Synthèse concaténative de la voix chantée

Luc Ardaillon, Axel Roebel
équipe analyse/synthèse des sons
IRCAM - UPMC
luc.ardaillon@ircam.fr

1

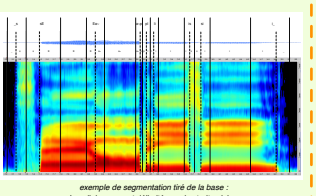
Les recherches présentées ici concernent la synthèse de voix chantée. Le but est d'obtenir, à partir d'un texte et d'une partition, une voix synthétique à la fois naturelle et expressive chantant la partition et le texte donnés en entrée.
La méthode utilisée pour cela est la concaténation et transformation d'unités [Bonada 2006] : une base de données est enregistrée par un chanteur, de laquelle des segments sont extraits puis assemblés afin de générer la synthèse. Les discontinuités entre ces segments doivent alors être supprimées.
Afin de rendre la synthèse naturelle et expressive, un module de contrôle doit permettre d'en générer les paramètres (hauteur, durées, intensité et timbre) en fonction des contextes donnés en entrée par la partition (mélodie, paroles) et certains choix esthétiques (style, accents expressifs, ...).
Les paramètres ainsi générés permettent alors de transformer les segments concaténés afin d'obtenir le résultat final de la synthèse.

Base de données

Enregistrements :

- Série de courtes phrases chantées lentement à **hauteur et intensité fixes** couvrant l'ensemble des **diphones de la langue française**

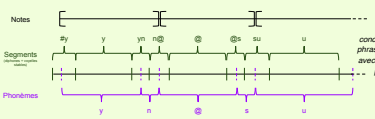
- Bases différentes enregistrées pour **2 styles** "lyrique" et "variété", afin de capturer les **qualités timbrales propres à chaque style**



*exemple de segmentation tiré de la base :
Les diphones sont délimités par les traits pleins,
Les frontières entre phonèmes sont indiquées en pointillés*

Stratégie de segmentation :

- Double segmentation :
(1) en **phonemes** => gestion précise des durées et placement des phonèmes
(2) en **diphones** "courts" + parties stables => gestion indépendante des parties co-articulées (transitions entre phonèmes) et stables (tenues de notes)



exemple de concaténation pour la phrase "une souris ...", avec segmentation en phonèmes et diphones

- **Segmentation automatique** en phonèmes et parties stables => permet de déduire les segments des diphones à utiliser pour la synthèse

Contrôle

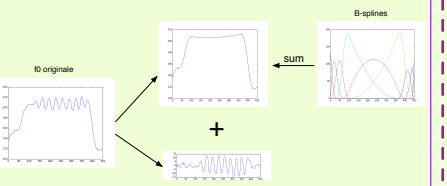
Objectifs :
- Synthèse naturelle et expressive
- Modélisation du style
- Contrôle paramétrique intuitif
- Indépendant du système de synthèse

Etude :
- Constitution d'un corpus (lyrique et variété française) avec 2 chanteurs professionnels pour étude et apprentissage des paramètres **en contexte**

Fréquence fondamentale

- 2 composantes principales :
  - courbe "lisse" avec transitions entre notes
  - vibrato
  (+ inflexions liées au contexte phonétique)

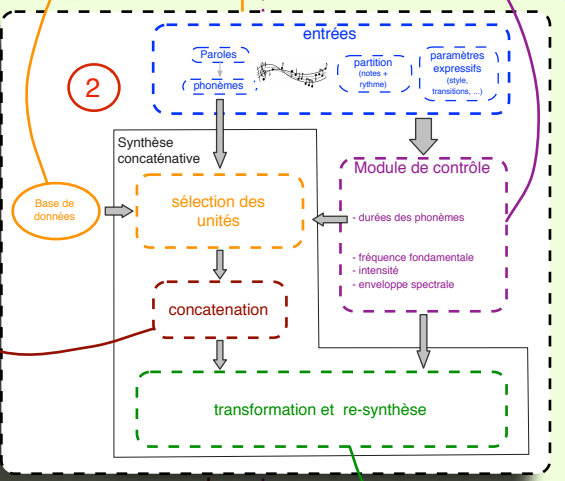
- Modélisation par B-splines avec approximation par méthode des moindres carrés => modèle paramétrique



3

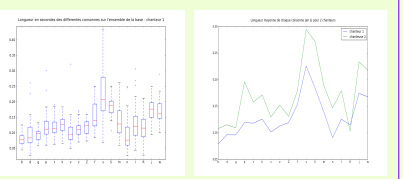
5

2



Durées et alignement des phonèmes

Consonnes : liées à l'articulation
- durée moyenne propre à chaque consonne pour un débit lent (dépendant du locuteur)



- durées varient en fonction du contexte phonétique (début de phrase, enchainement de consonnes, ...) et du tempo => à apprendre

Voyelles : liées aux notes
=> début des notes sur les onsets des voyelles
=> $durée_{voyelle} = durée_{note} - durée_{consonnes}$

autres paramètres:

- intensité / effort vocal
- enveloppe spectrale (formants et source)
=> correlation avec hauteur et intensité à établir

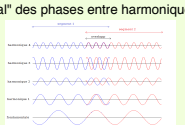
Concaténation

- Dans le **domaine spectral**

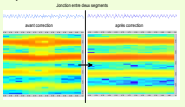
- 2 types de **discontinuités** à supprimer

4

1). Phase
Problème : alignement "vertical" des phases entre harmoniques différent d'un segment à l'autre



Solution : étaler ces différences dans le temps, pour chaque harmonique, avec un écart maximum imperceptible entre 2 trames



Temps entre centre de la trame et phase 0 de la fondamentale :
$dt = -\varphi_{centre}/\omega_0$ avec $\omega_0 = f_0 \times 2\pi$

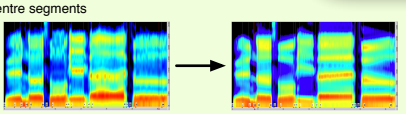
Différence de phase entre l'harmonique i et la fondamentale :
$\varphi_0^i = princarg(\varphi_{centre}^i + 2\pi \times f_0 \times dt)$

Ecart entre les trames n et n-1 :
$\Delta\varphi_0^i(n) = princarg(\varphi_0^i(n) - \varphi_0^i(n-1))$

Si $abs(\Delta\varphi_0^i(n)) > \Delta\varphi_{max}$:
$=> \varphi_0^i(n) = princarg(\varphi_0^i(n-1) + \Delta\varphi_{max} \times sign(\Delta\varphi_0^i(n)))$

2). Enveloppe spectrale
Problème : différences d'enveloppe d'un segment à l'autre pour un même phonème

Solution : interpoler les enveloppes sur les parties stables (tenues) et autour des jonctions entre segments



Transformation et re-synthèse

6

Pitch-shifting :
- dans le domaine spectral
- Algorithme **SHIP** [Roebel 2010] : Conservation des relations de phases entre harmoniques d'une trame à l'autre => haute qualité pour la voix (pas d'effet de "phasiness")

Time-stretching :
- Sur les parties stables pour notes tenues
- Sur les parties co-articulées pour contrôler la vitesse d'articulation

Re-synthèse :
IFFT + overlap-add

Expériences de synthèse et résultats

7

Objectifs :
- Evaluation du moteur de synthèse
- Identification des composantes à améliorer

Moyen :
- Re-synthétiser un extrait de chant enregistré dont on peut extraire certains paramètres (fréquence fondamentale et durées des phonèmes) et les imposer lors de la synthèse, en s'affranchissant du module de contrôle. Cela permet d'Isoler l'influence de chaque paramètre sur le résultat perçu.

Conclusion :
- Synthèse de bonne qualité (écoute informelle), ce qui semble valider l'ensemble du processus (contrôle mis à part)
- Durées des phonèmes et f0 essentielles pour un résultat naturel => importance d'un module de contrôle adapté
- Timbre "statique", sans variations expressives => adaptation de l'enveloppe spectrale nécessaire

Résultats préliminaires :
Synthèse de bonne qualité, avec jonctions transparentes sans discontinuités, et peu d'artefacts
Identification des points à améliorer

8

Deux principaux axes de recherche :
(1) contrôle des paramètres en fonction des contexte et du style, avec apprentissage
(2) adaptation de l'enveloppe spectrale (source et conduit vocal) en fonction des autres paramètres (hauteur, intensité) et de la qualité vocale souhaité