Synthèse de parole à partir du texte I

Text To Speech - TTS

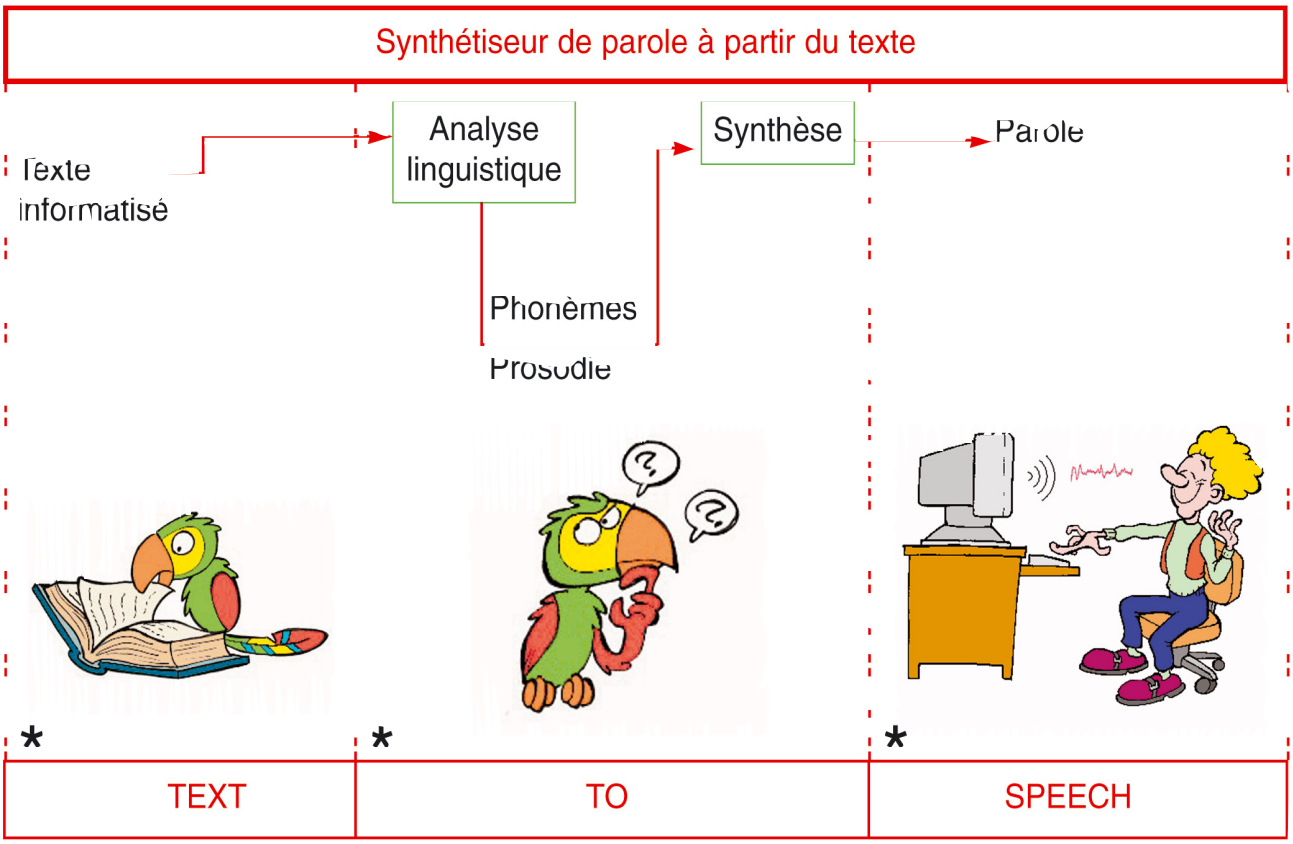
INTRODUCTION

Comme son nom l'indique, le but d'un synthétiseur de parole à partir du texte est de donner lecture d'un texte. Il s'agit d'un outil logiciel qui transformera un fichier texte en un fichier audio. Dans le cas où le document n'est disponible qu'en version "papier", il suffira de le scanner puis d'utiliser un logiciel de reconnaissance des caractères optiques afin d'en obtenir une version informatisée.

PRINCIPE

Comme illustrée ci-dessous, cette transformation du 'texte' en 'parole' se fera en deux étapes :

1. Analyse linguistique
2. Synthèse proprement dite



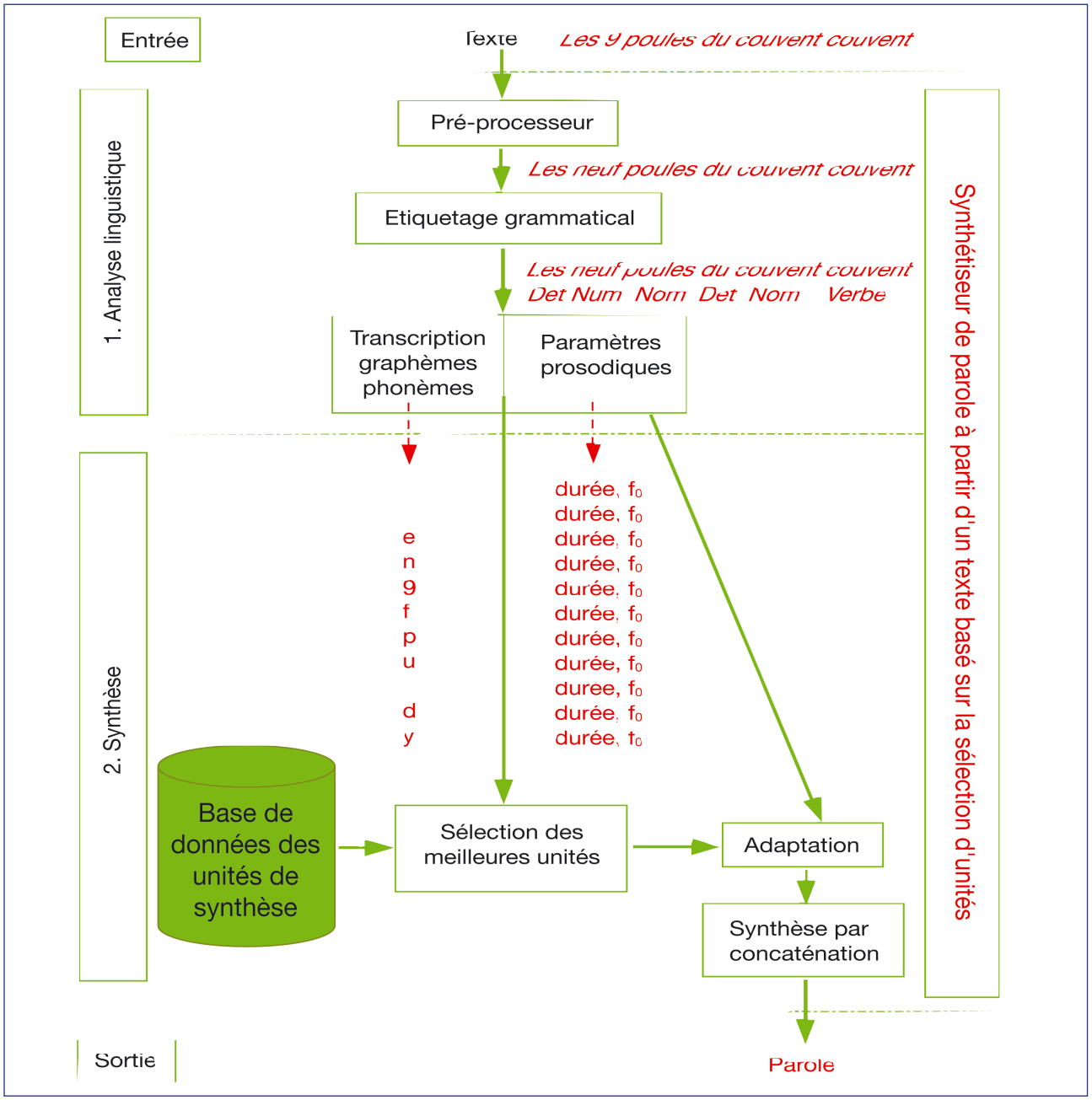
* Pictogrammes issus du site : http:/www.picto.qc.ca/

Synthèse de parole à partir du texte II

Text To Speech - TTS

Plus en détails

Le schéma suivant vous permet de suivre pas à pas les différents traitements qui sont effectués par le synthétiseur afin de transformer le texte de départ (que nous supposons sous forme informatisée) en un signal de parole.



Synthèse de parole à partir du texte III

Text To Speech - TTS

Un mot d'explication

Nous retrouvons sur ce schéma les deux types de traitement dont nous avons déjà parlé, à savoir : • Analyse linguistique
• Synthèse de parole

Pour chacun de ces types de traitement, cette rubrique nous permettra de mieux comprendre le rôle de chacun des blocs qui le constituent.

1. Analyse linguistique

L'analyse linguistique a pour but de transformer le texte "écrit" en une suite de "phonèmes", ou plus simplement de sons, à laquelle est associée une prosodie. Pour atteindre ce but, les opérations suivantes sont nécessaires :

1.1 Pré-processeur (Mise en forme du texte)

Les textes écrits contiennent bon nombre de pierres d'achoppement pour l'ordinateur qui devra en donner lecture. En effet, aux caractères alphabétiques viennent s'ajouter les caractères numériques (9, 58, …) et typographiques (; , ? : / …) ainsi que la combinaison des deux (9.627, 7/8, 5_, 32 65 / 44 45 16…) sans compter les abréviations, les images, les tableaux,… Tous ces écueils doivent être éliminés avant même de pouvoir commencer l'analyse proprement dite. C'est pourquoi le module ayant cette mission est appelé "pré-processeur" : il effectue un pré-traitement, une "mise en forme " de l'information pour pouvoir en faire l'analyse. Retenons que ce pré-processeur a pour mission principale de transformer le texte de base en une suite de graphèmes (suite de lettres).

Dans notre exemple, nous constatons que le chiffre 9 de la phrase "les 9 poules du couvent couvent" a été transformé en lettres (neuf).

1.2 Etiquetage grammatical

L'ordinateur dispose maintenant d'un texte "écrit en toutes lettres"(suite de graphèmes) qu'il va devoir convertir en phonèmes (suite de sons codés par un alphabet). Facile, me direz-vous, il suffit d'une table de conversion et le tour est joué ! Pas si simple, en réalité !! Prenons notre phrase-exemple "Les neuf poules du couvent couvent.". La graphie"couvent" y apparaît deux fois mais la prononciation associée est différente selon le cas et donc la transcription phonémique sera également différente. Comment lever cette ambiguïté ? Dans cet exemple, la différence de prononciation peut être identifiée par la catégorie grammaticale à laquelle le mot appartient puisque dans le premier cas, il s'agit d'un nom et que dans l'autre, il s'agit d'un verbe. Il est donc utile d'attribuer la bonne catégorie grammaticale à chacun des mots du texte afin de pouvoir en déduire la bonne prononciation. Cette fonction est remplie par l'étiqueteur grammatical et donnera, dans le cas de notre exemple, le résultat suivant :

Les	neuf	poules	du	couvent	couvent
Det.	Num.	Nom	Det.	Nom	Verbe

Synthèse de parole à partir du texte IV

Text To Speech - TTS

1.3 Transcription graphèmes – phonèmes

L'ordinateur dispose maintenant de l'information nécessaire et suffisante pour pouvoir convertir les graphèmes (la forme écrite du texte) en phonèmes.

Regardons ce que cela donne dans le cas de notre phrase-exemple :
le n9f pul dy kuva~ kuv

1.4 Paramètres prosodiques

Prêt pour la génération de la parole, alors ? Pas encore tout à fait ! En effet, la manière de prononcer une phrase est certes liée à la suite des phonèmes qui la composent mais également à la nature et au contexte de la phrase. Ainsi, l'intonation, le rythme et l'accentuation donneront une perception particulière d'une phrase donnée : c'est ce qu'on appelle la prosodie. Si nous effectuons un parallèle avec la musique, la prosodie correspond à la mélodie de la parole.

Reprenons notre phrase-exemple pour bien comprendre :

" *Les neuf poules du couvent couvent.* "

La prosodie de cette phrase sera toute différente selon que la phrase est de type :

- Affirmatif (relate un fait)
- Exclamatif (exprime la surprise, la stupéfaction, l'admiration,…)
- Interrogatif (exprime l'interrogation).

Le module prosodique a donc pour mission d'associer aux phonèmes une durée impliquant le rythme (plus la durée est courte, plus le rythme est rapide) ainsi qu'une fréquence fondamentale donnant l'intonation (par exemple, les questions se terminent souvent sur une fréquence fondamentale plus élevée, une sonorité plus aiguë que les phrases affirmatives).

A ce stade, nous disposons d'une suite de phonèmes ainsi que des paramètres prosodiques à respecter que nous pouvons symboliser comme suit, en reprenant notre phrase-exemple :

_	durée, f0
l	durée, f0
e	durée, f0
n	durée, f0
9	durée, f0
f	durée, f0
p	durée, f0
u	durée, f0
l	durée, f0
d	durée, f0
y	durée, f0
...	… …

Ces informations forment la **consigne** à respecter sur base de laquelle le module de synthèse va pouvoir générer la parole souhaitée.

Synthèse de parole à partir du texte V

Text To Speech - TTS

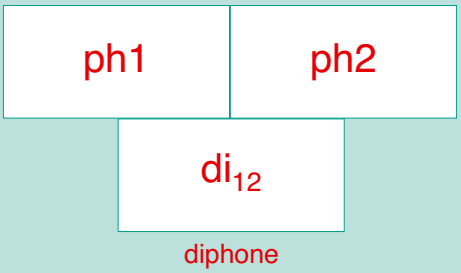
2. Synthèse

Le module de synthèse est celui où sont mises en œuvre les techniques de traitement du signal. Nous nous limiterons à la synthèse par concaténation d'unités et nous nous reporterons à la littérature pour ce qui est des autres techniques de synthèse. Ce type de synthèse est par ailleurs largement utilisé et fait l'objet de nombreuses recherches afin de l'optimiser.

L'idée qui prédomine, en concaténation d'unités, est de "coller" les uns aux autres des petits bouts de parole répondant à la consigne afin de produire le signal de parole attendu. Ces petits bouts de parole sont appelés des unités. Ces unités peuvent être de tailles diverses mais nous nous limiterons au cas des unités* de la taille du phonème. Autrement dit, nous travaillerons directement avec les informations fournies par le module de traitement linguistique.

*** Notions avancées**

En pratique, les unités utilisées sont souvent des diphones. Un diphone peut être caricaturé comme une unité de la taille du phonème mais à cheval sur deux phonèmes consécutifs. L'intérêt de ce type d'unité est qu'il matérialise bien les phénomènes de coarticulation, c'est-à-dire l'effet d'un phonème sur ses voisins.



diphone

Mais d'où viennent ces unités, ces phonèmes?

Au préalable, nous enregistrons une personne qui parle (qui lit un long texte par exemple) afin de disposer d'une grande quantité (une heure par exemple) de signal de parole (parole qui doit être naturelle et de bonne qualité). Ce signal est alors découpé en petits morceaux appelés unités. Ces unités sont ensuite stockées dans une base de données (sorte de réservoir informatisé).

Lors de la synthèse, il suffit de puiser, dans la base de données, les unités répondant le mieux à la consigne et de les concaténer (les coller) les unes aux autres. On obtient alors le signal de parole de synthèse qu'il suffit de diriger vers une sortie "son" pour pouvoir en apprécier le résultat.

Notons que lorsque les unités sélectionnées ne répondent pas tout à fait à la consigne, elles subissent une petite "chirurgie esthétique" appelée "traitement du signal" afin de "gommer" ces différences. C'est le rôle du bloc nommé "adaptation" .

Synthèse de parole à partir du texte VI

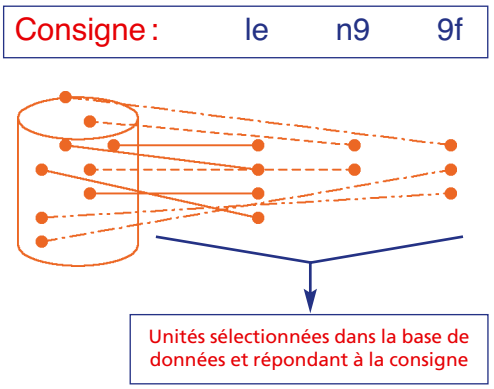
Text To Speech - TTS

Etapes de sélection d'unités : notions avancées

Le processus de sélection d'unités dans la base de données est un processus en deux étapes.

1. Etape de sélection des unités :

Cette étape consiste à retirer, de la base de données des unités, toutes les unités candidates (celles qui répondent à la consigne). Par exemple, si la consigne est de réaliser les diphones suivants : le n9 9f …, la première étape consistera à retirer de la base de données tous les diphones de ce type. Nous aurons donc plusieurs unités correspondant au diphone le, plusieurs unités correspondant au diphone n9 et ainsi de suite. C'est ce qui est représenté ci-dessous :

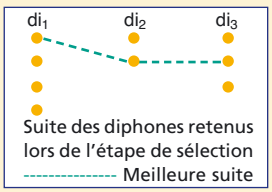


Coût de sélection

Un coût de sélection est calculé pour chaque unité sélectionnée. Ce coût est plus ou moins élevé selon que l'unité sélectionnée est plus ou moins éloignée de la consigne à réaliser.

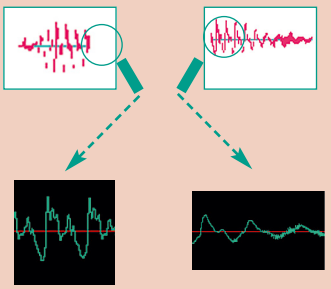
2. Etape de recherche de la meilleure suite d'unités :

L'étape suivante consiste à ne retenir que la meilleure suite d'unités. Deux critères interviennent. Tout d'abord, il faut que l'unité retenue soit la plus proche possible de l'unité à réaliser (consigne); c'est ce qui est caractérisé par le coût de sélection. D'autre part, il faut que les unités puissent être "collées" harmonieusement les unes aux autres afin d'éviter d'entendre des discontinuités dans la parole produite; c'est ce qui est caractérisé par le coût de concaténation. Cette opération revient finalement à trouver le meilleur chemin parmi les unités retenues afin de minimiser les coûts de sélection et de concaténation.



C'est ce qui est illustré ci-contre :

Coût de concaténation



Un coût de concaténation est également calculé entre les unités à "coller" afin que l'enchaînement se fasse de la manière la plus lisse possible et donne donc de la parole très naturelle.

Retour