utilisateurs, tout en leur permettant de créer de nouveaux types de données selon leurs besoins. Un second niveau consiste à standardiser le processus de génération des données elles-mêmes. Cela conduit par exemple à des recommandations comme celles de la Text Encoding Initiative[18] pour les annotations des corpus oraux. Certains projets, tel CHILDES pour l'analyse des situations de dialogues[19], proposent à la fois des normes de transcription et les outils conçus pour analyser les fichiers transcrits selon ces normes.

PARTAGE DES OUTILS

Des efforts ont également été entrepris pour créer des outils libres adaptés aux annotations des ressources audio et/ou vidéo des corpus comme Transcriber[20] ou ELAN du projet DOBES[21]. Toujours pour l'annotation, des outils de visualisation et d'analyse acoustique sont disponibles et largement utilisés, comme WaveSurfer[22] ou Praat[23]. Ces logiciels permettent l'analyse, la visualisation/annotation, la transformation et la synthèse de la parole. Mais ils sont limités, soit par un format propriétaire pour les données, soit par le langage de requêtes, soit pour la gestion des données.

LANGAGES DE REQUÊTE

Pour être exploitable par une large communauté d'utilisateurs, un corpus doit être muni d'un langage de requête qui soit à la fois simple et suffisamment expressif pour formuler des requêtes variées[24]. On peut distinguer deux grandes familles de systèmes utilisés pour stocker et rechercher de l'information structurée : les bases de données et les langages de balises de textes comme le XML[25]. Les langages de requête comme XSLT/XPath sont naturellement adaptés à la formulation des contraintes d'ordre hiérarchique, mais la syntaxe des requêtes se complique lorsqu'il s'agit d'exprimer des contraintes séquentielles. Les systèmes fondés sur le XML offrent une « extensibilité » limitée car ils nécessitent une recherche linéaire dans le sys-

18. Text Encoding Initiative : http://www.tei-c.org/
19. MacWhinney (2000).
20. Transcriber : http://trans.sourceforge.net/en/presentation.php ; *cf.* Barras *et al.* (1998).
21. DOBES : documentation sur les langues rares : http://www.mpi.nl/DOBES/
22. WaveSurfer : http://www.speech.kth.se/wavesurfer/ ; *cf.* Sjölander et Beskow (2000).
23. Praat : http://www.fon.hum.uva.nl/praat/ ; *cf.* Boersma (2001).
24. Lai et Bird (2004).
25. Gut *et al.* (2004), Cassidy et Harrington (2001).

tème de fichiers[26]. À l'inverse, les systèmes de bases de données sont capables de stocker de très grandes quantités d'informations et d'effectuer des requêtes relativement rapides sur celles-ci. Cependant, le modèle relationnel étant par nature moins adapté à la représentation des contraintes hiérarchiques et séquentielles que le XML, une requête donnée en XML se traduit de manière beaucoup plus complexe en SQL[27]. Si des langages intermédiaires plus simples comme LQL ont été proposés[28], les requêtes les plus complexes ne sont pas toujours formulables selon cette approche.

EXPLOITATION DES DONNÉES

Une fonctionnalité essentielle des plate-formes de gestion de corpus est la possibilité d'interfacer les données (éventuellement après filtrage par des requêtes) avec des outils de modélisation. Ainsi, alors que certains environnements de développement linguistique permettent de construire, de tester et de gérer des descriptions formalisées[29], d'autres se sont tournés vers les traitements statistiques[30]. L'apprentissage automatique pour les tâches de classification, de régression et d'estimation de densités de probabilités est aujourd'hui largement employé. Qu'elles soient déterministes ou probabilistes, ces méthodes nécessitent des accès directs aux données et à leurs descriptions. C'est pourquoi certains systèmes de gestion de corpus tentent de faciliter la communication entre leurs données et les machines d'apprentissage et d'inférence de règles, comme c'est le cas pour le projet EMU et le projet R[31].

IrcamCorpusTools : une plate-forme complète de gestion de corpus

Comme nous venons de le voir, si certains outils comme Praat apportent des solutions partielles permettant l'exploitation des corpus, peu de systèmes proposent une solution complète allant de la génération des données jusqu'aux requêtes sur celles-ci. Lorsque de tels systèmes existent, ils ont été le plus souvent conçus au départ pour une application

26. Cassidy et Harrington (2001).
27. Structured Query Language (Langage structuré de requête) ; cf. http://www.sql.org/
28. Nakov *et al.* (2005) ; sur LQL, cf. http://biotext.berkeley.edu/lql/
29. Bilhaut et Widlöcher (2006).
30. Cassidy et Harrington (2001).
31. *R Project* : http://www.r-project.org/

spécifique comme la synthèse de parole[32] ou l'observation de pathologies, comme c'est le cas pour le projet CSL (Computerized Speech Lab). Cela implique des limitations intrinsèques sur le type de données, sur leur représentation et donc sur leur capacité à être partagées. Ainsi le chercheur à la frontière des traitements automatiques du langage et de la parole est-il, pour le moment, contraint d'utiliser une batterie d'outils dédiés et fondés sur plusieurs langages de programmation, ce qui l'oblige à effectuer de nombreuses conversions de formats et interdit toute automatisation complète d'un processus.

LA PLATE-FORME IRCAMCORPUSTOOLS

Pour répondre aux besoins spécifiques de la parole, de son traitement et de l'analyse de corpus, la plate-forme IrcamCorpusTools (ICT) a été développée et est utilisée dans une grande variété d'applications. Elle s'inscrit à l'intersection de deux domaines de recherches complémentaires : la recherche linguistique et le développement de technologies vocales. Nous la présentons dans cette section en commençant par une vue générale du système et de son architecture. Puis nous présentons deux spécificités de la plate-forme : son langage de requête, qui prend simultanément en compte des contraintes d'ordre linguistique et des contraintes sur les signaux ; et le principe d'autodescription des données et des outils, qui permet de répondre à certaines des problématiques concernant les systèmes de gestion et de création de corpus de parole.

ARCHITECTURE DE LA PLATE-FORME

Afin de répondre à différentes demandes de recherche et de développement industriel, l'architecture d'origine[33] s'est naturellement orientée vers une solution extensible, modulaire et partagée par plusieurs utilisateurs et développeurs[34]. Cette mutualisation des outils et des données implique une certaine modularité tout en maintenant des contraintes de standardisation qui assurent la cohérence du système. La solution choisie repose sur le principe d'autodescription des données, et des outils permettant de définir une interface commune entre ces objets. Une vue générale de l'architecture de IrcamCorpusTools (ICT) est offerte par la figure 1. Elle fait apparaître la couche d'interface que nous introduisons entre les données et les outils, et qui est constituée par notre environnement Matlab. Cette architecture à trois niveaux est semblable à celle proposée

32. Taylor *et al.* (2001).
33. Beller et Marty (2006).
34. Veaux *et al.* (2008).

pour le système ATLAS[35]. Elle permet à différentes applications externes ou internes de manipuler et d'échanger entre elles des informations sur les données du corpus. Les différents éléments composant IrcamCorpusTools sont des instances (objets) de classes qui forment le cœur de la plate-forme. Ces classes sont représentées dans la figure 2 et nous les présentons maintenant en détail.

DESCRIPTEURS

L'activité de parole est intrinsèquement multimodale. La coexistence du texte, de la voix et de gestes (faciaux, articulatoires, etc.) génère une forte hétérogénéité des données relatives à la parole. Le système doit être capable de gérer ces données de différentes natures. Voici les types de données gérées par IrcamCorpusTools.

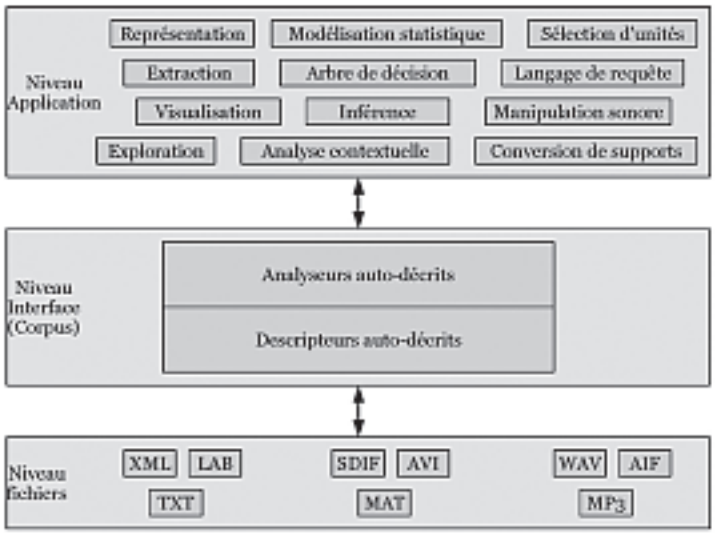

Figure 1 : *Vue d'ensemble de la plate-forme IrcamCorpusTools.*

Informations de type signal

Les signaux correspondent soit aux enregistrements provenant d'un microphone ou d'autres instruments de mesure (EGG, fMRI, ultrasons ou autres), soit à des résultats d'analyse de ces enregistrements. Ils peuvent être unidimensionnels ou multidimensionnels. Parmi les signaux les plus courants figurent ceux qui sont relatifs à la prosodie, comme la fréquence fondamentale f0, l'énergie, le débit de parole, le degré d'articula-

35. Bird et Liberman (2001).

tion mesuré à partir des formants (fréquence, amplitude, largeur de bande), et la qualité vocale (coefficient de relaxation, modèle LF, mesure du voisement), mais aussi ceux qui sont relatifs à l'enveloppe spectrale, donnés par différents estimateurs (FFT, MFCC, TrueEnvelope, LPC) et représentables sous la forme de coefficients autorégressifs (AR), de paires de lignes spectrales (LSF), de pôles, ou d'aires de sections du conduit vocal (LAR). Enfin, cette liste non exhaustive peut être augmentée de signaux issus d'autres modalités comme c'est le cas par exemple pour la mesure de l'aire glottique par caméra ultrarapide (voir, dans les exemples d'application d'analyse acoustique, l'étude de la qualité vocale).

Informations de type métadonnée

Ces informations peuvent, par exemple, servir à spécifier un contexte d'enregistrement (lieu, date, locuteur, consigne donnée, expressivité, genre de discours, etc.). Elles comprennent les transcriptions textuelles *a priori* (parole lue) ou *a posteriori* (parole spontanée). Elles permettent de définir n'importe quelle information sous la forme de mots/symboles ou de séquence de mots.

Informations de type annotation

Ces informations sont de nature textuelle et possèdent, de plus, un temps de début et un temps de fin, permettant d'attribuer une information de type linguistique à une portion de signal. Cette sorte de donnée est cruciale pour une plate-forme de gestion de corpus de parole, puisqu'elle permet de faire le lien entre les signaux et les catégories linguistiques, entre la physique (flux de parole continu) et le symbolique (unités de sens discrètes). Elles constituent souvent des dictionnaires clos, comme c'est le cas pour les phonèmes d'une langue ou pour d'autres étiquettes phonologiques (onset, nucleus, coda, etc.). Parmi ces informations, les segmentations phonétiques sont les plus courantes. Les annotations syntaxiques, de phénomènes prosodiques ou de mots, sont autant d'étiquettes qui peuvent être placées manuellement et/ou automatiquement. Elles définissent alors des segments, aussi appelés *unités*, dont la durée est variable : senone, semiphone, phone, diphone, triphone, syllabe, groupe accentuel, mot, groupe prosodique, phrase, paragraphe, discours, etc.

Informations de type statistique

Sur l'horizon temporel de chacune des unités, les signaux continus peuvent être modélisés par des valeurs statistiques. Ces valeurs, décrivant le comportement d'un signal sur cette unité, sont appelées *valeurs caractéristiques* : moyenne arithmétique et géométrique, variance, intervalle de

variation, maximum, minimum, moment d'ordre N, valeur médiane, centre de gravité, pente, courbure, etc.

UNITÉS

Les unités sont les objets permettant de relier les données entre elles. Elles sont définies pour chaque niveau d'annotation et regroupent les données symboliques ou acoustiques sur la base de la segmentation temporelle associée à ce niveau d'annotation. Les unités sont reliées entre elles par des relations de type séquentiel et/ou hiérarchique. Les relations hiérarchiques sont représentées sous la forme d'arbres (« phrase->mots->syllabes->phones », par exemple) dont les nœuds correspondent chacun à une unité. Afin de représenter des relations hiérarchiques multiples, une liste d'arbres est utilisée à la manière de Festival[36]. Les unités du niveau « phone » sont par exemple dans une relation de parenté avec celles du niveau « syllabe » et avec celles du niveau « mot » ; en revanche, les syllabes et les mots n'ont pas de relation de parenté entre eux (parce que, à cause des liaisons, certaines syllabes ne peuvent être liées de façon unique). Ces arbres permettent de propager les marques temporelles au sein d'une hiérarchie d'unités à partir d'un seul niveau d'annotation synchronisé avec le signal de parole (typiquement le niveau d'annotation issu de la segmentation phonétique). Inversement, à partir d'annotations indépendamment alignées, on peut construire les différentes hiérarchies entre unités, en se basant sur l'intersection des marques temporelles. Cela permet notamment de maintenir la cohérence des diverses données relatives aux unités, tout en autorisant des interventions manuelles à tous les niveaux. À l'inverse des relations hiérarchiques, les relations séquentielles entre unités ne sont définies qu'au sein d'un même niveau d'annotation.

FICHIERS

Nous avons choisi de stocker les différents descripteurs indépendamment les uns des autres afin de faciliter la mise à jour et l'échange des données du corpus[37]. Ces fichiers reposent sur plusieurs supports dont les formats les plus répandus sont :

– LAB, XML, ASCII, TextGRID, pour les données de type métadonnée et annotation ;

– SDIF, AVI, WAV, AIFF, AU, MP3, MIDI, pour les données de type signal ;

36. Taylor *et al.* (2001).
37. Müller (2005).

– MAT (Matlab), pour les données de type relation et statistique.

En revanche, les unités et leurs relations sont stockées dans un fichier unique. Une fonction permet de reconstruire les unités et leurs relations lorsqu'un descripteur (symbolique ou acoustique) a été modifié.

ANALYSEURS

Les analyseurs regroupent toutes les méthodes de génération ou de conversion des données. On peut les enchaîner si nécessaire, par exemple si on veut obtenir la moyenne de la fréquence fondamentale sur le groupe prosodique avoisinant une syllabe[38]. Certaines de ces méthodes sont *internes* au logiciel, d'autres utilisent des logiciels externes qui peuvent être exécutés par appel depuis IrcamCorpusTools. Grâce à l'interface du système de fichiers, les données engendrées par un tel logiciel sont automatiquement rendues accessibles au sein de notre environnement. D'un point de vue utilisateur, le caractère interne/externe ne fait aucune différence. Dans l'exemple cité précédemment, l'utilisateur peut remplacer un estimateur interne de la fréquence fondamentale (à titre d'exemple, par celui de Praat, de WaveSurfer ou de SuperVP[39]) sans avoir à changer d'environnement.

CORPUS

Un corpus peut être représenté comme un ensemble d'énoncés. Chacun de ces énoncés contient un ensemble d'analyses. Chacune de ces analyses comporte un ou plusieurs descripteurs. Par exemple, l'analyse « audio » comporte le descripteur « forme d'onde » qui n'est autre que le signal acoustique de la phrase enregistrée. Ces analyses sont donc synchronisées au niveau de la phrase dans un corpus. Mais une synchronisation plus fine existe aussi grâce à l'ajout d'unités décrites par l'analyse « segmentation ». Les objets « corpus » sont des interfaces avec le système de fichiers. Lorsqu'un analyseur est appliqué à un corpus, celui-ci fait appel à des fichiers d'entrée et de sortie. Il stocke ainsi toute création d'un fichier, qu'il relie à une configuration particulière de l'analyseur et des paramètres qui l'ont généré, et y attache les objets descripteurs. L'objet « corpus » est lui-même stocké dans un fichier XML, à la racine du système de fichier, ce qui permet à plusieurs personnes d'ajouter ou de supprimer des données dans un corpus sans que cela entraîne de conflit. En effet, l'objet « Corpus » conserve au fur et à mesure l'historique

38. Voir l'exemple donné *infra*, p. **NNN**.
39. Bogaards *et al.* (2004).

des opérations effectuées sur un corpus et lui confère donc un accès multi-utilisateur.

LANGAGE DE REQUÊTE

Certains outils de requête XML (Xpath, Xquery, NXT search) présentent une syntaxe complexe. Dans IrcamCorpusTools, nous privilégions l'expressivité du langage de requête. Une requête élémentaire est ainsi constituée :

1) du niveau dans lequel on effectue la recherche d'unité ;
2) d'une relation séquentielle par rapport à l'unité recherchée ;
3) d'une relation hiérarchique par rapport à l'unité recherchée ;
4) d'une condition à tester sur les données numériques associées aux unités.

Ces requêtes sont rapides car elles ne s'appliquent qu'aux données préalablement stockées en mémoire vive. De plus, elles peuvent être composées à volonté, afin de faire des recherches complexes entre les multiples niveaux d'unités.

PRINCIPE D'AUTODESCRIPTION

L'expressivité du langage de requêtes provient de la possibilité de mélanger des contraintes sur des données de types différents. Cela est rendu possible par le principe d'autodescription sur lequel repose IrcamCorpusTools. Chaque instance d'une classe (corpus, fichier, analyseur ou descripteur) est accompagnée de métadonnées décrivant son type, sa provenance, comment y accéder et comment la représenter. Cela permet une compréhension et une exploitation immédiates de tous les objets par tous les utilisateurs, mais aussi par le système lui-même. À l'instar du caractère interne/externe des analyseurs, l'hétérogénéité des données est invisible pour l'utilisateur, qui ne possède qu'un seul lexique restreint de commandes, avec lesquelles il peut rapidement se familiariser. Aucune donnée ne se « perd », car l'objet « corpus » garde une trace des différentes opérations réalisées sur lui et donc, des différentes analyses ayant engendré ses données. Cela permet notamment de conserver un historique de l'accès aux données. En effet, on peut toujours accéder à d'anciennes informations, même si la méthode d'accès à celles-ci a changé entre-temps. Enfin, n'importe quel utilisateur peut comprendre les données des autres et utiliser leurs analyseurs sur ses corpus, sans avoir à changer d'environnement. En résumé, le principe d'autodescription d'IrcamCorpusTools lui assure la pérennité des données, lui fournit un langage de requête expressif et lui confère la possibilité de mutualiser les données, les fichiers, les corpus et, surtout, les analyseurs. La mise en commun des outils est un

facteur déterminant pour le développement des recherches en TAL et en TAP, car leur complexité s'accroît rapidement.

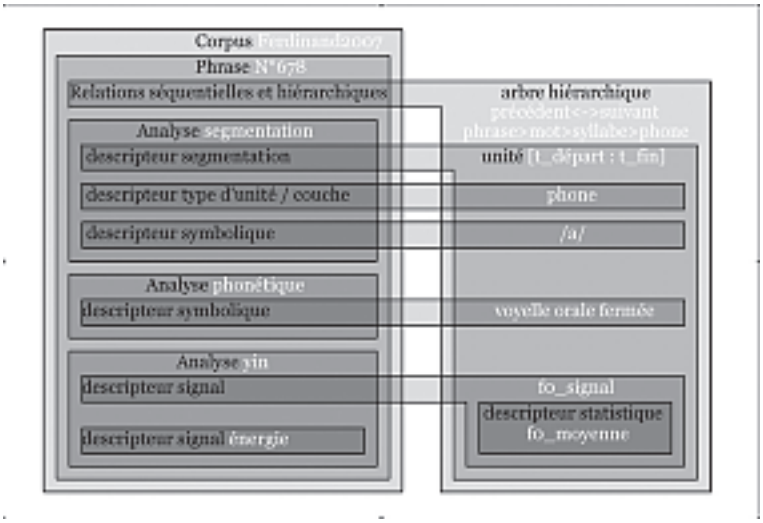

Figure 2 : *Exemple d'utilisation : une instance particulière.*

CRÉATION ET ANALYSE DE CORPUS

Conception de corpus

Si l'approche qui consiste à soumettre des hypothèses théoriques à l'épreuve de grands corpus oraux est de plus en plus répandue, c'est parce que la taille de ces corpus leur permet d'être considérés comme exhaustifs (sous certaines hypothèses)[40]. Pour le reste, l'approche traditionnelle consiste à créer des corpus en vue de valider certaines hypothèses théoriques prises en compte lors de la conception de ces corpus. Il en va de même pour la conception d'un synthétiseur de parole qui débute par une phase de conception de corpus, afin de minimiser les traitements ultérieurs. Nous avons élaboré un ensemble d'outils dans le dessein de sélectionner des ensembles de phrases respectant certaines contraintes linguistiques. Ces ensembles sont extraits de larges corpus textuels, par exemple Corpatext[41] de plus de 37 millions de mots. L'extraction est motivée par différentes recherches de couvertures maximales sous contraintes. Pour la synthèse TTS, l'ensemble des phrases retenues doit présenter le meilleur compromis entre une taille minimale et une couverture maximale des phonèmes par rapport à des contextes donnés (phonétique, lexical, syn-

40. Habert (2000).
41. Corpatext : http://www.lexique.org/public/corpatext.php

taxique, etc.). Ici, une couverture maximale pourra être interprétée comme ayant au moins un candidat pour chaque contexte ou bien comme ayant une distribution statistique des candidats reflétant une distribution naturelle (comme la distribution sur tout le corpus textuel, par exemple).

Décodage acoustico-phonétique

Pour permettre des études en linguistique de corpus, il est nécessaire qu'un certain nombre d'étapes soient automatisées. Dans le cadre de la synthèse de parole, de la modélisation prosodique et expressive, le décodage acoustico-phonétique est une étape essentielle en amont d'une chaîne de traitements linguistiques permettant de représenter la structure de la parole. Cette étape permet la segmentation d'un signal de parole en ses unités linguistiques minimales. Celles-ci sont ensuite regroupées en des unités linguistiques de dimensions supérieures (syllabes, groupes accentuels, groupes prosodiques). Une fois la conception du corpus réalisée (parole de laboratoire ou parole spontanée par exemple), les enregistrements sont automatiquement segmentés en phones à l'aide de l'analyseur IrcamAlign[42]. Ce dernier prend en entrée le signal de parole, sa transcription textuelle, ainsi qu'un dictionnaire constitué de modèles statistiques paramétriques (Hidden Markov Models, HMM[43]) de chacun des phones en contexte, appris sur le corpus multilocuteur BREF80[44]. À partir de la transcription textuelle et du dictionnaire, un modèle statistique de la phrase est constitué en prenant en compte les différentes variantes de prononciation. La meilleure séquence de phones peut alors être sélectionnée, puis alignée sur le signal de parole. Finalement, afin de détecter les erreurs éventuelles et de simplifier une phase de correction manuelle, un indice de confiance est associé automatiquement à chacun des phones segmentés.

Création des unités

Le système IrcamCorpusTools offre une grande modularité dans l'étape de spécification des unités, ce qui permet d'envisager un large champ d'application possible en étude de la parole. Il est ainsi possible de définir arbitrairement une structure de parole (tant au niveau des unités utilisées que de leurs attributs associés) à partir de considérations particulières au domaine d'étude considéré. Cette propriété se révèle nécessaire

42. Lanchantin *et al.* (2008).
43. Rabiner (1989).
44. Lamel *et al.* (1991).

dans l'étude des phénomènes rattachés à la parole, que ce soit pour définir des structures de la parole à partir de théories phonologiques spécifiques au sein d'une langue, pour représenter la variabilité des structures observées entre les langues, ou bien pour définir des niveaux d'analyses supplémentaires pour des domaines d'études spécifiques (acquisition du langage, pathologie, etc.). À partir de la segmentation phonétique présentée précédemment, la représentation de la structure phonologique segmentale et suprasegmentale de la parole dans IrcamCorpusTools est décrite avec les éléments suivants : le phonème et ses attributs phonologiques, la structure syllabique (onset/rhyme, nucleus/coda), la syllabe et ses attributs phonologiques, le groupe accentuel, le groupe prosodique et le discours.

Analyse, synthèse et transformation de voix dans SuperVP et AudioSculpt

L'analyse, la synthèse et la transformation de voix requièrent finalement des logiciels dits de traitement du signal adaptés aux signaux sonores, musicaux, et spécialement au signal de la voix. L'équipe « Analyse-synthèse » des sons de l'Ircam développe en particulier les logiciels SuperVP et AudioSculpt pour la musique mais aussi pour la voix. Le logiciel SuperVP est un « moteur » de traitement du signal qui peut être utilisé seul ou appelé par d'autres logiciels comme AudioSculpt, Xspect ou Max/MSP.

LE LOGICIEL SUPERVP

Le logiciel SuperVP[45] (pour Super Vocoder de Phase[46]) est une implémentation très évoluée d'un vocodeur de phase généralisé, qui est utilisable soit en mode autonome (en ligne de commande) soit comme bibliothèque. Il est appelé comme moteur de calcul pour les traitements et les analyses dans les logiciels AudioSculpt, Xspect ou Max/MSP, et dans un grand nombre de projets de recherche et de production musicale. Il a été perfectionné sur de nombreux points et, en particulier, pour améliorer la qualité des transformations de la parole. Il est aussi utilisé ou distribué sous licence par plusieurs sociétés commerciales.

45. Bogaards *et al.* (2004), Roebel (2003).
46. http://forumnet.Ircam.fr/

SuperVP permet de faire une grande quantité d'analyses, de traitements et de synthèses. Il inclut plusieurs méthodes d'estimation d'enveloppe spectrale (AutoRegressive ou LPC, Cepstre, Cepstre Discret, TrueEnvelope[47], formants). Il offre diverses méthodes d'analyse telles que celle des partiels harmoniques et inharmoniques ou de la fréquence fondamentale (F0), plusieurs techniques de détection des transitoires (il permet de les traiter spécifiquement), d'estimation sinusoïdes/bruit et voisé/non voisé. SuperVP permet de faire toutes sortes de filtrages, par bandes, par surfaces, par enveloppe spectrale, par phase, etc. Divers traitements sont possibles comme le « déplacement de fréquences » (*frequency shift*), la suppression de bruits aléatoires et tonaux (*denoiser*), la transposition, la dilatation/contraction, etc.

SuperVP offre des méthodes de synthèse, telles que la synthèse source-filtre et la synthèse croisée généralisée, qui sont particulièrement adaptées à la transformation de voix, par exemple par application de l'enveloppe spectrale d'amplitude et/ou de phase d'une voix sur une autre. En particulier, SuperVP inclut les traitements les mieux adaptés pour la parole, notamment des filtrages temporels (LPC) et fréquentiels (par FFT), la transposition et la modification de durée avec conservation de l'enveloppe spectrale et de la forme d'onde (méthode dite *shape invariant*[48]). Il permet également d'utiliser le résultat d'une analyse de voisement non seulement en temps mais même en fréquence dans les transformations. Tous ces traitements spéciaux sont indispensables pour préserver la qualité de la voix. Les entrées et sorties d'analyse de SuperVP sont dans le format standard SDIF qui facilite leur gestion, leur maintenance et leur compatibilité avec les autres logiciels.

LE LOGICIEL AUDIOSCULPT

AudioSculpt[49] est une application graphique interactive d'analyse et de traitement musical des signaux sonores qui utilise essentiellement SuperVP comme moteur de traitement du signal. Ce logiciel permet l'étude très détaillée d'un son, de son spectre, de sa forme d'onde, de sa fréquence fondamentale et de son contenu en « partiels ». Toutes les analyses (comme le filtrage, la dilatation/contraction du temps et la suppression du bruit) peuvent être éditées, stockées (notamment en format SDIF) et employées pour guider le traitement dans l'application, ou peuvent servir d'entrée spectrale pour des environnements de composition.

47. Roebel et Rodet (2005).
48. Quatieri et McAulay (1992).
49. http://forumnet.ircam.fr ; *cf.* Bogaards *et al.* (2004).

Au cœur du logiciel se trouve une représentation très flexible du « sonagramme » du son suivant les diverses méthodes d'estimation spectrale de SuperVP. Une fois que le sonagramme a été obtenu, des filtres ou des traitements peuvent être dessinés directement dessus. AudioSculpt permet ainsi de « sculpter » un son de manière visuelle.

AudioSculpt comporte une classe unique de filtres spectraux de gain, appelés les filtres de « surface », qui permettent l'amplification ou l'atténuation d'une région arbitraire du plan temps-fréquence. D'autres traitements, qui peuvent être appliqués à une section ou à la totalité du son, incluent : passe-bande/rejection, transposition, dilatation, *écoute interactive en vitesse lente sans transposition* (et même « gel » spectral), suppression de bruit, écrêtage, filtrage *break-point* et par formants, etc.

Tous les traitements s'accompagnent d'un objet graphique sur le sonagramme aussi bien que d'un objet dans le « séquenceur de traitements », qui est un des outils les plus intéressants d'AudioSculpt. Ces objets peuvent être déplacés, modifiés, copiés, collés ou répliqués en temps et en fréquence. Le concept de séquenceur de traitements est très utile, et il permet aussi de se concentrer sur certains traitements et de les essayer seuls en débranchant momentanément les autres. Le traitement final complet sera ainsi réglé de façon optimale.

Outre la musique, il est aussi largement utilisé dans d'autres applications comme le *design* sonore, la postproduction cinéma et vidéo, la recherche et le développement scientifique ou la musicologie[50].

Exemples d'applications d'analyse, de transformation et de synthèse de voix

ANALYSE ACOUSTIQUE : ÉTUDE DE LA QUALITÉ VOCALE

En parallèle et indépendamment de la structure spécifiée du corpus, il est possible d'associer des analyses acoustiques produites par des analyseurs. Dans le cas de l'analyse et la synthèse de l'expressivité, l'une des caractéristiques prosodiques importantes est la qualité vocale. C'est pourquoi nous cherchons, par inversion du conduit vocal, à estimer une source d'excitation du conduit vocal. De nombreuses méthodes existent déjà[51], mais la difficulté réside dans leurs validations. En effet, il n'existe

50. La documentation est accessible à l'adresse suivante : http://support.Ircam.fr/forum-ol-doc/audiosculpt.

aucun moyen de mesurer *in vivo* cette source d'excitation. Cependant, à l'aide de la vidéo-endoscopie à haute vitesse, il est possible de filmer la glotte à un taux de 4 000 images/secondes. Sur cette suite d'images, la variation temporelle de l'aire de la glotte est calculée automatiquement[52]. La courbe de variation de l'aire permet alors l'estimation d'un débit glottique[53], qui est donc comparé à la source d'excitation estimée qui lui est fortement corrélée[54]. Dans ce contexte, il est donc indispensable de pouvoir manipuler des informations tant visuelles qu'acoustiques s'exprimant dans une ou plusieurs dimensions. La plate-forme IrcamCorpusTools permet une visualisation synchronisée de ces différentes informations et facilite ainsi l'interprétation des données multimodales. L'ensemble des paramètres glottiques estimés est alors calculé sur le corpus par un analyseur acoustique, et accessible par le langage de requête. Les étapes de construction d'un corpus et la spécification de ses différents niveaux d'analyse pertinents d'un point de vue linguistique, et l'estimation de divers signaux relatifs à la parole, permettent un grand nombre d'études linguistiques et/ou statistiques sur des régularités au sein de ces corpus.

CARACTÉRISATION ACOUSTIQUE DU PHÉNOMÈNE DE PROÉMINENCE

La proéminence est un phénomène prosodique majeur pour l'analyse et la modélisation de la prosodie[55]. L'analyse de ce phénomène peut se conduire en trois temps : dans une première étape, des outils statistiques sont utilisés pour permettre l'émergence des corrélats acoustiques de la proéminence et permettre leur détection automatique ; dans une deuxième étape, les proéminences détectées automatiquement sont utilisées pour faire émerger un ensemble de formes de la proéminence ; enfin, ces formes prosodiques sont étudiées par des linguistes pour réaliser une correspondance forme/fonction. Nous nous arrêterons ici seulement sur la première étape de cette étude : l'émergence automatique des corrélats acoustiques de la proéminence et sa détection automatique. Une modélisation statistique des corrélats acoustiques de la proéminence est rendue possible grâce à la complémentarité des possibilités d'IrcamCorpusTools et des méthodes statistiques implémentées en Matlab[56].

À titre d'exemple, voici les étapes menant à une caractérisation acoustique de la proéminence reposant sur la hauteur f0 moyenne des

51. Vincent *et al.* (2005), Henrich (2001).
52. Degottex *et al.* (2008a).
53. Maeda (1982).
54. Degottex *et al.* (2008b).
55. Rosenberg et Hirschberg (2007), Avanzi *et al.* (2008).
56. Obin *et al.* (2008c).

unités « syllabe », relativisée par rapport à la hauteur f0 moyenne des syllabes adjacentes ou par rapport à la hauteur f0 moyenne du « groupe prosodique » parent. Ainsi, nous pouvons accéder aux unités syllabes de la phrase 678 du corpus Ferdinand2007, ainsi qu'à leurs f0 moyennes comme précédemment[57] :

```
» syls = loadfeatures(corpus, 678, _syllabe_) ;
» f0_mean_syl = mean(segment(f0, syls)) ;
```

Grâce aux relations entre unités, on accède au groupe prosodique parent de chaque syllabe et à leurs f0 moyennes respectives :

```
» prosos = getparent(corpus, syls, _prosodic_);
» f0_mean_proso = mean(segment(f0, prosos));
```

Enfin, des valeurs relatives sont déterminées pour chaque syllabe, en divisant leurs hauteurs moyennes par la hauteur moyenne de leur groupe prosodique parent respectif (la fonction gv() extrait les valeurs des objets) :

```
» f0_mean_syl_rel = gv(f0_mean_syl) / gv(f0_mean_proso);
```

Cette étape montre comment les relations hiérarchiques entre les différentes unités de la phrase permettent à une unité d'un niveau donné d'hériter des données associées de ses « parents » ou d'agréger les données associées à ses « enfants ». On peut utiliser ces procédés d'analyse pour tout type de signaux et sur un corpus entier (ou sur la réunion de plusieurs corpus). Dans l'étude présentée, nous avons utilisé ces procédés pour générer une description du signal de parole pour toutes les syllabes.

Cette description comprend :

a) plusieurs corrélats acoustiques (fréquence fondamentale, durée, intensité, information spectrale et information spectrale perceptive) ;

b) plusieurs mesures sur ces corrélats au niveau de la syllabe (valeur moyenne, valeur maximale, etc.) ;

c) plusieurs empans de relativisation mis en œuvre pour relativiser la valeur observée sur une syllabe donnée, en fonction des valeurs observées dans son contexte (aucun contexte, syllabes adjacentes, groupe accentuel précédent, groupe prosodique précédent). Vis-à-vis d'une annotation de la proéminence manuelle ou découlant automatiquement de cette description, il est alors possible de filtrer les syllabes présentant une proéminence grâce à la requête :

57. Voir *supra*, p. NNN.

» syl_pro = getunits(corpus, _syllabe_, {_prominence_,_is_,_P_}) ;

On récupère, de la même façon, les syllabes ne présentant pas de proéminence. La figure 3 montre cette distinction binaire dans un espace réduit de la description dont les axes sont la durée et la f0 moyenne relativisée.

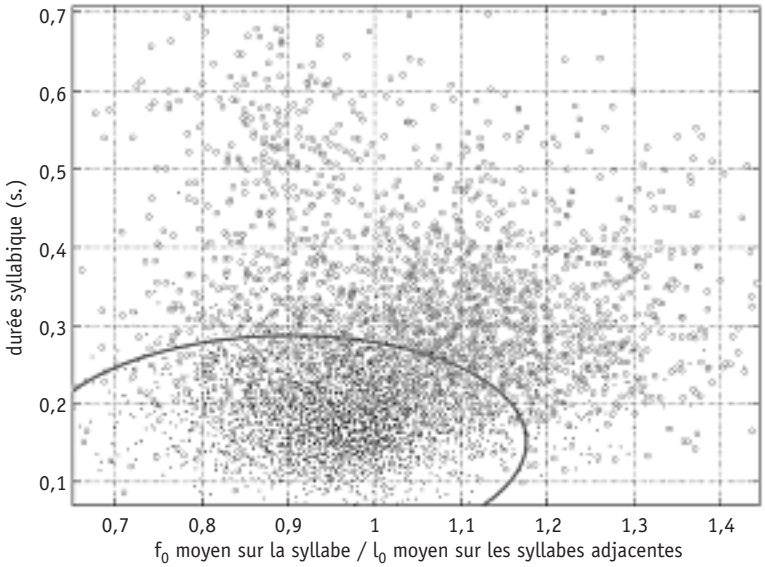

Figure 3 : *Distribution des syllabes proéminentes (cercles clairs) et non proéminentes (points sombres) dans un espace prosodique choisi. Le trait plein gras représente leur courbe de séparation quadratique.*

Les corrélats acoustiques de la proéminence sont appris grâce à des outils statistiques (machines d'apprentissage) de Matlab, dont le but est de déterminer l'ensemble ordonné des corrélats acoustiques observés sur la syllabe, qui permet la meilleure discrimination entre les syllabes proéminentes et les syllabes non proéminentes. La puissance du langage de requête d'IrcamCorpusTools a ainsi permis la caractérisation et la modélisation de la proéminence sur un corpus de voix parlée monolocuteur[58]. Grâce à la facilité d'intégration d'analyseurs externes, cette méthode a été confrontée à d'autres sur des corpus de parole spontanée[59]. Enfin, elle a permis la mise en place d'une méthode de caractérisation automatique des

58. Obin *et al.* (2008c).
59. Obin *et al.* (2008a).

genres de discours[60] : en plus de la simple lecture, le genre de discours d'interview, de discours politique, de dialogue, d'aide vocale, etc.

TRANSFORMATION DE VOIX

Sur la base des outils présentés dans les paragraphes précédents, diverses applications de traitement de voix parlée et de voix chantée ont été développées par l'Ircam. Nous exposerons ci-après la transformation de type et de nature (par exemple, transformation entre voix d'homme et voix de femme, ou chuchotement, voix rauque, etc.), la transformation de l'expressivité (joie, colère, doute, ironie, etc.) et la conversion d'identité (transformer la voix d'un locuteur pour qu'elle paraisse avoir été émise par un autre locuteur).

Transformation de type et de nature

Deux des caractéristiques les plus importantes d'une parole enregistrée sont la hauteur (ou *pitch*, assimilable à la fréquence fondamentale) et le timbre. La hauteur est liée à la longueur et à l'épaisseur des plis vocaux qui distinguent notamment les voix masculines graves des voix aiguës de femme et d'enfant. Mais une voix de femme relativement grave et une voix masculine légère, dont les hauteurs seraient semblables, diffèrent en général par leur *timbre*, ou leur *couleur*, par exemple une voix plus claire ou plus sombre, plus ou moins nasale, tendue, enrouée, avec plus ou moins de souffle, etc. Après avoir étudié les caractéristiques de différents types et natures de voix, nous avons développé la bibliothèque logicielle IrcamVoiceTrans[61]. IrcamVoiceTrans est fondée sur la bibliothèque SuperVP et permet de changer le type de voix entre les hommes, femmes, enfants, personnes âgées, etc., en temps réel. D'autres modifications de la voix originale peuvent donner l'impression d'une voix de fumeur ou d'une voix rauque, chuchotée, etc., ou même excitée ou ennuyée[62]. Un exemple typique d'application est une vidéo où les voix des quatre personnages (homme, vieille femme, adolescent et enfant) sont dérivées, par transformation d'âge et de genre, de la voix d'une seule locutrice (jeune femme)[63]. Cette technologie peut aussi être employée pour modifier la voix produite par notre système de synthèse à partir du texte IrcamTTS (voir ci-dessous). Les applications comprennent la musique, les installations, le théâtre, le spectacle vivant en général, et le multimédia, jeu

60. Obin *et al.* (2008b).
61. Farner *et al.* (2009).
62. Farner *et al.* (2008).
63. http://recherche.ircam.fr/equipes/analyse-synthese/demos.html

vidéo, animation, vidéo et films. Par exemple, pour le film *Tirésia* de B. Bonello nous avons transformé la voix de femme de l'héroïne en voix d'homme, et pour le film *Les Amours d'Astrée et de Céladon* d'Éric Rohmer, la voix d'homme de Céladon en voix de femme. Pour le film *Vatel* de Roland Joffé, c'est la prononciation anglaise de l'acteur Gérard Depardieu qui a été corrigée, par modification de l'accent tonique ou de certaines voyelles et consonnes. Certaines de ces transformations sont déjà opérationnelles en temps réel aussi dans l'environnement Max/MSP (objets SuperVP~) et dans l'application SuperVP-TRaX, distribués dans le Forum des logiciels de l'Ircam[64].

Modélisation et transformation contextuelles de l'expressivité

On entend par *expressivité* les aspects, essentiels dans la communication humaine, transportés par la parole comme l'émotion (joie, colère, peur, etc., chacune d'elles pouvant être « introvertie » ou « extravertie »), les affects, les intentions (ironie, doute, etc.), aussi bien que des notions comme surprise positive ou négative, discrétion, excitation ou confusion. Alors que les logiciels de synthèse à partir du texte (*text-to-speech* ou TTS[65]) ne fournissent que des phrases ayant une expressivité neutre, souvent un style de lecture, la plupart des applications, en particulier artistiques, nécessitent au contraire que la voix soit expressive. De la même façon, pour des applications en jeux vidéo, animation et film, il est indispensable de pouvoir modifier la façon de prononcer tel ou tel segment ou phrase, ou d'obtenir des effets qui ne correspondent pas à une voix habituelle, comme le fait un acteur. Celui-ci peut typiquement, sur une portion de phrase, modifier sa voix pour obtenir quantité d'effets comme un certain sous-entendu, une voix de « Mickey Mouse », une voix grinçante, une sorte de rire ou de gémissement, etc.

L'analyse et la synthèse de l'expressivité dans la parole sont donc un nouvel enjeu pour la communauté de la parole. Elles permettent de rendre les systèmes de reconnaissance vocale plus robustes et d'accroître le registre des synthétiseurs de la parole à partir du texte. De plus, elles sont un outil pour les psychologues qui étudient les émotions et les éventuelles pathologies qui y sont liées. Notre approche est avant tout motivée par le désir de modifier, à l'instar d'un acteur, l'expressivité d'une phrase parlée, qu'elle soit synthétisée ou bien naturelle.

Dans un projet récent, nous avons développé des solutions permettant de produire automatiquement ces modifications. Toutes ces transfor-

64. http://forumnet.ircam.fr
65. Voir *supra*, p. **NNN**.

mations doivent être d'une très haute qualité permettant d'étendre leur usage à des domaines d'application, restés jusque-là à l'écart, comme le film d'animation et le doublage de film.

Nous avons donc enregistré des acteurs exprimant un même texte avec différentes expressivités et avec différents niveaux d'intensité expressive. Ces corpus servent à l'établissement de modèles de jeux d'acteur. Ces modèles sont dépendants de variables contextuelles et linguistiques[66] comme le phonème ou le degré de proéminence. Par exemple, les variations acoustiques liées à l'expressivité dépendent du degré de proéminence (particulièrement pour le débit de parole)[67]. Et la transformation du degré d'articulation nécessite la connaissance du contexte phonétique[68]. La capacité d'IrcamCorpusTools de gérer différents types de données y est donc pleinement exploitée. Les variables linguistiques sont utilisées par un *réseau bayésien* pour estimer des densités de probabilités conditionnelles des variables acoustiques relatives aux cinq dimensions de la prosodie (hauteur, débit de parole, intensité, degré d'articulation et qualité vocale[69]). La comparaison de ces densités conduit à différents facteurs de transformation utilisés par le logiciel SuperVP pour modifier l'expressivité d'une phrase neutre. Des exemples sonores sont disponibles sur le Web[70].

Conversion d'identité

Comme en ce qui concerne l'expressivité[71], la plupart des applications, en particulier artistiques, nécessitent de donner à divers personnages, ou à des avatars, des voix bien différenciées et spécifiques du personnage. Il s'agit de produire, par un logiciel, une transformation artificielle de la voix naturelle d'un locuteur pour obtenir la voix d'un personnage différent (apprise sur l'enregistrement d'une voix réelle), ce qui est appelé *conversion d'identité*. En plus des applications artistiques, du doublage et du jeu vidéo, cette technologie trouve des applications dans d'autres domaines comme les avatars et les agents conversationnels. En effet, la méthode de doublage utilisée traditionnellement repose toujours sur le simple enregistrement d'acteurs. Dans les dessins animés comme dans les jeux vidéo, ou même dans le doublage de films, l'utilisation de techniques de modification de voix, au contraire, permet de créer les voix de plusieurs personnages avec l'enregistrement d'un seul acteur. Ainsi,

66. Beller et Rodet (2007).
67. Beller *et al.* (2006), Beller (2007b).
68. Beller (2007a), Beller *et al.* (2008).
69. Pfitzinger (2006).
70. http://recherche.Ircam.fr/equipes/analyse-synthese/beller
71. Voir *supra*, p. **NNN**.

l'Ircam peut proposer l'ensemble des technologies requises pour apporter une solution complète à l'utilisation de la voix dans les applications multimédias en général. Dans le domaine du film, plusieurs applications sont possibles. La première est une modification de voix lorsque l'acteur ne peut pas le faire lui-même. Cette technique a été employée, entre autres, dans le film *Vatel* de Roland Joffé pour améliorer la prononciation de l'anglais de Gérard Depardieu. Il serait même possible de modifier, voir de créer par synthèse à partir du texte, telle ou telle réplique d'un acteur sans avoir à le faire revenir en studio pour un enregistrement. Une autre application est la modification du timbre d'un acteur pour des contraintes liées au rôle. Cette technique a été testée pour le rôle de Klaus Maria Brandauer dans le film *Vercingétorix*. Enfin, la synthèse de la voix pourrait permettre de faire entendre la voix spécifique d'une personne non disponible (acteur non disponible après le tournage par exemple) ou disparue comme cela a été demandé à l'Ircam pour la voix du général de Gaulle ou celle du poète Jean Cocteau. Enfin, dans d'autres domaines comme le théâtre et la musique, des metteurs en scène et des compositeurs souhaitent pouvoir utiliser dans leurs créations des traitements et des synthèses de voix. Dans le cas du théâtre, il s'agirait, par exemple, de générer pendant les répétitions ou le spectacle une voix en temps réel qui servirait d'*alter ego* de l'acteur ou qui démultiplierait la voix de l'acteur. Ou encore de transformer en temps réel la voix d'un acteur en celle d'un autre personnage.

La conversion d'identité d'une voix source en une voix cible consiste à modifier le *timbre*, la hauteur (fréquence fondamentale ou *pitch*) et la durée de chaque phonème prononcé[72]. La méthodologie de conversion d'identité consiste d'abord à apprendre une fonction de transformation à partir d'enregistrements des locuteurs source et cible (par exemple, par la technique dite *mélange de gaussiennes*). Cette transformation peut être ensuite appliquée à la voix du locuteur source au moyen d'une technique d'analyse-synthèse comme SuperVP[73].

SYNTHÈSE DE LA PAROLE À PARTIR DU TEXTE

Le langage de requête de IrcamCorpusTools[74] peut être utilisé de manière manuelle par une succession de lignes de commandes comme dans l'exemple précédent. Mais il peut aussi être invoqué de manière automatique à différents niveaux, de manière à construire des arbres de

72. Villaviciencio *et al.* (2006).
73. Voir supra, p. **NNN**.
74. Voir supra, p. **NNN**.

données par de multiples décisions successives. Un synthétiseur de parole à partir du texte (*text-to-speech* ou TTS) a été construit de cette façon. Les procédés employés dans les systèmes TTS à base de corpus sont aujourd'hui bien connus[75], mais nous présentons ce système, IrcamTTS, car c'est une application du langage de requête. Après une phase d'apprentissage automatique impliquant de nombreuses requêtes sur les données symboliques de plusieurs niveaux (phones, diphones, syllabes, mots, groupes prosodiques, etc.), un arbre de décision est construit de manière à fournir en ses feuilles de nombreux sous-ensembles d'unités du niveau « diphone », qui peuvent d'ailleurs appartenir à des corpus différents. Chacune de ces feuilles correspond à des sous-ensembles d'unités acoustiquement homogènes. À l'étape de synthèse, ces sous-ensembles sont accessibles *via* une succession de requêtes construites à partir du texte à synthétiser. Il en résulte des sous-ensembles d'unités candidates. On sélectionne, parmi ces sous-ensembles, les unités qui minimisent une distance de concaténation, grâce à la programmation dynamique[76]. Cette distance est, elle aussi, apprise automatiquement, et permet de favoriser le naturel des transitions au niveau segmental, mais aussi au niveau suprasegmental. Comme le langage de requête d'IrcamCorpusTools permet de stipuler des contraintes acoustiques, celles-ci peuvent être définies et ajoutées, de manière à influencer la prosodie finale de la phrase de synthèse. Ces contraintes prosodiques peuvent, par exemple, être fournies par un modèle de la proéminence ou par un modèle de l'expressivité[77]. Dans la requête, il est également possible de bannir certaines unités, afin que l'algorithme de sélection en choisisse d'autres parmi les sous-ensembles candidats possibles : ainsi l'utilisateur, ou le programme appelant la requête, peut modifier ou améliorer la synthèse résultante.

VOIX CHANTÉE

L'Ircam a travaillé depuis longtemps sur la synthèse de la voix chantée. Ainsi, l'air de la Reine de la nuit de l'opéra de Mozart *La Flûte enchantée*[78] a entièrement été synthétisé par le logiciel Chant avec la technique dite *formes d'ondes formantiques*.

---

75. Hunt et Black (1996).
76. Viterbi (1967).
77. Voir respectivement supra p. **NNN** et **NNN**.
78. On peut l'entendre à l'adresse : http://recherche.ircam.fr/equipes/analyse-synthese/reine.html

De même, l'Ircam a développé une méthode de synthèse temps-réel[79] de chœurs parlés et chantés[80] pour la plate-forme logicielle Max/MSP, utilisée notamment dans l'opéra *K* de Philippe Manoury.

Enfin, pour le film *Farinelli* de Gérard Corbiaud, la voix du célèbre castrat a été créée par ordinateur par l'équipe « Analyse-synthèse » de l'Ircam[81]. Une soprane et un haute-contre ont été enregistrés, et leur voix montées note à note suivant les parties qu'ils pouvaient chanter au mieux. Ces voix ont été transformées, d'une part pour les rendre égales (et supprimer l'impression de passage d'une voix d'homme à une voix de femme à chaque transition de chanteur), d'autre part pour obtenir la réalisation artistique d'une voix de castrat exceptionnelle, sujet essentiel du film[82].

Transformation de voix parlée en voix chantée

Cette transformation a été conçue notamment pour l'œuvre *Lolita* de Joshua Fineberg (créée en 2006), fondée sur le livre éponyme de Vladimir Nabokov. Le travail est conçu comme un opéra, mais qui se produirait dans l'esprit du narrateur. Toutes les voix chantées entendues par l'assistance sont le résultat de traitements par ordinateur de la voix parlée du narrateur, graduellement transformée en voix de femme[83]. De cette façon, le morceau est un opéra véritablement imaginaire : c'est l'opéra imaginé dans l'esprit du narrateur.

La technique utilisée est une modification *shape invariant* par le logiciel SuperVP qui permet de grandes transpositions et de longs étirements de la voix parlée[84]. L'objectif final est un système automatique de transformation en temps réel de voix parlée en voix chantée, où les syllabes sont automatiquement assignées aux notes correspondantes de la composition. Dans son état actuel, le système n'est pas encore capable de réaliser l'alignement automatique des syllabes parlées et des notes. Pour cela, nous devons encore adapter à cette fin le système d'alignement (présenté p. **NNN**), comme nous l'avons fait pour le chant *fado* à l'occasion de l'œuvre *Com que voz* de Stefano Gervasoni.

Les notes de la composition exigent une transposition de la voix parlée en voix chantée, jusqu'à trois octaves au-dessus et une octave au-dessous. De même, pour donner aux voyelles de la voix parlée les durées des notes de la composition, il faut les étirer d'un facteur qui peut attein-

79. Elle est accessible dans le club d'utilisateurs de l'Ircam : http://forumnet.ircam.fr/
80. Schnell *et al.* (2000).
81. Depalle *et al.* (1995).
82. Depalle *et al.* (1994).
83. Roebel et Fineberg (2007).
84. Roebel et Rodet (2005).

dre la valeur huit ou dix. Ces deux transformations sont effectuées avec une grande qualité par le logiciel SuperVP. Enfin l'enveloppe spectrale du narrateur est transformée en celle d'une chanteuse.

Analyse de voix chantée

L'œuvre *Com que voz* de Stefano Gervasoni confronte le chant portugais traditionnel *fado* au chant en écriture contemporaine d'un ténor. Des enregistrements de la chanteuse de *fado* Christina Branco ont aussi été traités à l'Ircam. Pour cela le logiciel IrcamAlign[85] a été adapté à la phonétique de la langue portugaise et à la voix chantée. Ainsi les syllabes des enregistrements ont pu être segmentées et ont été utilisées comme matériau musical par le compositeur.

Conclusion

Dans ce chapitre, nous avons présenté des études et des développements, en particulier informatiques, d'analyse, de traitement et de synthèse de la voix et de la parole. En premier lieu, nous avons montré comment l'utilisation de grands corpus permet d'extraire des informations essentielles sur la structure de la parole et son contenu linguistique et acoustique. Gérer de tels corpus nécessite des outils particuliers. Dans l'équipe « Analyse-synthèse des sons », nous avons développé un tel outil, qui ouvre des possibilités remarquables. IrcamCorpusTools, une plate-forme extensible pour la création, la gestion et l'exploitation des corpus de parole, permet d'interfacer facilement des données hétérogènes avec des analyseurs internes ou externes, en utilisant le principe d'autodescription des données et des analyseurs. En outre, l'autodescription des données garantit leur pérennité, favorise l'introduction de nouveaux types et leur confère une plus grande visibilité. De même, l'autodescription des analyseurs assure l'extensibilité de la plate-forme ainsi que sa modularité et la mutualisation des corpus. La plate-forme IrcamCorpusTools est capable de gérer les relations hiérarchiques multiples et séquentielles entre des unités. Un langage de requête simple et expressif donne un accès immédiat aux données de ces unités. Ces fonctionnalités appliquées à différents corpus de parole (parole contrôlée et parole spontanée pour des études de la prosodie et/ou de l'expressivité) intéressent directement les

85. Voir supra, p. **NNN**.

recherches à la frontière entre le traitement automatique des langues et le traitement automatique de la parole. En guise d'exemple, un processus de synthèse de parole expressive à partir du texte peut être entièrement réalisé, depuis sa genèse jusqu'au résultat sonore. L'intégration de différentes exploitations rassemblées au sein d'une même plate-forme illustre les avantages de l'interopérabilité. C'est pourquoi nous avons le projet de distribuer publiquement IrcamCorpusTools à des communautés de chercheurs.

Pour le traitement du signal sonore lui-même, deux outils ont été présentés, SuperVP et AudioSculpt. Ils permettent de sculpter littéralement le son, comme un plasticien hors temps réel, ou même dans le temps réel de la parole, ce qui est essentiel sur scène notamment mais aussi pour la rapidité des mises au point.

Enfin nous avons montré comment ces recherches et logiciels sont utilisés dans la création musicale, artistique en général, et dans les multimédias où de tels outils sont de plus en plus demandés et mis en œuvre[86].

RÉFÉRENCES BIBLIOGRAPHIQUES

– Avanzi M., Lacheret-Dujour A. et Victorri B. (2008), « ANALOR. A tool for semi-automatic annotation of French prosodic structure », *Speech Prosody*.
– Barras C., Geoffrois E., Wu Z. et Liberman M. (1998), « Transcriber : A free tool for segmenting, labelingand transcribing speech », *LREC*, p. 1373-1376.
– Beller G. (2007a), « Influence de l'expressivité sur le degré d'articulation », *RJCP*, Rencontres jeunes chercheurs de la parole.
– Beller G. (2007b), « Transformation de la parole dépendante de l'expressivité et du texte », Journée des sciences de la parole.
– Beller G., Marty A. (2006), « Talkapillar : outil d'analyse de corpus oraux », *RJC-ED268*, Paris-III-Sorbonne-nouvelle.
– Beller G., Obin N. et Rodet X. (2008), « Articulation degree as a prosodic dimension of expressive speech », *Speech Prosody*, Campinas.
– Beller G. et Rodet X. (2007), « Content-based transformation of the expressivity in speech », *ICPhS*, Saarbruecken.
– Beller G., Schwarz D., Hueber T. et Rodet X. (2006), « Speech rates in French expressive speech », *Speech Prosody*, *SproSig*, ISCA, Dresde.
– Bilhaut F. et Widlöcher A. (2006), « LinguaStream : An integrated environment for computational linguistics experimentation », Trente, Italie.

86. REMERCIEMENTS. Le développement d'IrcamCorpusTools et des outils et logiciels décrits dans cet article, sont partiellement financés par le projet RIAM VIVOS (VIVOS : http://www.vivos.fr) sur la création de voix expressives pour des applications multimédias, par le projet ANR RHAPSODIE (RHAPSODIE : http://rhapsodie.risc.cnrs.fr) sur l'élaboration de corpus prosodiques de référence en français parlé et par le projet RIAM AFFECTIVE AVATARS, nouvelle génération d'avatars expressifs pilotés par la voix.

– Bird S., Day D., Garofolo J., Henderson J., Laprun C. et Liberman M. (2000), « ATLAS : A flexible and extensible architecture for linguistic annotation », *arXiv*.
– Bird S. et Liberman M. (2001), « A formal framework for linguistic annotation », *Speech Commun.*, vol. 33, n° 1-2, p. 23-60.
– Boersma P. (2001), « Praat, a system for doing phonetics by computer », *Glot international*, vol. 5-9, p. 341-345.
– Bogaards N., Roebel A. et Rodet X. (2004), « Sound analysis and processing with AudioSculpt 2 », *International Computer Music Conference* (ICMC), Miami (E.-U.).
– Cassidy S. et Harrington J. (2001), « Multi-level annotation in the Emu speech database management system », *Speech Communication*, vol. 33, p. 1-2, Elsevier Science Publishers B. V., Amsterdam, p. 61-77.
– Chafe W. (1992), « The importance of corpus linguistics to understanding the nature of language », *in* J. Svartvik (éd.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82*, Berlin-New York, Mouton de Gruyter, p. 79-97.
– Cunningham H., Maynard D., Bontcheva K. et Tablan V. (2002), « GATE : A framework and graphical development environment for robust NLP tools and applications », *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.
– De Cheveigné A. et Kawahara H. (2002), « YIN, a fundamental frequency estimator for speech and music », *JASA*, vol. 111, p. 1917-1930.
– Degottex G., Bianco E. et Rodet X. (2008a), « Measure of glottal area on high-speed videoendoscopy », *Production Workshop : Instrumentation-based Approach*, Paris.
– Degottex G., Bianco E. et Rodet X. (2008b), « Usual to particular phonatory situations studied with highspeed videoendoscopy », *International Conference on Voice Physiology and Biomechanics*.
– Depalle_P., Garcia G. et Rodet X. (1994), *A Virtual Castrato, International Computer Music Conference* (ICMC), Arrhus.
– P. Depalle, G. Garcia et X. Rodet (1995), « Reconstruction of a castrato voice : Farinelli's voice », *IEEE WASPAA*, Mohonk Mountain House (NY).
– Durand J. et Tarrier J.-M. (2006), « PFC, corpus et systèmes de transcription », *Cahiers de grammaire*, vol. 30, p. 139-158.
– Farner S., Rodet X. et Ach L. (2008), *Voice Transformation and Speech Synthesis for Video Games*, Paris Game Developers Conference, Paris.
– Farner S., Roebel A. et Rodet X. (2009), *Natural Transformation of Type and Nature of the Voice for Extending Vocal Repertoire in High-Fidelity Applications*, soumis à l'AES.
– Gussenhoven C. et Jacobs H. (2004), *Understanding Phonology*, Arnold, 2005.
– Gut U., Milde J.-T., Voormann H. et Heid U. (2004), « Querying annotated speech corpora », *Speech Prosody*, Nara, Japan.
– Habert B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? », *Linguistique sur corpus*, Perpignan, p. 11-58.
– Henrich N. (2001), *Étude de la source glottique en voix parlée et chantée*, PhD thesis, université Paris-VI.

– Hunt A. J. et Black A. W. (1996), « Unit selection in a concatenative speech synthesis system using a large speech database », *ICASSP*, IEEE Computer Society, Washington (DC), p. 373-376.
– Lai C. et Bird S. (2004), « Querying and updating treebanks : A critical survey and requirements analysis ».
– Lamel L., Gauvain J.-L. et Eskénazi M. (1991), « Bref, a large vocabulary spoken corpus for French », *EuroSpeech*.
– Lanchantin P., Morris A., Rodet X. et Veaux C. (2008), « Automatic phoneme segmentation with relaxed textual constraints », *Proc. of LREC*, Marrakech.
– Laroche J. et Dolson M. (1999), « New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications », *Journal of the AES*, vol. 47, n° 11, p. 928-936.
– Quatieri T. F. et McAulay R. J. (1992), « Shape invariant time-scale and pitch modification of speech », *IEEE Transactions on Signal Processing*, 40 (3), p. 497-510.
– MacWhinney B. (2000), *The CHILDES Project : Tools for Analyzing Talk*, troisième édition, vol. I : *Transcription Format and Programs*, Lawrence Erlbaum Associates, Mahwah (NJ)
– Maeda S. (1982), « A digital simulation method of the vocal-tract system », *Speech Communication*.
– Müller C. (2005), « A flexible stand-off data model with query language for multi-level annotation », *ACL*, p. 109-112.
– Nakov P., Schwartz A., Wolf B. et Hearst M. (2005), « Supporting annotation layers for natural language processing », *ACL*, p. 65-68.
– Obin N., Goldman J., Avanzi M. et Lacheret-Dujour A. (2008a), « Comparaison de 3 outils de détection automatique de proéminence en français parlé », Journées d'études de la parole, Avignon.
– Obin N., Lacheret-Dujour A., Veaux C., Rodet X. et Simon A.-C. (2008b), « A method for automatic and dynamic estimation of discourse genre typology with prosodic features », soumis à *Interspeech*, Brisbane, Australia.
– Obin N., Rodet X. et Lacheret-Dujour A. (2008c), « French prominence : a probabilistic framework », *Proc. of ICCASP*, Las Vegas (NV), P. 3993-3996.
– Oostdijk N. (2000), « The spoken Dutch corpus : Overview and first evaluation », *LREC*, 887-893.
– Pfitzinger H. (2006), « Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction », *in* Hoffmann, H., Mixdorff, R. (éd.), *Speech Prosody*, n° 40, Abstract Book, Dresde, p. 6-9.
– Rabiner L. (1989), « A tutorial on hidden Markov Models and selected applications in speech recognition », *IEEE*, vol. 77, 2, p. 257-286.
– Roebel A. (2003), « Transient detection and preservation in the phase vocoder », *International Computer Music Conference (ICMC)*, Singapour, p. 247-250.
– Roebel A. et Rodet X. (2005), « Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation », *Proc. of the 8th Int. Conf. on Digital Audio Effects (DAFx05)*, p. 30-35.
– Roebel A. et Fineberg J. (2007), « Speech to chant transformation with the phase vocoder », *Interspeech*, Anvers.

– Roebel A., Villaviciencio F. et Rodet X. (2007), « On cepstral and all-pole based spectral envelope modeling with unknown model order », *Pattern Recognition Letters, Special issue on Advances in Pattern Recognition for Speech and Audio Processing*, accepté pour publication.
– Rodet X., Escribe J. et Durigon S. (2004), « Improving score to audio alignment : Percussion alignment and precise onset estimation », *Proc Int. Conf on Computer Music* (*ICMC*), p. 450-453.
– Rosenberg A. et Hirschberg J. (2007), « Detecting pitch accent using pitch-corrected energy-based predictors », *Interspeech*, Anvers, p. 2777-2780.
– Schnell N., Peeters G., Lemouton S., Manoury, P. et Rodet X. (2000), *Synthesizing a Choir in Real-time Using Pitch Synchronous Overlap Add (PSOLA)*, ICMC : International Computer Music Conference, Berlin.
– Sjölander K. et Beskow J. (2000), « WaveSurfer. An open source speech tool », International Conference on Spoken Language Processing.
– Taylor P., Black A. W. et Caley R. (2001), « Heterogeneous relation graphs as a mechanism for representing linguistic tnformation », *Speech Communication*, vol. 3, p. 153-174.
– Veaux C., Beller G. et Rodet X. (2008), « IrcamCorpusTools : An extensible platform for speech corpora exploitation », *LREC*, Marrakech, Maroc.
– Villavicencio F., Roebel A. et Rodet X. (2006), « Improving LPC spectral envelope extraction of voiced speech by true envelope stimation », *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. I, p. 869-872.
– Villavicencio, F., Roebel, A. et Rodet, X. (2007), « All-pole spectral envelope modelling with order selection for harmonic signals », *IEEE International Conference on Acoustics, Speech and Signal processing* (*ICASSP'07*), 1, I-49 I-52.
– Villavicencio, F., Roebel, A. et Rodet, X. (2008), « Extending efficient spectral envelope modeling to mel-frequency based representation », *IEEE International Conference on Acoustics, Speech, and Signal processing* (*ICASSP'08*), p. 1625-1628.
– Villavicencio F., Roebel A. et Rodet X. (2009), « Applying improved spectral modeling for high-quality voice conversion », soumis à *ICASSP*.
– Vincent D., Rosec O. et Chonavel T. (2005), « Estimation of LF glottal source parameters based on an ARX model », *Interspeech*.
– Viterbi A. J. (1967), « Error bounds for convolutional codes and an asymptotically optimum decoding algorithm », *IEEE TIT*, vol. 13 (2), p. 260-269.