

Simulations de voix et de parole pour Prosodie - Genève 2002

Eric Keller
Laboratoire d'analyse informatique de la parole (LAIP)
Université de Lausanne
eric.keller@imm.unil.ch

Point de départ: la simulation de locuteurs naturels

Notre objectif à long terme est la synthèse de parole pour locuteurs virtuels (reconstruction historique, cinéma virtuel, etc.).

Plusieurs limites de connaissances nous séparent de cet objectif.

Notamment, la parole de nos synthétiseurs n'est pas encore naturelle. Ceci est largement dû aux trois raisons suivantes (Keller, 2001 "Volume COST 258"):

- Les limites intrinsèques des systèmes concaténatifs temporels
- L'état insuffisant de développement des systèmes paramétriques
- La compréhension toujours insuffisante des effets sur le signal de divers aspects de la parole, de l'expressivité et de la voix.

Limites des systèmes concaténatifs

La technologie concaténative temporelle en utilisation générale (y-compris Mbrola et les systèmes à bases de données étendues) est caractérisée par des limites intrinsèques:

- Une synthèse limitée à la manipulation des paramètres *fzéro*, durée et (parfois) amplitude
- Chaque nouvelle voix *et chaque nouveau style de parole* requiert une nouvelle base de données (typiquement 3-36 mois de travail pour chaque nouvelle BD)
- Dans le cas de systèmes diphoniques ou polyphoniques, l'aspect concaténatif oblige l'utilisation de parole bien articulée, facile à segmenter et simple à enchaîner.

Limites des systèmes paramétriques

Actuellement, il n'existe pas encore d'alternative satisfaisante aux systèmes concaténatifs.

- Les systèmes paramétriques de la première génération (les synthèses formantiques [p.ex. Klatt, Stevens]) préservent insuffisamment les caractéristiques vocales des locuteurs.
- Les systèmes paramétriques de la deuxième génération sont seulement naissants (p.ex. Systèmes sinusoidaux de type HNM, voir page suivante).
- Actuellement, il y a un manque de connaissances sur...
 - La simulation d'effets coarticulatoires dans le signal, y-compris la gestion non-segmentale de spécifications en amont (p.ex., le modèle phonologique non-segmental de John Local [York])
 - L'identification dans le signal, la manipulation et la reproduction contrôlée de caractéristiques individuelles de la voix, de l'expressivité et des émotions
- Ces connaissances sont préalables à la simulation satisfaisante de la parole conversationnelle.

Etat actuel des systèmes paramétriques de la deuxième génération

Les systèmes synthèses sinusoidales (p.ex. HNM "Harmonics and Noise") sont au premier rang des systèmes paramétriques de la deuxième génération. Ces systèmes sont souvent conçus de manière hybride:

1. Reproduction d'un système *concaténatif* au niveau de la gestion temporelle et diphonique.
2. Utilisation de la technologie *HNM* pour (a) les manipulations de *f0*, ainsi que (b) le lissage aux transitions segmentales et diphoniques, et/ou polyphoniques.
3. Etat actuel de la HNM développée par l'auteur en 2002 (cadre COST 277): 1. et 2(a). En cours de développement: 2(b).
4. Développement ultérieur (2003+): implantation graduelle de la gestion co-articulaire et des distinctions de caractéristiques individuelles

A terme, ce développement devrait permettre la simulation réaliste de personnages virtuels.

Les simulations pour ce colloque

Dans les pages suivantes, nous commenterons les simulations que nous avons préparées pour ce colloque sous trois thèmes:

- Manipulations *f0* et durée
- Copie-synthèse et transposition de paramètres
- Synthèse complète

Manipulations f0 et durée: Procédure

Ces manipulations impliquent les étapes suivantes:

1. Entrée du signal
2. Analyses locales de Fourier afin d'obtenir
 - La fréquence fondamentale (f0)
 - L'enveloppe spectrale
 - Le bruit résiduel
3. Manipulation des raies harmoniques, sans modification de l'enveloppe spectrale, rajout du bruit résiduel après la modification
4. Manipulation des durées dans le domaine temporel (suppression ou rajout de cycles de f0)
5. Concaténation des régions localement modifiées
6. Sortie du signal

Manipulations de la fréquence fondamentale et de la durée - exemples

Fichier original < Wyss Groult >	fréquence fondamentale (mélodie) 15% en- dessous de la fréquence fondamentale originale	fréquence fondamentale (mélodie) 15% au- dessus de la fréquence fondamentale originale
Débit 15% en-dessous du débit original	< Wyss Groult >	< Wyss Groult >
Débit 15% au-dessus du débit original	< Wyss Groult >	< Wyss Groult >
Manipulations extrêmes (presqu'une octave d'écart)	f0 40% en-dessous < Wyss (± -35 Hz) Groult (± -82 Hz) >	f0 50% au-dessus < Wyss (± +57 Hz) Groult (± +128 Hz) >

Commentaires: Les comparaisons entre les deux locuteurs montrent que les manipulations f0 et durée de la voix plus élevée et plus irrégulière de Mme Groult sont relativement difficiles à accomplir. Les analyses détaillées des représentations spectrales des signaux de Mme Groult suggèrent que ceci serait surtout dû à: (1) limites de la résolution spectro-temporale de voix à f0 élevée, (2) difficulté de manipulation de voix à paramétrisation fort irrégulière, et (3) le bruit de fond de l'enregistrement.

Copie-synthèse (CS) et transposition diphoniques (TR): procédure

La copie-synthèse et la transposition de paramètres prosodiques permettent de vérifier la modification de la f0 et des durées dans le cadre d'un système diphonique simplifié. Ceci diffère d'une synthèse complète dans ce sens qu'on écarte les problèmes de concaténation de diphones provenant de différents contextes. Nous avons procédé comme suit (d'autres approches sont possibles).

1. (CS et TR) Extraction automatisée et édition manuelle de la segmentation phonétique
2. (CS et TR) Création d'une base de données diphoniques
3. (CS et TR) Récupération des diphones requis pour l'énoncé actuel dans la phrase d'origine
4. (Soit CS >) Entrée du signal, extraction automatique de la f0 et obtention de la durée
5. (Soit TR >) Génération de la f0 et de la durée par LAIPTTS_F, ou f0 stable (monotonie)
6. (CS et TR) Manipulation de f0 dans le domaine spectral
7. (CS et TR) Manipulation de la durée dans le domaine temporel (suppression/duplication de cycles)
8. (CS et TR) Concaténation des éléments modifiés
9. (CS et TR) Sortie du signal

Copie-synthèse et transposition: exemples

Copie-synthèse: f0 et durées obtenues directement d'une phrase d'entrée (f0 par l'algorithme d'autocorrélation développé par l'auteur):

- Wyss: original
- Groult: original
- copie-synthèse
- Mbrola (pas strictement comparable)

Transposition d'une f0 stable (monotonie):

- Wyss:
- Groult:

Transposition de paramètres prosodiques

- Wyss: HNM
- Groult: HNM
- Mbrola (pas strictement comparable)
- Mbrola (pas strictement comparable)

Commentaires:

- Différences de qualité pour les deux locuteurs
- L'extraction des trajets f0 par notre algorithme AC fournit des résultats satisfaisants

La synthèse complète: procédure

Une synthèse complète implique l'abstraction, la modélisation et la simulation en synthèse de paramètres linguistiques, prosodiques et phonétiques. Pour LAIPTTS, ceci est accompli en trois stades: (a) traduction entre représentations graphique et phonétique (la "phonétisation"), (b) traitement prosodique et (c) génération du signal ("codage").

- Notre synthèse du français n'est pas actuellement en mesure de simuler des conversations spontanées de personnes, car...
 - Le système a été modélisé sur la parole lue, non sur la parole conversationnelle
 - Le système est essentiellement au stade de 1997 et n'incorpore pas les résultats de recherche des 5 dernières années
 - Le système n'incorpore pas encore la génération de signaux.

A fins d'illustration, nous avons produit la simulation d'une partie de la conversation entre les Mmes Groult et Fayard, en utilisant (1) notre système de 1997 et (2) la génération de signaux par Mbrola (Fayard: F2, Groult: F4).

Synthèse complète: exemple

Extrait de l'interview:

Commentaires: débit et f0 trop réguliers, voix Mbrola peu naturelles, parole trop formelle, absence d'indices d'interactivité, d'emphase et d'expressivité, pas de manipulation de l'amplitude, simulation insatisfaisante des entrecoupages, etc. etc.

Conclusions

- **Méthodologie:** dans certaines limites, une simple HNM à traitement local (sans traitement sur le temps de paramètres spectraux) permet d'effectuer d'assez bonnes modifications f0. La résolution temporo-spectrale, la précision du modèle de l'enveloppe et de la séparation du bruit, ainsi que l'état actuel de l'algorithme, imposent des limites au traitement de "voix difficiles".
- **Recherche:** La simulation de la parole conversationnelle et expressive requièrera l'implantation d'un grand nombre de connaissances supplémentaires
- **Développement futur:** complétion du système diphonique avec concaténation spectrale - modélisation progressive de l'interaction des paramètres spectraux et temporels - à long terme, un système de plus en plus paramétrisé.