IMAGE ET REALITE VIRTUELLE

Parole et langage

G. Bailly

Sujet d'examen du 25 février 2004
2 heures - Cours et documents autorisés

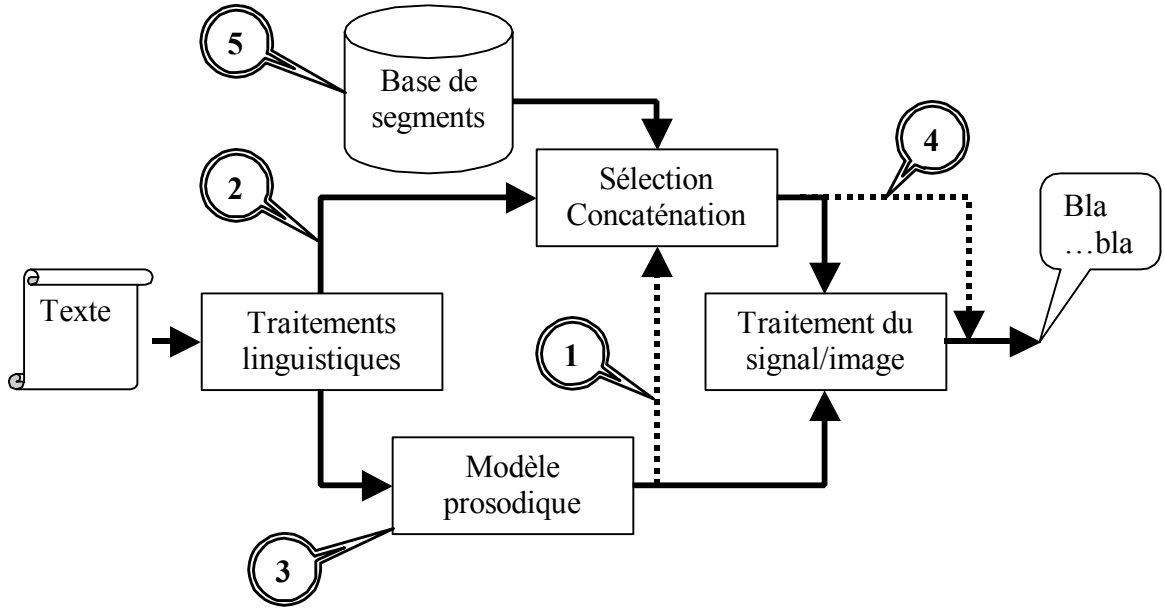

Figure 1 : Schéma synoptique d'un système de synthèse par concaténation d'unités stockées.

Synthèse de la parole – 10pts

Vous devez mettre en oeuvre un système de synthèse opérant par concaténation d'unités pré-stockées. Le schéma synoptique classique d'un tel système est donné Figure 1. Les questions posées ci-dessous réfèrent aux « bulles » (pointant sur les objets d'intérêt) correspondantes sur la figure.

1. Le modèle prosodique calcule typiquement des paramètres « supra-segmentaux » (mélodie, durées des sons… mais aussi mouvements de la tête ou des mains) qui organisent les segments en fonction de la structure linguistique et de diverses fonctions discursives (attitudes, émotions…) que l'on veut entendre/voir véhiculer dans le discours synthétique. Le module « Traitement du signal/image » déforme alors les segments de parole naturelle (stockés dans la base de segments) sélectionnés et concaténés par le module de « Sélection/concaténation» de manière à ce que les signaux de synthèse obéissent bien à l'évolution de ces paramètres.
   Pourquoi peut-on envisager d'injecter ces paramètres dans le module de « Sélection/concaténation» ? Expliquer comment l'opération de sélection pourra incorporer cette information. Donnez un exemple (par exemple sur la durée des sons). En quoi cela peut-il conduire à améliorer le travail du module de « Traitement du signal/image » ? Donnez un exemple (par exemple sur la mélodie).
2. Au delà de la simple chaîne phonétique à prononcer, peut-on injecter des informations sur la structure linguistique et les diverses fonctions discursives directement dans le module de « Sélection/concaténation» sans passer par le « Modèle prosodique » ? Expliquer comment l'opération de sélection pourra incorporer cette information.
3. Dans ce cas le « Modèle prosodique » est-il toujours nécessaire ? Qu'est-ce que cela implique sur la base de segments ?

4. On supposant que la base de segments est gigantesque et que l'opération de sélection/concaténation intègre effectivement toutes les informations que l'on réussi à extraire du texte sur la structure linguistique et les diverses fonctions discursives désirées, pourquoi on peut éventuellement se passer aussi du module de « Traitement du signal/image ». Quels sont les petits détails que ce module peut cependant encore régler ?
5. Quels sont les problèmes posés par la collection d'une base de segments de plusieurs heures ? Développez notamment ceux posées par la consistance de la qualité de la voix, de son expression faciale… et par la couverture linguistique des textes qu'il aura à prononcer. Comment procéderez-vous pour choisir ces textes ? Que proposez-vous pour éviter l'« effet de liste » (l'intonation a tendance a être monotone lorsqu'on demande à un locuteur de lire des milliers de phrase). Que proposez-vous pour contrôler a posteriori la qualité de sa prononciation (a-t-il bien dit ce que vous lui avez donné à lire sans se tromper… la fatigue aidant, la concentration et la motivation diminue !) ?

Reconnaissance de la parole – 6pts

Vous devez concevoir un système qui permet d'identifier des mots épelés (consistant donc en l'énoncé des 26 lettres de l'alphabet /a/, /be/, /se/…/zɛd/) prononcés sans pauses entre les lettres constitutives (ex : image prononcé comme /iɛmaʒeœ/). Vous disposez de plusieurs exemplaires de prononciation de chaque lettre par divers locuteurs.

1. Expliquer le principe de la reconnaissance par double programmation dynamique
2. Diviser le jeu de lettres en plusieurs classes, à l'intérieur desquelles il y aura le plus d'erreurs de reconnaissance (de confusions possibles)
3. Quels sont les avantages et désavantages à juxtaposer des modèles de lettres plutôt que d'utiliser des modèles de mots (tous épelés évidemment)
4. Pourquoi cette technique d'épellation des mots constituent en général un problème de reconnaissance plus difficile que leur prononciation normale ?
5. Expliquer comment le choix du lexique peut influer sur la remarque précédente.
6. Considérant le jeu de lettres (/be/, /se/, /de/, /ʒe/, /pe/, /te/, /ve/), suggérez un moyen de réduire les confusions au sein de ce groupe. Quels sont les indices phonétiques optimaux permettant d'effectuer cette réduction ?

Animation faciale (4pts)

Vous devez concevoir un agent conversationnel capable de parler tout en exprimant quelques expressions faciales (étonnement, colère, sourire…)

1. Quel est l'intérêt pour un système de dialogue personne-système de disposer d'un agent doté de ces capacités d'expression ?
2. Quels sont les problèmes posés aux systèmes basés-image opérant par superposition de « patchs » sur une vidéo de fond ?
3. Quels sont les problèmes posés aux systèmes basés-modèle ? Comment allez-vous répartir/gérer les compétences des modèles de contrôle, de forme et d'apparence ? Qui gérera la négociation entre mouvements faciaux nécessaires à la production des sons de parole et ceux nécessaires aux expressions faciales ? Comment comptez-vous combiner des mouvements antagonistes (ex sourire tout en prononçant sur un /u/)?
4. Comment comptez-vous recueillir des expressions réalistes… sans pour autant heurter l'éthique ?
5. Comment pourrez-vous vérifier que l'agent que vous avez réalisé véhicule effectivement/efficacement ces expressions ?