Validation d'un processus de création de voix contextuelle et intégration de nouvelles langues à une application de synthèse vocale grand public

Camille Lecorgne

Validation d'un processus de création de voix contextuelles
&
Intégration de nouvelles langues à une application de synthèse vocale grand public

LECORGNE
Camille

UFR LLASIC

Mémoire de master 2 professionnel - 20 crédits – Sciences du langage
Spécialité Industries de la langue – Parcours TALEP
Sous la direction de Mme Véronique AUBERGÉ et Chiara MAZZA

Année universitaire 2014-2015

Validation d'un processus de création de voix contextuelles
&
Intégration de nouvelles langues à une application de synthèse vocale grand public

LECORGNE
Camille

UFR LLASIC

Mémoire de master 2 professionnel - 20 crédits – Sciences du langage
Spécialité Industries de la langue – Parcours TALEP
Sous la direction de Mme Véronique AUBERGÉ et Chiara MAZZA

Année universitaire 2014-2015

Remerciements

Je tiens tout d'abord à remercier ma tutrice chez Voxygen, Chiara, pour ses précieux conseils, sa bonne humeur et son implication dans son rôle de tutrice (et pour son thé).

Merci à Voxygen de m'avoir permis de vivre cette expérience professionnelle enrichissante. Merci aussi pour leur accueil chaleureux et leur bonne humeur.

Merci à Laure pour son écoute, pour m'avoir aidée à chaque fois que j'en ai eu besoin, pour avoir su se montrer patiente et m'avoir expliqué les choses encore et encore.

Merci à Sofiane pour le temps qu'il m'a consacré, pour ses explications riches et ses petits passages appréciés dans les bureaux.

Merci à Paul Bagshaw pour ses conseils, sa bienveillance, son humour et son accent british.

Merci à Noémie pour ses imitations, pour les premiers mois agréablement passés dans le même bureau et pour avoir pu admettre que la Bretagne c'est pas si terrible.

Merci à L'équipe Rennaise pour sa bonne ambiance, son mélange de langues et de cultures et pour les tablettes de chocolat.

Merci à L'équipe de Pleumeur-Bodou pour son accueil et sa sympathie.

Merci à Mme Aubergé pour m'avoir fait connaître la synthèse vocale et notamment Voxygen, pour m'avoir fait partager sa passion et son enthousiasme et pour nous avoir toujours encouragés.

Merci aux camarades robustes et élégants de Grenoble, en particulier Anaïs, mon binôme, pour son aide immense tout au long du master.

Merci, enfin, à ma famille, mes amis et Loïc qui m'ont soutenue et supportée tout au long de ce master et durant ce stage. Merci surtout à ma sœur pour sa présence, et parce qu'elle a su rendre ces six mois formidables.

UNIVERSITE Stendhal GRENOBLE 3

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : LECORGNE PRENOM : Camille

DATE : 01/09/2015 SIGNATURE [redacted]

Mise à jour mars 2013

Table des matières

Introduction

Aussi loin que remonte l'histoire de la synthèse vocale, un de ses enjeux majeurs à toujours été d'en optimiser sa qualité, Voxygen s'attelle donc aujourd'hui à poursuivre ce travail. Basée sur la concaténation de segments de parole préalablement enregistrés, la technologie Voxygen utilise un système de synthèse par corpus (aussi appelée synthèse à partir du texte ; text-to-speech) qui permet de vocaliser n'importe quelle entrée textuelle et dont la plus petite unité sélectionnée est le diphone. La qualité d'une voix de synthèse étant rigoureusement liée au corpus enregistré, il est primordial de le constituer de façon à ce qu'il renvoie un niveau de synthèse optimal.

Pour obtenir une synthèse de haute qualité, l'entreprise a mis en place la création de ce qu'elle appelle des voix contextuelles, conçues pour vocaliser des corpus limités à un domaine restreint constitués de parties fixes, à restituer telles quelles, et de parties variables. Pour restituer correctement les parties variables et pouvoir couvrir l'ensemble des combinaisons possibles tout en assurant des concaténations propres avec les parties fixes, des outils de création de corpus ont été mis en place. D'autre part, toujours dans un souci de qualité, Voxygen travaille en particulier avec une unité appelée sandwich vocalique qui permet de protéger en son sein des phonèmes impropres à la concaténation.

Le procédé de création de voix contextuelles ayant déjà été expérimenté avec succès sur le français, la mission de stage qui m'a été confiée avait pour objectif la validation du processus dans un ensemble de domaines applicatifs et dans différentes langues, mais aussi son application éventuelle à des langues en cours de développement. Ce travail présente donc comment nous avons pu vérifier ce processus sur différentes langues, dans quelles mesures il varie selon les langues et les domaines d'application et en quoi les unités sélectionnées impactent sur le niveau de synthèse.

Aussi, nous présenterons dans un premier temps l'entreprise Voxygen, ses différents domaines d'action, puis nous nous intéresserons à la synthèse vocale en générale et à la technologie utilisée par Voxygen. Dans un deuxième temps nous nous pencherons particulièrement sur le sandwich vocalique comme unité et son influence sur la qualité de la synthèse. Dans un troisième temps nous montrerons l'importance du choix des corpus selon différents points de vue, et enfin, en quatrième partie, nous évoquerons les notions de pluridisciplinarité et de plurilinguisme, que j'ai choisies pour qualifier mon stage chez Voxygen, illustrées par les travaux effectués.

Voxygen, présentation de l'entreprise et de sa technologie de synthèse vocale

I. Voxygen

L'entreprise Voxygen est une Start-up fondée en 2011 grâce à un essaimage de la part des membres de l'équipe d'Orange Labs en synthèse vocale, anciennement France Telecom R&D, qui constitue aujourd'hui l'équipe dirigeante de l'entreprise. Elle emploie environ 30 personnes entre son siège à Pleumeur-Bodou, et ses bureaux à Rennes, à Paris, ainsi qu'au Sénégal. Ses principaux concurrents sont Nuance (États-unis) et Acapela (Belgique).

Le travail que nous allons présenter ici a été mené dans l'équipe Recherche & Développement de Voxygen, branche qui concentre la majeure partie de son activité, dans ses bureaux rennais qualifiés de département de Développement linguistique.

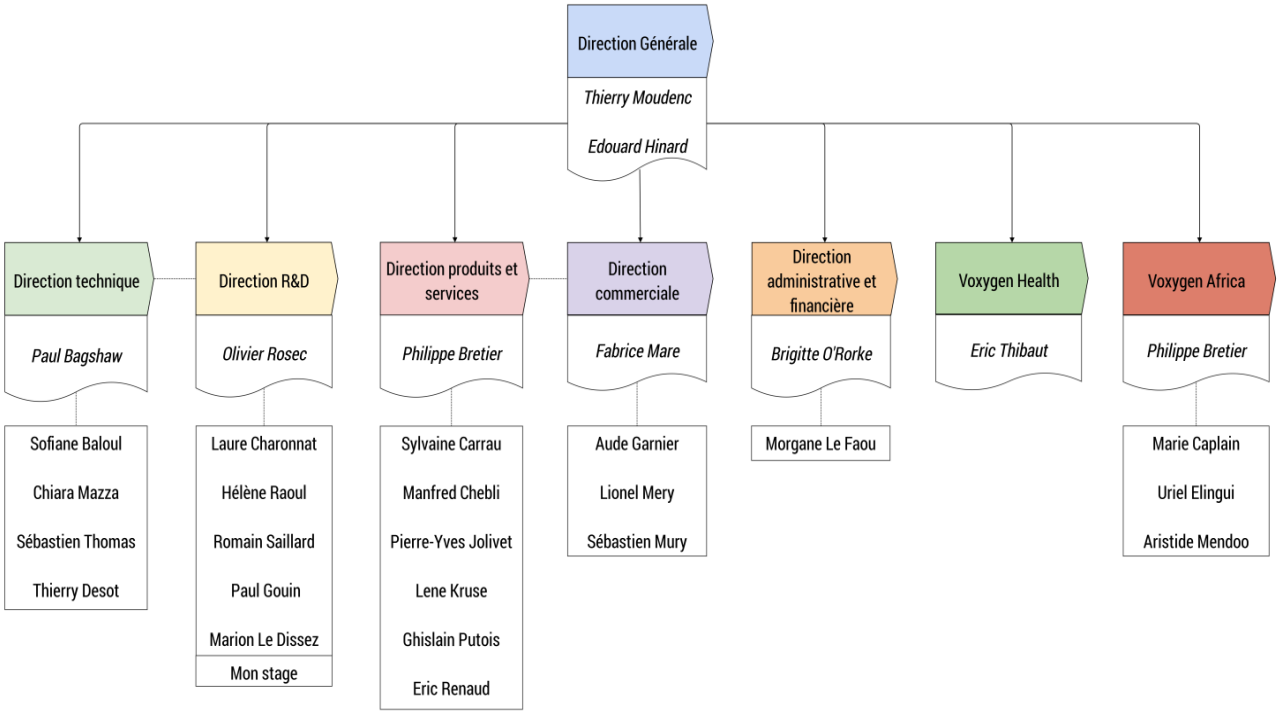


Figure 1 : Organigramme de l'entreprise

Après 20 ans de recherche dans les laboratoires France Télécom et Orange, l'équipe a pu mettre en place une technologie de synthèse vocale basée sur la sélection d'unités pour la création de voix multi-expressives et multi-lingues. La technologie utilisée par Voxygen est capable de vocaliser n'importe quelle entrée textuelle, nous verrons plus loin grâce à quels procédés.

L'entreprise propose plusieurs produits et services, autant pour les particuliers que pour les professionnels, et est notamment présente chez des groupes tels que Météo France, EDF ou la SNCF. Pour cela, elle offre un catalogue de langues varié comprenant plus de 100 voix dont 70 voix professionnelles, une dizaine de voix à accent et 25 voix fun. La couverture en termes de langue est relativement large puisqu'il existe des voix françaises, anglaises (britanniques et américaines), italiennes, espagnoles, arabes, allemandes et prochainement portugaises et néerlandaises. De plus, les recherches mèneront prochainement au développement de plusieurs langues africaines (Wolof, Pulaar, etc.).

Voxygen propose également :

- un service de création de voix sur mesure qui permet à n'importe qui de demander n'importe quel type de voix dans une des langues disponibles,
- un webreader qui permet de vocaliser les sites web en temps réel,
- l'expressive speech studio qui permet de produire des messages vocaux n'importe quand avec toutes les voix disponibles,
- l'expressive speech cloud qui permet au client de vocaliser ses données directement avec les voix hébergées par Voxygen, etc.

II. La synthèse vocale

La synthèse vocale, ou text-to-speech (TTS) en anglais ; littéralement « du texte à la parole », est une technologie permettant de vocaliser n'importe quelle donnée à la seule condition qu'elle soit de nature textuelle. On passe donc du texte à un signal de parole. Pour cela, deux étapes de traitements sont nécessaires :

- L'étape de traitements linguistiques (ou hauts-niveaux),
- L'étape de traitements acoustiques (ou bas-niveaux).

La première étape consiste à analyser le texte d'entrée pour lui attribuer des descriptions linguistiques et prosodiques puis à le convertir en une séquence phonétique destinée, d'une part, à établir les unités qui constitueront la cible, et d'autre part, à être enregistrée par un locuteur. L'analyse linguistique intervient après une étape de pré-traitements qui permet de lisser le texte en réécrivant des abréviations, des nombres ou des acronymes. Par exemple : svp > s'il vous plaît, 1920 > mille neuf cent vingt, etc. Cette étape terminée, une analyse morpho-lexicale est réalisée, permettant, grâce à un lexique et à des règles, de décomposer les mots inconnus présents dans le texte d'entrée en morphèmes.

L'étape lexicale n'étant pas suffisante pour lever toutes les ambiguïtés, on effectue alors une analyse syntaxique fonctionnant à base de règles grammaticales décrites en amont, afin d'attribuer à chaque terme une étiquette morpho-lexicale.
Exemple :

La	-->	Déterminant, article défini, singulier, féminin
fille	-->	Nom, singulier, féminin
marche	-->	Verbe, indicatif présent, 3ème personne, singulier
dans	-->	Préposition
la	-->	Déterminant, article défini, singulier, féminin
rue	-->	Nom, singulier, féminin

Grâce à ces différentes analyses, une description prosodique du texte est donnée. Cette description apporte des informations relatives au rythme d'élocution ainsi qu'à l'intonation. Aussi, cette dernière est déterminée en fonction de ce que l'on qualifiera de groupe de souffle, une séquence phonétique délimitée par deux pauses. Nous verrons plus loin que selon les langues les groupes de souffles ne sont pas traités de la même manière (partie 3_I).

C'est alors qu'intervient la deuxième étape de traitement, l'étape de traitements acoustiques. Ici, il s'agit de vocaliser le texte, on produit donc un signal de parole à partir de la séquence phonétique engendrée par les traitements linguistiques. C'est à cette étape acoustique que les procédés divergent selon les méthodes de synthèse utilisées. Nous présenterons donc succinctement ces différentes techniques puis nous nous attarderons plus particulièrement sur celle utilisée par Voxygen dans la partie suivante.

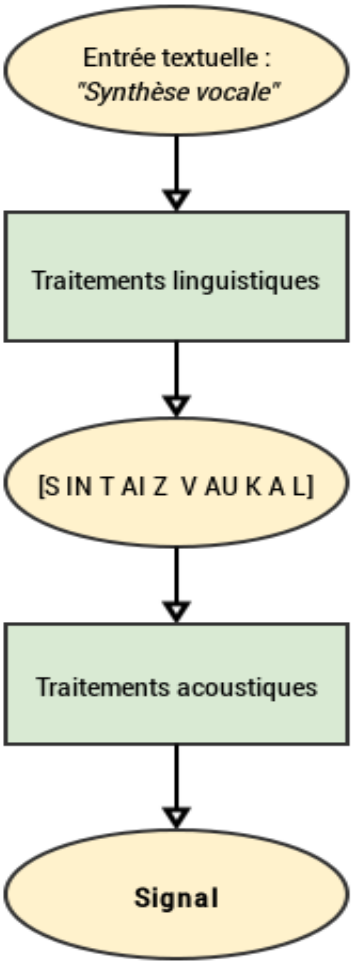


Figure 2 : schéma général de la synthèse vocale à partir du texte

- La synthèse articulatoire : on peut dire que cette méthode de synthèse descend de la fameuse machine parlante de Von Kempelen qui consistait à produire des sons grâce à des mécanismes imitant le conduit vocal humain. En effet, la synthèse articulatoire produit un signal de parole en imitant le fonctionnement de ce dernier, à la seule différence qu'elle en fait une reproduction numérique. Malheureusement, les processus de calcul étant tellement lourds au vu des résultats peu satisfaisants obtenus, cette technique n'est plus utilisée aujourd'hui.

- La synthèse par règles (ou synthèse par formants) : cette méthode de synthèse consiste à décomposer le signal de parole en différents paramètres (amplitude, fréquence et bande passante) qui sont ensuite utilisés pour calculer, grâce à des règles, la trajectoire temporelle des formants qui composent ce signal. Mais cette technique est aujourd'hui

peu usitée du fait des voix robotiques et peu naturelles obtenues.

- La synthèse par HMM : La synthèse par modèles de Markov cachés (Hidden Markov Models) apparaît comme l'évolution de la synthèse par formants. En effet, elle est régie par des règles issues d'une représentation formantique du spectre. La différence réside cependant dans la complexité et la puissance de calcul nécessaires, beaucoup plus importantes que dans la synthèse par formants. Ici, la représentation formantique est beaucoup plus complète puisqu'elle fait appel à des paramètres plus nombreux, et d'autre part, les règles de prédiction de trajectoire des formants sont remplacées par des algorithmes d'apprentissage très puissants. Ce sont ces apprentissages qui font la différence. En effet, on apprend des modèles HMM à des arbres de décision qui les regroupent selon leurs caractéristiques acoustiques et des facteurs contextuels. Les traitements linguistiques effectués en amont permettent d'obtenir des probabilités de trajectoires en fonction du contexte. Cette technique de synthèse a longtemps été la plus prometteuse grâce à sa flexibilité. En effet, cette méthode se contente d'un corpus réduit et possiblement bruité pour renvoyer une synthèse intelligible. Toutefois, le manque de naturel des voix obtenues pousse encore les méthodes à évoluer.

- La synthèse par concaténation de diphones : cette technique de synthèse consiste à construire une base des diphones présents dans la langue, sans tenir compte des paramètres prosodiques. Le principe de cette technique est donc de concaténer les diphones phonétiquement attendus puis, dans un deuxième temps, de lisser la prosodie en ajustant la hauteur et l'intensité de la voix. (Cette technique est présentée dans la partie suivante 1_III).

- La synthèse par sélection de diphones : à la différence de la synthèse par concaténation de diphones, cette technique, utilisée par Voxygen et présentée dans la partie suivante 1_III, consiste à construire une base des diphones présents dans la langue en tenant compte de la prosodie directement dans les diphones sélectionnés.

III. La technologie utilisée par Voxygen

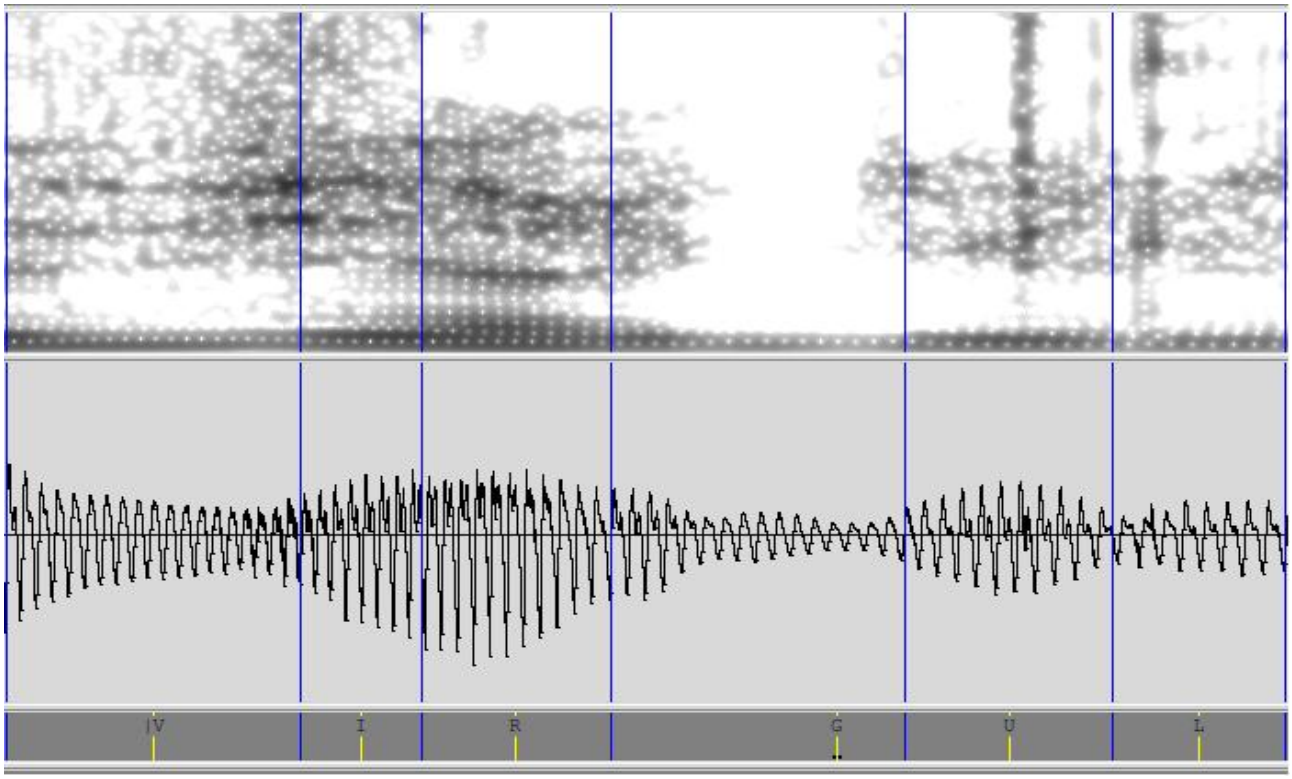
Nous allons, dans cette partie, nous intéresser à la technique de synthèse la plus usitée aujourd'hui puisqu'elle est aussi la plus naturelle, celle qui restitue le mieux la parole et la plus fidèle à la voix d'origine ; la synthèse par concaténation, qui, au gré de ses évolutions est aujourd'hui appelée synthèse par sélection d'unités ou synthèse par corpus. C'est donc naturellement que l'ancienne équipe France Télécom, aujourd'hui Voxygen, a choisi d'orienter ses recherches et de construire son architecture selon cette méthode.

Pour commencer, nous allons nous pencher sur la technique de synthèse considérée comme le point de départ de la synthèse par sélection : la synthèse par concaténation de diphones. Cette technique de synthèse a vu le jour principalement pour deux raisons. Premièrement, les chercheurs se sont rendu compte de l'importance des transitions entre phonèmes et de l'impact qu'elles pouvaient avoir sur la perception du signal (intelligibilité, naturel). Deuxièmement, les méthodes de synthèse utilisant des règles ou imitant l'appareil phonatoire pour construire des sons ne permettent pas de fabriquer un signal suffisamment naturel et intelligible.

De ce deuxième constat est née l'idée de construire des bases de données sonores directement à partir des enregistrements d'un locuteur. Les unités acoustiques utilisées sont donc issues de parole réelle. Le signal est alors fabriqué en concaténant les diphones présents dans la base, selon le texte produit. Le choix du diphone comme unité de sélection est venu du premier constat évoqué plus haut ; l'impact des transitions entre les phonèmes. En effet, le diphone s'étendant du milieu (partie stable) du phonème courant au milieu du phonème suivant, on prend appui sur les zones stables des phonèmes, ce qui facilite les concaténations.

Malheureusement, les problèmes engendrés par cette technique sont principalement liés à la prosodie restituée par le signal. En effet, en utilisant une unité acoustique si courte, la prosodie obtenue par concaténation est très monotone. On applique donc une prosodie calculée par des algorithmes de traitements du signal, mais le résultat obtenu ne renvoie pas l'image d'une parole naturelle et suffisamment intelligible. De plus, le fait d'utiliser le diphone engendre des discontinuités acoustiques pour chaque concaténation et une mauvaise prise en compte des phénomènes de coarticulation portant sur plus d'un diphone, renvoyant une synthèse artificielle et parfois difficilement intelligible. [Cadic, D., 2011]

La figure ci-dessous présente le mot «virgule» découpé en phonèmes. En bleu les marqueurs de frontières de phonèmes et en jaune les marqueurs de milieu de phonèmes. A partir de cette figure, nous pouvons illustrer un découpage en diphones :



#-V | V-I | I-R | R-G | G-U | U-L | L-#

Figure 3 : découpage d'un signal (| représente le marqueur de milieu de phonème ou plus précisément, sa zone stable, sur laquelle la concaténation est faite.)

Au vu du manque de cohérence et de naturel lié à la synthèse par concaténation de diphones, la nécessité de prendre en compte des unités plus longues a rapidement émergé. Les études se sont donc orientées vers l'utilisation des triphones, quadriphones, syllabes ou même de mots entiers, mais un problème de volume des enregistrements et des corpus nécessaires a empêché les techniques d'aboutir. C'est alors que l'idée d'utiliser des unités de longueurs variables voit le jour. C'est cette direction que Voxygen à choisi de suivre.

Bien qu'elle reste une technique de concaténation d'unités puisque son principe fondamental repose sur la juxtaposition d'unités, la synthèse utilisée par Voxygen est plutôt appelée synthèse par sélection d'unités ou synthèse par corpus. En effet, la plus petite unité pouvant être sélectionnée reste le diphone pour sa restitution des transitions entre phonèmes, ce qui empêche des

concaténations sur des zones non stables, mais le fait de sélectionner des unités de longueurs variables implique d'être en mesure de sélectionner l'unité la plus adaptée au contexte attendu. Plus précisément, après la constitution d'un corpus d'enregistrement, un locuteur enregistre un nombre de phrases conséquent et pertinent dans une langue donnée afin de couvrir potentiellement tous les sons présents dans la langue, ce qui constitue la base de données sonores du système, en somme, le point de départ de la synthèse vocale. La particularité de Voxygen est sa spécialisation dans la création de voix expressive. Pour cela, elle intègre à son processus de création de voix des règles de spécificité prosodique telles que la position syllabique, l'intonation montante ou descendante, la structure syllabique, etc, qui permettent l'utilisation d'une même unité dans différents contextes, tout en garantissant une prosodie naturelle. Chaque unité est représentée dans des contextes linguistiques et prosodiques différents. La difficulté réside donc, pour l'algorithme de sélection utilisé, dans la quantité de données à traiter et les choix à effectuer, tant au niveau linguistique, qu'aux niveaux phonétique et prosodique. L'algorithme est donc construit de façon à sélectionner les unités en fonction de ce que l'on appelle le coût cible et le coût de concaténation selon les unités présentes dans la base de données. Nous expliquerons plus précisément dans la partie 2_I de ce travail en quoi consiste le coût de concaténation, nous allons ici nous pencher sur la notion de coût cible.

Comme évoqué plus haut, le système de synthèse utilisé par Voxygen prend comme unité élémentaire le diphone. Le coût cible attribué aux unités est donc basé sur les diphones contenus dans la base. Pour chaque entrée textuelle proposée, plusieurs candidats potentiels pour chaque unité sont présélectionnés en fonction de leurs caractéristiques linguistiques, prosodiques et syntaxiques et de leurs caractéristiques transitoires (c'est sur ces caractéristiques liées à la concaténation que le coût de concaténation est attribué), ce qui constitue ce que l'on appelle un treillis d'unités.

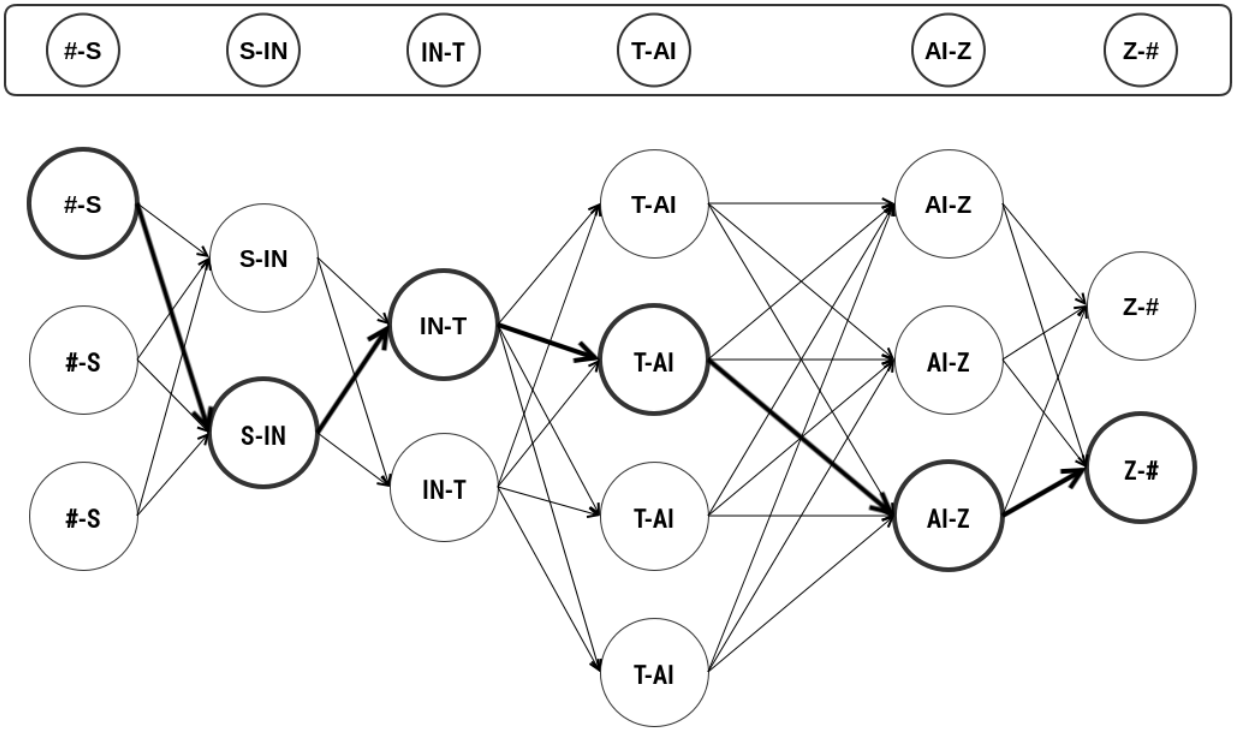


Figure 4 : exemple d'un treillis d'unités (diphones) pour la synthèse du groupe de souffle « Synthèse ». En gras le chemin le moins coûteux, emprunté par l'algorithme de sélection.

Le coût cible est alors attribué à chaque candidat du treillis en fonction de l'écart entre ses caractéristiques et celles de la cible attendue et le coût de concaténation en fonction des distorsions plus ou moins importantes engendrées, que nous détaillerons dans la partie 2_I de ce travail. Évidemment, deux diphones présents côte à côte dans la base, se verront attribuer un coût de concaténation nul (ce qui ne sera pas forcément le cas du coût cible), il est donc possible de trouver des unités beaucoup plus longues que le diphone selon le contenu de la phrase d'entrée. Théoriquement, il serait tout à fait possible de trouver dans la base une phrase entière, ce qui renverrait un signal de synthèse totalement naturel puisqu'extrait tel quel de la base. L'intérêt d'un tel système de synthèse, qui permet de sélectionner des unités de longueurs variables en attribuant ces coûts, est évidemment de pouvoir sélectionner les unités les plus longues possibles afin de limiter les distorsions liées aux concaténations (les phénomènes de coarticulation sont donc mieux pris en compte) et de favoriser telle ou telle unité en fonction de ses caractéristiques prosodiques. Ceci permet de construire un signal de parole synthétique de meilleure qualité et relativement naturel grâce à la prise en compte de la prosodie directement dans la sélection des unités.

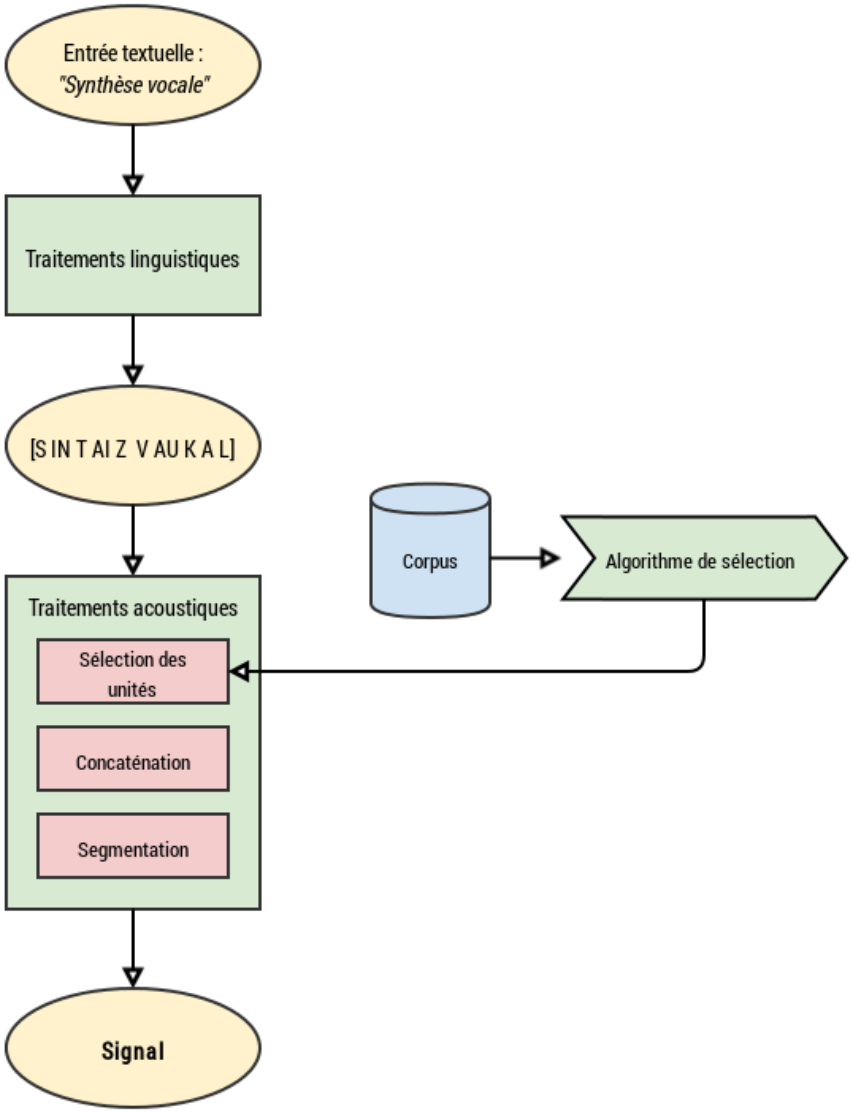


Figure 5 : schéma de la synthèse vocale par corpus

Cette technique n'est donc plus une utopie aujourd'hui grâce aux mémoires informatiques toujours plus importantes et aux puissances de calcul des processeurs actuels. En effet, les besoins sont réels en matière de mémoire afin de pouvoir stocker un volume très important d'enregistrements et les algorithmes utilisés nécessitent une puissance de calcul très élevée. La qualité de la synthèse étant directement liée à la taille du corpus, il est nécessaire de pouvoir couvrir le plus grand nombre possible d'unités. En effet, et c'est là que notre travail tend à se concentrer, plus le corpus est vaste et varié, plus les unités sélectionnées sont potentiellement longues et adaptées au contexte, et plus la qualité de la parole de synthèse sera potentiellement bonne.

C’est donc toujours dans un souci de qualité que l’entreprise Voxygen met en place la création de voix appelées “Slot’N’Fill” que nous détaillerons en troisième partie de ce travail et qu’elle introduit l’unité appelée “sandwich vocalique”, qui fera l’objet de la deuxième partie de ce travail.

Pour une meilleure compréhension des parties suivantes, voici l'alphabet phonétique de Voxygen du français, de l'américain et de l'espagnol, avec l'équivalence de chaque phonème dans l'Alphabet Phonétique International (API).

Alphabet phonétique français :

API	Voxygen	API	Voxygen
p	**P**	w	**W**
t	**T**	ɥ	**H**
k	**K**	a	**A**
b	**B**	i	**I**
d	**D**	ɔ	**O**
g	**G**	o	**AU**
f	**F**	ə	**E**
s	**S**	ø	**EU**
ʃ	**CH**	œ	**OE**
v	**V**	ɛ	**AI**
z	**Z**	e	**EI**
ʒ	**J**	u	**OU**
m	**M**	y	**U**
n	**N**	œ̃	**UN**
ɲ	**NJ**	ɛ̃	**IN**
l	**L**	ɔ̃	**ON**
ʁ	**R**	ɑ̃	**AN**
j	**Y**		

Alphabet phonétique US :

API	Voxygen	API	Voxygen
p	**p**	l	**l**
t	**t**	l̩	**l=**
k	**k**	r	**r**
b	**b**	j	**y**
d	**d**	hw / w	**w**
dˤ	**d^**	iː	**ii**
g	**g**	i	**i**
tʃ	**ch**	ɛ	**e**
dʒ	**jh**	ɑː	**aa**
s	**s**	æ	**a**
z	**z**	ə	**@**
ʃ	**sh**	ɜ	**@@**
ʒ	**zh**	ɜ	**uh**
f	**f**	uː	**uu**
v	**v**	ʊ	**u**
θ	**th**	ɔː	**oo**
ð	**dh**	ɒ	**o**
h	**h**	eɪ	**ei**
m	**m**	aɪ	**ai**
m̩	**m=**	ɔɪ	**oi**
n	**n**	oʊ	**ou**
n̩	**n=**	aʊ	**au**
ŋ	**ng**		

Alphabet phonétique espagnol :

API	Voxygen	API	Voxygen
a	**A**	k	**K**
e	**E**	g	**G**
i	**I**	f	**F**
o	**O**	z	**Z**
u	**U**	s	**S**
a :	**AA**	ʝ	**Y**
e :	**EE**	x	**X**
i :	**II**	ʧ	**CH**
o :	**OO**	m	**M**
u :	**UU**	n	**N**
w	**W**	ɲ	**NN**
j	**J**	l	**L**
p	**P**	ʎ	**LL**
b	**B**	ɾ	**R**
t	**T**	r	**RR**
d	**D**		

Le sandwich : fragile à l'intérieur, robuste à l'extérieur

I. Définition

Avant d'entrer réellement dans une définition précise de la notion de sandwich, nous devons donner une précision phonémique. En effet, Il est important d'introduire deux notions avant de passer à la suite : le phonème fragile et le phonème robuste.

Les phonèmes sont tous différents et présentent tous des variantes de prononciation, d'ouverture, de voisement, etc. Il est donc important de les classifier en amont, afin de bien comprendre leurs mécanismes, permettant ou non la concaténation. En effet, tous les phonèmes ne se comportent pas de la même manière à la synthèse et la concaténation n'aura pas le même impact sur tous les phonèmes. D'autre part, la classification des phonèmes diffère selon les langues, il est donc primordial d'adapter cette classification.

La classification des phonèmes se fait en fonction de leurs caractéristiques phonétiques et articulatoires : voisement, ouverture, lieu d'articulation. En fonction de ces caractéristiques, on peut définir leur facilité de concaténation, selon laquelle on applique un principe de coût appelé «coût de concaténation». Le coût de concaténation consiste à attribuer un coût entre 1 et 10 à chaque phonème, 1 étant attribué aux phonèmes robustes, 5 aux liquides et au /e/ muet et 10 aux phonèmes fragiles, pour permettre à l'algorithme de sélection des unités de toujours choisir le «chemin» le moins coûteux. Ce coût est donc calculé en fonction des «*distorsions acoustiques qui résulteraient d'une concaténation*» [Cadic, D., 2011]. Ainsi, l'algorithme fera en sorte de n'effectuer des concaténations que sur des phonèmes non coûteux, les phonèmes dits «robustes», et d'épargner les phonèmes dits «fragiles».

Durant sa thèse au sein de l'équipe Recherche et Développement du groupe France Télécom, [Cadic, D., 2011] a réparti les classes phonétiques du français selon leur fragilité et leur robustesse

(présentées des plus robustes aux plus fragiles) :

- *les occlusives sourdes [p] [t] [k]*
- *les autres consonnes sourdes [f] [s] [ʃ]*
- *les occlusives voisées [b] [d] [g]*
- *les consonnes nasales et fricatives voisées [m] [n] [v] [z] [ʒ]*
- *les liquides [l] [ʁ]*
- *les semi-voyelles et le schwa [j] [w] [ɥ] [ə]*
- *les voyelles [a] [ɔ] [o] [ɛ] [e] [ø] [œ] [ɑ̃] [ɔ̃] [ɛ̃] [œ̃] [i] [y] [u]*

Les dénominations données par Voxygen sont légèrement différentes mais représentent les mêmes classifications :

- Consonnes plosives sourdes : [P] [T] [K]
- Consonnes plosives voisées : [B] [D] [G]
- Consonnes fricatives voisées : [Z] [J] [V]
- Consonnes fricatives sourdes : [S] [CH] [F]
- Consonnes nasales : [M] [N]
- Consonnes liquides : [L] [R]
- Semi-voyelles : [Y] [W]
- Voyelles orales : [A] [O] [AU] [AI] [EI] [EU] [OE] [I] [U] [OU] [UI]
- Voyelles nasales : [AN] [ON] [IN] [UN]
- /e/ muet (ou schwa) : [E]

Ces classements répertorient les phonèmes du français selon leur capacité à supporter ou non la concaténation. Ainsi, les consonnes, en générale, rendent une concaténation propre, on leur attribue donc un coût de concaténation de 1, ce sont ces phonèmes qui sont considérés comme les plus robustes. Les liquides, les semi-voyelles et le /e/ muet supportant moins bien la concaténation, ont un coût de 5 et enfin les voyelles qui ont un coût de concaténation de 10. Les phonèmes considérés comme fragiles sont donc les voyelles, même si les liquides et les semi-voyelles sont considérées comme fragiles selon *«le contexte phonétique et le niveau de protection envisagé»* [Cadic, D., 2011].

Les phonèmes fragiles sont considérés comme étant « impropres » à la concaténation. En effet, leur caractère trop lisse, pas assez bruité, ne permettent pas une concaténation propre.

Typiquement, les frontières de ces phonèmes sont marquées très nettement, ce qui rend difficile une transition (concaténation) avec un autre phonème. De plus, les formants, caractéristiques des voyelles, sont très marqués sur le signal, présentent des différences de hauteur et varient d'intensité à plusieurs niveaux ce qui rend la concaténation très difficile. En effet, pouvoir obtenir une concaténation propre impliquerait de trouver un phonème de hauteur et d'intensité égales. En concaténant sur des phonèmes fragiles, le risque est de provoquer ce que l'on appelle des « sauts » dans le signal, ce qui renverrait une synthèse désagréable et surtout très robotique, parfois presque inintelligible.

D'autre part, à la différence des phonèmes fragiles, les phonèmes dits « robustes » sont tout à fait adaptés à la concaténation puisque leur environnement sonore est souvent bruité avec des frontières beaucoup moins nettes ou bien plosif et comporte donc une occlusion. En effet, dans un environnement bruité, la concaténation et ses effets seront moins audibles, car protégés par le bruit. De plus, à la différence des voyelles, les phonèmes plosifs (par exemple /t/, /p/) sont caractérisés par une occlusion puis une plosion, sans différence de hauteur, ce qui facilite et permet de définir spécifiquement l'endroit où la concaténation se fera : dans le silence au moment de l'occlusion. Les phonèmes robustes sont donc plus adaptés à la concaténation puisque le bruit et les frontières floues protègent leurs transitions des effets indésirables audibles.

Maintenant que les notions de fragilité et de robustesse des phonèmes ont été introduites, nous allons nous pencher, dans un premier temps sur ce que représente le sandwich et ce qu'il prend en compte en essayant de le définir de façon précise.

Comme évoqué plus haut dans la première partie de ce travail, le diphone ne constitue pas une unité suffisamment robuste pour l'utiliser de façon systématique dans la synthèse par concaténation. Le «sandwich vocalique» (que nous appellerons ensuite «sandwich» pour plus de simplicité) est donc introduit comme unité phonétique nouvelle par [Cadic, D., 2011].

Cette unité désigne *«toute séquence de phonèmes fragiles entourées par deux phonèmes robustes»* [Cadic, D., 2011]. Le sandwich est donc un ensemble de phonèmes fragiles, présents les uns à la suite des autres dans un corpus, et dont les deux extrémités (premier et dernier phonèmes) sont des phonèmes robustes. Les phonèmes robustes servent à protéger de la concaténation les phonèmes fragiles contenus dans l'ensemble. L'expression régulière suivante montre, logiquement, de quelle façon est considéré le sandwich :

$$C\ (W^*\ V\ W^*)^+\ C$$

où :

C désigne l'ensemble des phonèmes robustes, y compris le silence,

W désigne l'ensemble des phonèmes fragiles, hormis les voyelles,

V désigne l'ensemble des voyelles,

** désigne la présence zéro, une ou plusieurs fois de l'unité qu'il quantifie,*

+ désigne la présence une ou plusieurs fois de l'unité qu'il quantifie.

On se rend donc bien compte ici qu'un sandwich contient au moins une voyelle, d'où l'adjectif vocalique, ce qui permet de le différencier du cluster consonantique, qui lui ne contient que des consonnes.

Chouette

CH W AI T

Ce sandwich illustre exactement l'expression régulière donnée par [Cadic, D., 2011] puisque nous pouvons noter la présence de deux phonèmes robustes [CH] et [T] (représentés par le C dans l'expression régulière) protégeant des phonèmes fragiles [W] (également représenté par le W dans l'expression régulière) et [AI] (représenté par le V dans l'expression régulière). Nous pouvons également noter la présence de la semie-voyelle [W] qui engendre un phénomène de coarticulation sur le phonème suivant [AI]. Il est donc primordial, pour l'intelligibilité et le naturel de la synthèse, de protéger ce phonème de la concaténation.

Ainsi, si le sandwich contient au minimum une voyelle, on peut en déduire qu'il peut en contenir plusieurs et même *«traverser les frontières de mots»* [Cadic, D., 2011] comme dans les exemples ci-dessous (en considérant les liquides comme fragiles) :

Je venais finir ton travail.

J EU V EU N AI F I N I R T ON T R A V A Y

Les sandwiches [# J EU V], [N AI F], [N I R T] et [T ON T] traversent les frontières de mots.

Il veut parler au ministre.

I L V EU P A R L EI AU M I N I S T R

Le sandwich [L EI AU M] traverse non seulement deux frontières de mots mais compte également deux voyelles dans son ensemble. De plus, on peut remarquer ici la présence d'un cluster consonantique [S T R], ce qui nous permet de voir clairement la différence entre ces deux unités.

Le sandwich constitue donc une unité plus importante, en termes de taille, que les unités utilisées auparavant (diphone, triphone...), présentées dans la première partie de ce travail. On cherche à sélectionner l'unité la plus longue possible en faisant bien attention à ne pas couper le sandwich sur un phonème fragile. Néanmoins, si la sélection d'un sandwich est impossible, des unités plus courtes telles que le diphone (unité minimale) peuvent être sélectionnées.

II. Motivation de son utilisation et sélection des unités

Comme abordé plus tôt, le diphone est l'unité représentative minimale utilisée dans la synthèse par concaténation mais elle n'est pas suffisante, d'une part, pour couvrir tous les contextes linguistiques et prosodiques d'un corpus, et d'autre part, sa petite taille engendre des discontinuités phonétiques dérangeantes. Les corpus d'enregistrements construits pour la synthèse sont souvent composés de milliers de phrases qui peuvent supporter une couverture par des unités plus longues que le diphone comme le triphone, la syllabe ou même le mot ou le sandwich. Nous allons ici nous intéresser à cette dernière unité.

Cette unité, de longueur variable en fonction des corpus et des parties à couvrir, permet une couverture des unités en contexte. En effet, chaque unité est alors caractérisée non seulement par son contenu phonétique, linguistique et prosodique mais aussi par son environnement phonétique, linguistique et prosodique.

Outre les questions contextuelles, l'intérêt d'une telle unité réside dans sa capacité à protéger les phonèmes fragiles des concaténations. En effet, ce coût étant directement corrélé à la qualité de la synthèse, on peut facilement affirmer que le fait d'utiliser le sandwich réduira ce coût de concaténation et permettra donc d'appréhender une synthèse de meilleure qualité.

Par ailleurs, la sélection des unités dans le processus de synthèse dépendant de la constitution

du corpus et donc de la couverture de la langue, il est nécessaire, pour que le système sélectionne des sandwiches, que ceux-ci soient présents dans la base de données. De plus, comme évoqué plus haut, les unités ne doivent pas seulement remplir des contraintes phonétiques dans leur contenu mais également des contraintes phonétiques, linguistiques et prosodiques dans leur environnement. Pour cela, il est préférable, surtout dans le cas de Voxygen qui met un point d'honneur à proposer des voix multi-expressives, que les unités soient toujours multi-représentées.

L'utilisation du sandwich vocalique comme unité est, en conséquence, motivée par sa capacité à protéger et anticiper les discontinuités phonétiques et prosodiques liées aux nombreuses concaténations nécessaires à la synthèse, en protégeant les parties fragiles par des parties robustes. De plus, le sandwich permet, de par son inclusion des phonèmes fragiles en son milieu, de mieux prendre en compte les phénomènes de coarticulation le plus souvent engendrés par les voyelles, les semi-voyelles et les liquides, phonèmes considérés justement comme des phonèmes fragiles.

III. Mise en pratique

A. Mavoa

Pour illustrer le principe de prise en compte du sandwich dans la sélection des unités et en démontrer concrètement l'utilité et l'importance, nous allons nous pencher sur un des travaux effectués durant ce stage : l'écriture de scénarios pour l'application Mavoa.

Figure 6 : Page d'accueil de l'application Mavoa

Commençons par exposer le principe de l'application en question : Mavoa. Mavoa est une application mise en place par un stagiaire étudiant en informatique, employé par Voxygen en 2014 et qui permet à l'utilisateur, par l'intermédiaire d'un site internet, de s'enregistrer sur un corpus court (3 ou 4 phrases) pour entendre, une fois l'enregistrement terminé, la synthèse d'une phrase non prononcée, et ce, avec sa propre voix. La vocation de cette application, exposée à la Cité des Sciences de Paris, est de faire la démonstration de la technologie de synthèse vocale de Voxygen.

Figure 7 : Page de choix des scénarios en fonction des langues

Figure 8 : Page d'enregistrement des phrases

La mise en place de cette application repose sur plusieurs étapes. Dans un premier temps, il s'agit d'écrire, à la main et non automatiquement, des scénarios (ce qui constituera les phrases à enregistrer : le corpus) en fonction de la phrase de synthèse que l'on souhaite renvoyer à l'utilisateur. Cette étape étant la pièce maîtresse de l'illustration de l'importance des sandwiches, nous y reviendrons plus tard dans cette partie.

Lorsque l'utilisateur a terminé d'enregistrer le corpus, apparaissent à l'écran les phrases enregistrées ainsi que la phrase de synthèse enrichies de segments de couleurs montrant à l'utilisateur de quelle façon la phrase a été construite, en fonction de quelles parties du corpus. Pour cela, il convient d'écrire un script informatique afin d'afficher correctement les différents segments. Nous expliquerons plus en détail le principe de construction et de fonctionnement de ces scripts dans la dernière partie (partie 4_II_C) de ce travail.

Enfin, la troisième étape, certainement la plus importante, consiste à mettre en place les outils nécessaires à la création du dictionnaire qui permettra de synthétiser la phrase souhaitée.

Pour tout processus de création de voix, il convient de créer un dictionnaire afin de construire la synthèse. Mais pour construire ce dictionnaire, le système doit faire appel à ce que l'on appelle une BDS_REF (base de données de signal segmenté de référence). La particularité de cette

application réside dans le fait qu'un dictionnaire est créé automatiquement à chaque enregistrement de l'utilisateur, et ce donc, pour chaque scénario. De plus, l'application Mavoa propose des scénarios dans différentes langues (français, anglais britannique, anglais américain, italien et espagnol), il est donc nécessaire de créer des dictionnaires différents selon les langues. Une BDS_REF est alors créée pour chaque scénario. Précisons toutefois que la création de la BDS_REF est un peu particulière dans ce cadre. En effet, dans le cas d'une création de voix «normale», une BDS_REF, qui représente les données théoriques du système de synthèse vocale « Baratinoo », est générée automatiquement en fonction de la langue et du script locuteur, puis une fois les enregistrements réalisés, la BDS locuteur, qui représente les données « réelles », et le dictionnaire sont créés. Ici, au contraire, On utilise seulement une BDS_REF pour chaque scénario et le dictionnaire est calculé directement à chaque enregistrement.

La technologie de synthèse de Voxygen fonctionne grâce au système de synthèse vocale appelé «Baratinoo» et au module de création de dictionnaire «Sound2Dico». Pour créer une BDS_REF, il est donc indispensable de tenir ces deux systèmes à jour. Et pour que cette BDS_REF soit en état de fournir les informations nécessaires, il faut créer ce que l'on appelle un fichier xml de voix qui spécifie tous les paramètres de la voix dans tous les modules de Baratinoo. En admettant que ces pré-requis soient respectés, la BDS_REF peut être créée. A partir des fichiers textes des phrases à enregistrer, on fait appel au module txt2enr de Sound2Dico pour créer ce qu'on appelle les fichiers «enr», ainsi qu'au module txt2phn pour générer les fichiers «phn».

- enr : les fichiers enr sont des fichiers d'enrichissement qui agissent sur la prosodie. En effet, dans ces fichiers sont répertoriées, pour chaque phrase contenue dans les fichiers txt et pour chaque phonème de ces phrases, de façon numérique, des informations sur :
    - l'environnement phonétique de chaque phonème (phonème précédent et phonème suivant)
    - la position syllabique
    - la structure syllabique
    - le marqueur mélodique
    - la valeur de la fréquence fondamentale à la marque de milieu de phonème
    - la durée (en milliseconde) de la marque début à la marque milieu de phonème
    - la durée (en milliseconde) de la marque milieu à la marque fin de phonème.

Ces informations sont communes à toutes les langues mais chaque langue dispose cependant d'informations supplémentaires que nous ne détaillerons pas ici.

- phn : les fichiers phn sont des fichiers donnant la transcription phonétique des phrases contenues dans les fichiers txt à partir desquels ils sont générés.
    - Des fichiers de référence sont également nécessaires à la création des dictionnaires :
    - la liste des phonèmes de la langue avec leur type (voyelle, consonne), leur lieu d'articulation, leur mode de production (voisé ou non) et le coût de concaténation attribué
    - un fichier caractérisant la voix
    - la liste de tous les diphones du corpus les uns à la suite des autres avec leur identifiant numérique unique

Grâce à ces fichiers, les dictionnaires pourront être créés automatiquement à chaque enregistrement.

Nous allons maintenant nous pencher plus spécialement sur la création du corpus : les scénarios. Comme évoqué plus haut, Mavoa intègre des scénarios dans plusieurs langues différentes. Ici, le travail d'écriture confié durant le stage a porté sur l'espagnol et le français, nous nous attacherons donc à décrire des procédés mis en place dans ces langues.

Le processus de création de scénarios se fait «à la mano», aucune phase n'est automatisée dans la constitution des corpus Mavoa. L'objectif est donc de choisir la phrase qui sera synthétisée avec la voix de l'utilisateur et écrire un corpus de trois ou quatre phrases contenant les meilleures unités pouvant être sélectionnées pour la synthèse (l'idéal étant de sélectionner des sandwiches pour obtenir une synthèse de meilleure qualité), en respectant des contraintes diverses :

- les segments (au minimum des diphones) doivent être découpés sur des phonèmes robustes
- doivent avoir la même position syllabique (début, milieu, fin) dans la phrase à enregistrer et dans la cible
- doivent apparaître dans le même contexte (par exemple, devant une virgule, en début de phrase, etc) dans la phrase à enregistrer et dans la cible
- pour l'espagnol, par exemple, il est également conseillé de se soucier du type de voyelle. En effet, il existe des voyelles longues et des voyelles courtes et une voyelle courte ne pourra pas être remplacée par une voyelle longue et inversement. (ce qui n'est pas valable en français par exemple puisque la durée des voyelles n'a pas de qualité phonétique dans la représentation des phonèmes de Voxygen).

Dans l'idéal, il serait intéressant de pouvoir contrôler les valeurs des enr de chaque unité sélectionnée pour pouvoir les faire coïncider avec la cible, mais cela ne serait possible qu'après enregistrement des phrases et représenterait un travail très fastidieux.

Mis à part ces contraintes syntaxiques, les segments sélectionnés doivent évidemment renvoyer la même phonétique que les segments de la cible. En effet, si le processus d'écriture n'est pas automatique, le processus de synthèse, lui, l'est. Il s'agit donc de «prédire» les meilleures unités afin que le système Baratinoo puisse sélectionner ce qui synthétisera la phrase cible dans les phrases du corpus. Une des difficultés de ce travail d'écriture a été de construire à la main des phrases contenant les bonnes unités dans une langue étrangère, ce qui ralenti sans doute le travail. Mais la principale difficulté a été de respecter les contraintes citées ci-dessus. En effet, les contraintes sur la position syllabique, le contexte phonétique ou encore le type de voyelle rendent le travail compliqué et réduisent les possibilités. Cependant, agissant sur la prosodie de l'énoncé elles sont très importantes dans la synthétisation de l'expression, point d'honneur des travaux de l'entreprise.

De plus, et c'est le point crucial de notre exposé, le découpage des unités est l'enjeu majeur du fonctionnement de l'application. Comme déjà explicité, la plus petite unité sélectionnée est le diphone (milieu du phonème courant jusqu'au milieu du phonème suivant), mais des unités plus longues peuvent être prises en compte, par exemple, le sandwich. Dans le travail d'écriture de scénarios pour Mavoa, l'intérêt est d'écrire des phrases contenant les segments les plus longs possibles potentiellement «concaténables» dans la synthèse. En effet, le corpus étant court, il est beaucoup plus intéressant de sélectionner des unités longues pour obtenir une synthèse de bonne qualité. Si seul le diphone était sélectionné, les concaténations seraient beaucoup trop nombreuses et la synthèse pas assez naturelle. Ici, le sandwich permet de restituer une parole beaucoup plus naturelle puisque les unités sont directement issues du signal enregistré.

Voici trois exemples de scénarios espagnols qui illustrent bien la notion de sandwich (notons toutefois que la notion de sandwich dans le cadre de Mavoa est un peu particulière puisque nous nous permettons d'y inclure des phonèmes robustes. L'intérêt ici est de sélectionner des unités longues, commençant et se terminant par des phonèmes robustes afin de protéger d'éventuels phonèmes fragiles de la concaténation) :

¿ Puedo imitar tu voz, oyes ?

Puedo imaginar tu voluntad.
¿ El hombre sabe imitar también, oyes ?
Tiene una hermosa voz, como Miguel.

Digo cosas que nunca has dicho.

Digo con mis palabras lo que has dicho.
Hay cosas que no sabes de mi.
Se que nunca has distribuido los papeles.

Quien vive sin pensar, no puede decir que vive.

Dices que vives en Pensacola ?
Quien viaje sin prensar, puede descubrir.
Sin informaciones, no puedo predecir que vive.

Le système de synthèse étant basé sur le diphone, chaque sandwich est coupé sur un milieu de phonème et le sandwich suivant commencera sur la deuxième moitié du même phonème. Par exemple, dans le premier scénario nous avons [T U BB] puis [BB OO Z]. Nous pouvons bien nous rendre compte sur la figure ci-dessous, représentant le résultat qu'obtient l'utilisateur après enregistrement, que les phonèmes sont coupés au milieu.

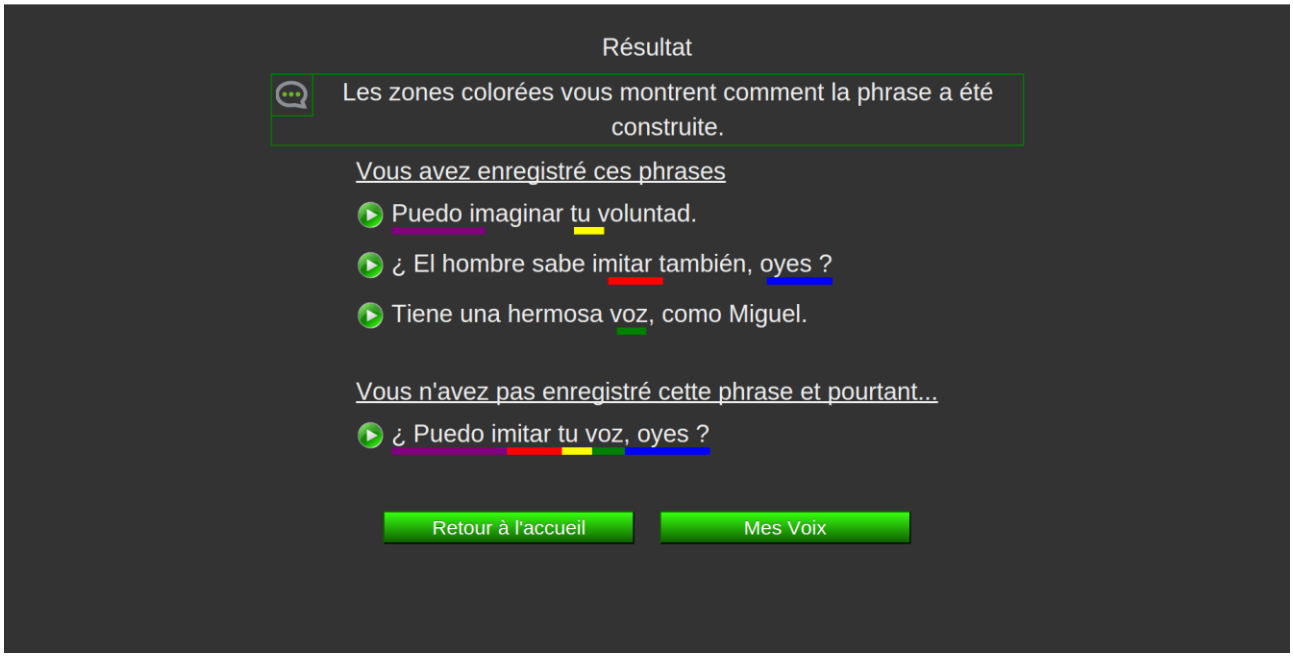


Figure 9 : Résultats des enregistrements et de la synthèse. Affichage du découpage et de la sélection des unités.

Nous pouvons voir dans ces scénarios qu'aucun diphone n'est sélectionné seul mais que plusieurs sandwiches sont sélectionnés à la suite pour permettre au système de concaténer des unités présentes côte à côte dans la base, ce qui restitue une parole complètement naturelle. Nous pouvons également remarquer que les contraintes syntaxiques ont été respectées pour permettre de renvoyer une prosodie naturelle et sans artefact. Par exemple, «Puedo» est sélectionné dans le même contexte (début après silence) que l'unité attendue dans la cible.

De plus, nous pouvons préciser que les phonèmes du français et ceux de l'espagnol ne sont, par définition, pas les mêmes, les phonèmes considérés comme fragiles sont donc également différents. En effet, par exemple, en espagnol le [R] est considéré comme fragile, ce qui n'est pas le cas en français. D'où l'importance, dans la phrase «¿ El hombre sabe imitar también, oyes ?», de couper le sandwich [T AA R T] sur le t de «también» et non sur le «r» de «imitar».

Voici la représentation des phonèmes de l'espagnol :

- Voyelles orales [A] [E] [I] [O] [U] [AA] [EE] [II] [OO] [UU]
- Semi-voyelles [W] [J]
- Consonnes plosives sourdes [P] [T] [K]
- Consonnes plosives voisées [B] [D] [G] [BB] [DD] [GG] [YY]
- Consonnes fricatives sourdes [F] [Z] [S] [X] [CH]

- Consonnes fricatives voisées[Y] [V] [ZZ] [SS]
- Consonnes nasales [M] [N] [NN] [NG]
- Consonnes liquides [L] [LL] [R] [RR]

Avec un coût de concaténation de 5 sur les semi-voyelles, 10 sur les voyelles, 1 sur les consonnes. Cependant quelques coûts sont différents du français : en espagnol, sur les consonnes nasales, le [NN] et le [NG] ont un coût de 1 mais le [M] et le [N] ont un coût de 5. Et en ce qui concerne les liquides, le [L] et le [LL] ont un coût de 5 mais le [R] et le [RR] ont un coût de 10 et sont donc considérés comme des phonèmes fragiles. De plus, on peut remarquer aux doubles graphèmes que l'espagnol comporte beaucoup de phonèmes allongés ([AA], [BB], [ZZ], etc), ce qui n'est pas le cas dans la représentation du français.

Une barrière technique se fait néanmoins ressentir dans le processus de synthèse de Mavoa. En effet, il est impossible d'affirmer, et justement parce que le corpus de Mavoa est construit à la main, que les unités que nous attendions pour construire la cible sont bien celles qui ont été sélectionnées par le système Baratinoo. Le seul moyen de vérifier cette sélection est de réaliser la synthèse en interne et de vérifier les valeurs d'enr. Par ailleurs, si l'on décide d'effectuer cette vérification, ce que nous avons fait pour l'espagnol, et qu'il s'avère que le système n'a pas sélectionné ce que nous attendions, un des moyens d'arriver à ses fins est de «tricher» en changeant les valeurs d'enr qui empêchent le système de sélectionner telle ou telle unité. Voici un exemple de cas qui nous a obligé à modifier les valeurs d'enr :

¿ Puedo imitar tu voz, oyes ?
Puedo imaginar tu voluntad.
¿ El hombre sabe imitar también, oyes ?
Tiene una hermosa voz, como Miguel.

Nous avions une coupure sur le «r» de «imaginar» alors que nous attendions une coupure sur le «t» de «también» pour obtenir la sandwich [T AA R T]. En calculant les coûts cibles et les coûts de concaténation, nous nous sommes rendu compte que le problème venait du coût cible sur l'enr de la position syllabique. En effet, en espagnol, une règle pénalise la sélection d'une syllabe initiale d'un mot de plusieurs syllabes pour la synthétiser sur un mot monosyllabique. Nous avons donc changé la valeur de l'enr du «t» de «también» pour forcer le système à prendre cette unité en la considérant comme monosyllabique. L'autre solution que nous avons testée pour forcer le système

à sélectionner les unités que nous attendions à tout simplement été d'exclure, dans les fichiers phn, les phonèmes non désirés en les notant d'un point d'exclamation.

#_|_P_W_*EE_DD_O_|_I_M_A_X_I_N_*AA_!R_|_T_U_|_BB_O_L_U_N_T_*AA_DD_#

B. SNF

La notion de sandwich étant un peu particulière dans le cas de l'application Mavoa puisqu'il s'agit plus de chercher à couvrir les unités les plus longues possibles, nous allons maintenant nous pencher sur le travail de création de voix contextuelle réalisé sur le français et l'anglais américain. A la différence de l'application Mavoa, la création du corpus et la sélection des unités dans cette tâche ont été réalisées de façon totalement automatique, nous pourrons donc réellement apprécier la recherche de qualité par l'utilisation du sandwich dans un processus automatique.

L'objectif des voix contextuelles, ou Slot'N'Fill (SNF), ou synthèse de phrases à trous, est de construire des voix de très haute qualité. Le principe est simple, en se situant dans un domaine précis et restreint (ici, par exemple, nous avons travaillé sur des phrases destinées à donner des informations au conducteur d'une voiture à propos de sa vitesse, de sa consommation, de l'heure, etc.), on se rend compte que certaines parties des phrases reviennent très fréquemment, l'objectif est donc de «séparer» ces parties fréquentes, appelées des parties fixes, des parties variables, toujours différentes. Dans le cas des phrases pour l'automobile, toutes les parties variables sont des données numériques. Voici quelques exemples de phrases :

Vous roulez à la vitesse de 3 kilomètres par heure.
Vous roulez à la vitesse de 4 kilomètres par heure.
Vous roulez à la vitesse de 5 kilomètres par heure.
Vous roulez à la vitesse de 6 kilomètres par heure.

et ainsi de suite... Nous pouvons bien nous rendre compte ici de la redondance des parties fixes et de «l'évolution» des parties variables. Les parties fixes sont répertoriées, sous une forme phonétique spécifique dans un fichier xml de type PLS (Prononciation Lexicon Specification, W3C) que nous détaillerons dans la partie 3_I de ce travail, et découpées sur le dernier phonème robuste avant la partie variable ou sur le premier phonème robuste après la partie variable.

«Vous roulez à la vitesse de 3 kilomètres par heure.»

Cette phrase contient deux parties fixes :

- *«Vous roulez à la vitesse de»* phonétisé ainsi dans le PLS : 7##7##7##7##7##DEU"
- *«kilomètres par heure.»* phonétisé ainsi dans le PLS : K7##7##7

(Ici, le symbole /7/ représente n'importe quelle suite de phonèmes, et /##/ représente une frontière entre deux mots.)

Nous pouvons voir les phonèmes robustes, surlignés, sur lesquels sont coupées les phrases. On peut donc considérer ces parties fixes comme des sandwiches puisqu'elles sont de longueur variable et protègent des phonèmes de la concaténation. L'intérêt de cette méthode est de couvrir seulement les parties variables ainsi que le diphone (ou plus dans le cas de /DEU"/ par exemple) de concaténation entre la partie fixe et la partie variable. En effet, les parties fixes sont enregistrées par le locuteur telles quelles et ne subiront aucune concaténation lors de la synthèse puisqu'elles ressortiront telles qu'elles ont été enregistrées. Cette méthode confère à la synthèse une plus grande amplitude et un gain sur le naturel de la prosodie grâce à la réduction des discontinuités liées aux concaténations. Une fois que ce document est édité, un corpus contenant toutes les variantes possibles (tous les nombres) de chaque phrase est généré. C'est à partir de ce corpus qu'est construit le script locuteur qui constitue un condensé de toutes les unités à couvrir et qui devra être lu par ce dernier lors de l'enregistrement. Un algorithme est chargé de trouver toutes les unités à couvrir, dans tous les contextes syntaxiques et prosodiques nécessaires et renvoie un fichier de statistiques illustrant la quantité de diphones, triphones, clusters consonnantiques et sandwiches présents dans le script. (Nous expliquerons plus précisément en partie 3_I de ce travail de quelle manière et suivant quels procédés le script condensé est obtenu). On peut alors noter que le sandwich tient une place très importante dans la couverture de ce genre de corpus restreint et redondant. Nous pouvons également affirmer que le sandwich est une promesse de qualité de la synthèse puisqu'il est ici sélectionné spécialement dans le cadre d'une création de voix de très haute qualité.

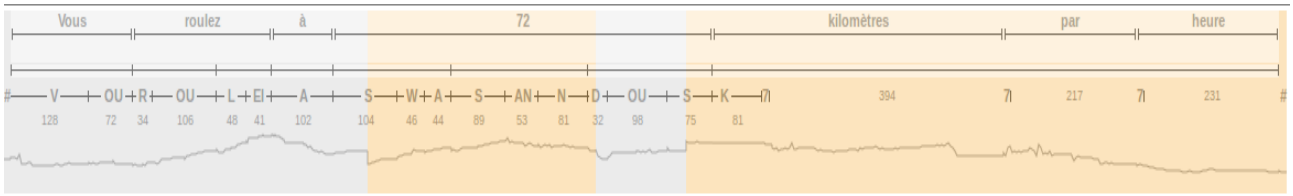


Figure 10 : synthèse d'une phrase pour l'automobile (SNF) en français avec découpage en unité sélectionnées.

Nous pouvons voir ici les différentes unités sélectionnées dans le signal, grâce à l'alternance des couleurs.

Nous avons pu voir dans cette partie dans quelles mesures l'utilisation du sandwich était importante dans la restitution d'une synthèse de qualité. De plus, les travaux réalisés durant ce stage ont permis de mettre en évidence la particularité de cette unité à plusieurs niveaux. A partir de ce constat, nous pouvons nous interroger sur les mesures mises en œuvre pour constituer les corpus, et par projection les unités, qui représenteront le mieux la langue selon le domaine d'application. C'est ce qui constituera la troisième partie de ce travail.

De l’importance du choix des corpus

La création d’une nouvelle voix, et plus loin encore, d’une nouvelle langue, dans un système de synthèse vocale impliquent la constitution préalable de ce que l’on appelle un corpus. Ce corpus très volumineux est constitué d’un ensemble de phrases correspondant à la représentation d’une langue, et dans le cas d’une voix dédiée à une application, il représente le domaine complet. A partir de ce corpus, un script condensé est mis en place ; il correspond à ce que l’on souhaite couvrir dans une langue et une application données et est ensuite mis en forme pour l’enregistrement. Cependant, ces deux étapes de création diffèrent selon le type de synthèse, la cible que l’on vise, le domaine applicatif.

Nous allons donc voir dans cette partie comment, selon le type de synthèse recherché, les corpus sont construits et les scripts de lecture optimisés. Nous verrons donc, dans un premier temps, trois procédés différents appliqués à trois applications différentes. Nous expliquerons dans le même temps la notion d’entonnoir dont nous avons choisi de qualifier le processus de création de corpus et de scripts condensés. Dans un deuxième temps, nous montrerons en quoi la création d’une voix se joue en deux dimensions : la dimension phonétique et la dimension prosodique. Enfin, en troisième partie, nous montrerons, à travers les travaux réalisés, comment la cible prévue pour la synthèse influe sur la construction du corpus et du script condensé.

I. Différents procédés pour différentes mises en oeuvre

La vocation première de Voxygen est de créer des voix dites «complètes», capables de vocaliser n’importe quelle entrée textuelle dans une langue donnée, on parle de synthèse «Full TTS». L’objectif d’une telle mise en oeuvre est de pouvoir couvrir tout le vocabulaire de la langue, dans tous les contextes envisageables. Pour cela, la création de ce que l’on appelle un corpus est nécessaire. Ce corpus est construit à partir de différentes sources :

- sous-titres de films,
- SMS,

- romans,
- journaux,
- chroniques historiques,
- wikipédia...

Deux obstacles se présentent néanmoins :

- la difficulté à trouver des sources qui soient nombreuses, variées et faisant partie du domaine publique,
- la difficulté à «nettoyer» les corpus.

En effet, les journaux, bien que constituant une immense source dans des langages divers et variés et relativement accessibles, ne permettent pas, par exemple, d'obtenir de données écrites à la première ou à la deuxième personne. C'est pourquoi des sources de type conversationnel sont introduites.

Par ailleurs, les corpus recueillis n'étant pas utilisables tels quels, il est nécessaire de les «nettoyer». Ce travail de nettoyage consiste à normaliser le format d'encodage des textes (UTF-8), à lisser le texte en ajoutant des balises de mises en forme, à supprimer d'éventuelles fautes d'orthographes, etc.

Le but de ce processus est d'obtenir un corpus assez grand (plusieurs millions de phrases) pour représenter potentiellement toute la langue, il est donc primordial de varier les sources et ainsi créer le plus d'effets de langage possibles. Cependant, il est nécessaire de préciser qu'il ne s'agit que d'une «représentation». En effet, la langue étant un concept infini, présentant un nombre infini de variantes, il serait impossible de la représenter dans sa totalité. Cela impliquerait de pouvoir enregistrer toute la langue et la définition même de la synthèse vocale par concaténation n'aurait plus lieu d'être.

La deuxième étape du processus de création de voix complète est le «parsing». Cette étape consiste à passer le corpus de phrases dans un module de parsing de Baratinoo qui va analyser la structure morphosyntaxique du texte, le découper en groupes de souffle et attribuer à chacun des statistiques de couverture et des valeurs prosodiques. Ces groupes de souffle sont définis différemment selon les langues. Par exemple, en français, ils sont définis par la ponctuation, ainsi,

le parsing découpera le texte à chaque fois qu'il rencontrera un point, une virgule, un point d'interrogation, etc. En anglais, par contre, les groupes de souffle sont découpés grâce à un arbre de décisions appris sur des corpus annotés manuellement car le découpage systématique sur des ponctuations comme en français n'est pas envisageable.

Ensuite, pour chaque groupe de souffle, le parsing va donner toutes les séquences de diphones, sandwiches, clusters consonantiques et triphones présentes dans le corpus en attribuant à chacun des phonèmes une suite d'ENR (voir partie 2_III_A) et seulement aux voyelles, ce que l'on appelle un contexte. Le contexte de chaque voyelle est prédéfini dans des fichiers dits de «regroupement» qui permettent, pour chaque langue et selon la cible que l'on souhaite atteindre, de réduire l'univers à couvrir (par conséquent le nombre de phrases). Les regroupements permettent de rassembler certaines unités ayant des caractéristiques proches. On n'agit ici que sur les voyelles car ce sont les phonèmes que l'on souhaite représenter le plus précisément pour les protéger des concaténations et ce sont également les phonèmes qui portent le mieux la prosodie. Par exemple, en français, on regroupe toutes les unités (voyelles) en fonction de leur position syllabique et de leur marqueur mélodique en ne les discriminant pas par leur structure syllabique. C'est-à-dire que pour un ensemble de phonèmes ayant la même position syllabique et le même marqueur mélodique, on estime qu'en en couvrant un seul, cela suffit, peu importe que la structure syllabique de ces phonèmes soit différente. Le corpus ainsi obtenu après le parsing est appelé la «pioche». On lui donne ce nom afin d'exprimer le phénomène d'extraction qui est réalisé. En effet, on «pioche» les phrases qui nous intéressent pour la couverture des unités afin de créer le script condensé.

La troisième étape du processus est l'extraction des statistiques, c'est l'une des étapes les plus importantes puisqu'elle nous informe sur les unités qui seront sélectionnées dans le script condensé. De plus, s'agissant des unités de couverture, cette étape illustre tout à fait le cheminement de ce travail. A partir de la pioche obtenue par le parsing, des statistiques sur la fréquence d'apparition des unités vont être calculées. Ces statistiques, donnant la fréquence d'apparition des sandwiches, des diphones, des clusters consonantiques et des triphones dans chaque groupe de souffle, vont permettre d'attribuer un «score» à ces groupes de souffle, c'est ce qui constitue la dernière étape du processus.

En effet, la dernière étape consiste à créer le script condensé (plusieurs milliers de phrases, 3500 en français par exemple). On impute donc, à chaque groupe de souffle un score en fonction des unités nouvelles qu'il apporte, et de l'importance que ces unités représentent selon ce que l'on

souhaite : unités rares, unités fréquentes, valorisation des sandwiches, valorisation des diphones, etc. Le score calculé exprime donc le potentiel de couverture de chaque groupe de souffle.

Par exemple, si l'on décide de donner plus d'importance aux unités fréquentes, plus un groupe de souffle contient des unités fréquentes, plus il aura un score élevé. En fonction de ces scores, on va pouvoir sélectionner les phrases candidates pour le script condensé grâce à un outil de validation renvoyant tous les groupes de souffle avec leurs unités restant à couvrir (cette étape de validation peut être réalisée automatiquement ou par l'intervention d'un opérateur). Cet outil permet néanmoins de refuser une phrase si elle ne paraît pas pertinente.
A partir du score de chaque groupe de souffle, on utilise deux méthodes de sélection :

- la méthode glouton : elle consiste à sélectionner d'abord les groupes de souffle les plus intéressants (selon les critères choisis), ceux ayant le score le plus élevé.
- la méthode cracheur : elle consiste à sélectionner d'abord les groupes de souffle les moins intéressants (selon les critères choisis), ceux ayant le score le plus faible.

Dans un premier temps on passe un glouton qui permet de couvrir d'abord les unités qui augmentent le plus la couverture puis un cracheur qui permet de couvrir les unités qui diminuent le moins la couverture. *«Une première phrase est sélectionnée selon un critère ; par exemple, le nombre d'unités à couvrir. Cette phrase fait alors partie de la couverture et ses unités sont retirées de l'ensemble des unités à couvrir. Le processus recommence ; la seconde phrase, dans cet exemple, comportera un maximum d'unités non-encore couvertes. Le processus s'arrête lorsque toutes les unités sont couvertes.» [François, H., 2002]*
«L'algorithme glouton commence avec une couverture vide. Les premières phrases amènent un recouvrement partiel, jusqu'à parvenir à la couverture totale des unités. La continuation de l'algorithme ne ferait qu'apporter de la redondance. L'algorithme cracheur part d'une couverture égale à l'ensemble des phrases, donc totale. Nous ne cherchons pas ici d'unités absentes de la base. Les phrases sont supprimées une à une jusqu'à ce que l'on arrive à un seuil où la suppression d'une phrase ferait nécessairement perdre la couverture totale des unités. La continuation de l'algorithme permet de répondre à des demandes de recouvrement partiel.» [François, H., 2002]
Le but de l'utilisation des deux méthodes est de ne pas se retrouver, à la fin du passage du glouton avec un corpus contenant les unités intéressantes plusieurs fois. Le passage du cracheur permet donc d'éliminer des phrases qui ne seraient pas utiles puisqu'elles ne diminueraient pas significativement la couverture. On donc peut facilement imaginer d'après cela que la couverture

des unités dans un processus de création de voix complète n'atteint jamais les 100%. En effet, comme mentionné plus haut, il est impossible de couvrir la langue dans son intégralité, le taux de couverture est donc toujours inférieur à 100%. Le script condensé représente donc finalement seulement un sous-ensemble de ce que l'on souhaite couvrir : la langue.

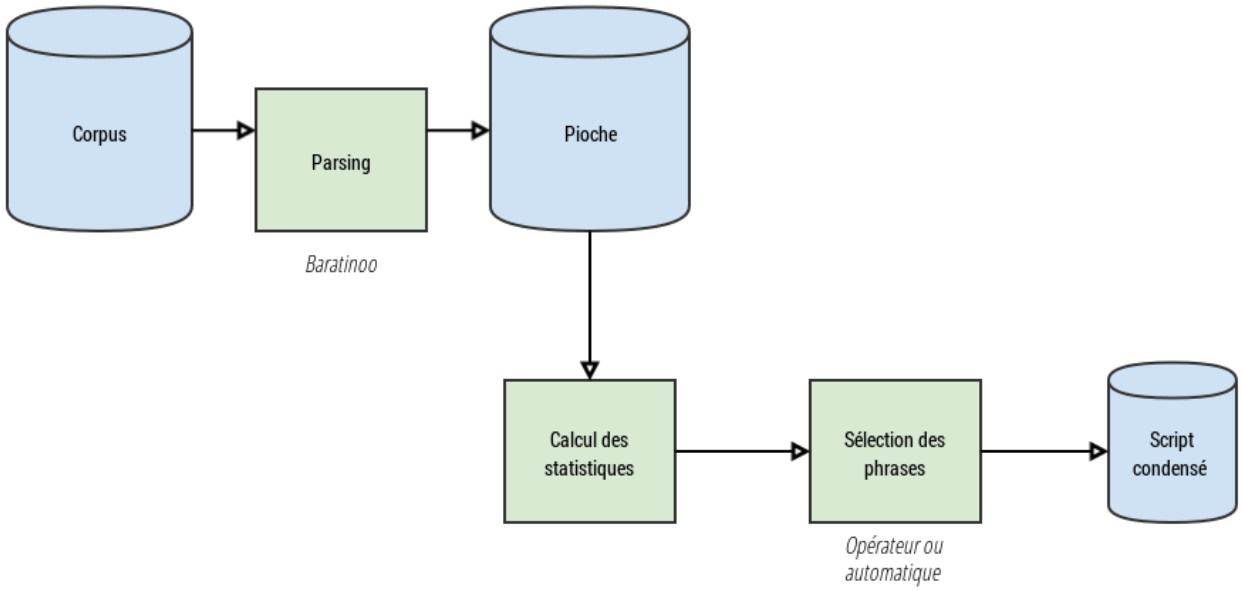


Figure 11 : schéma du processus de création du script condensé

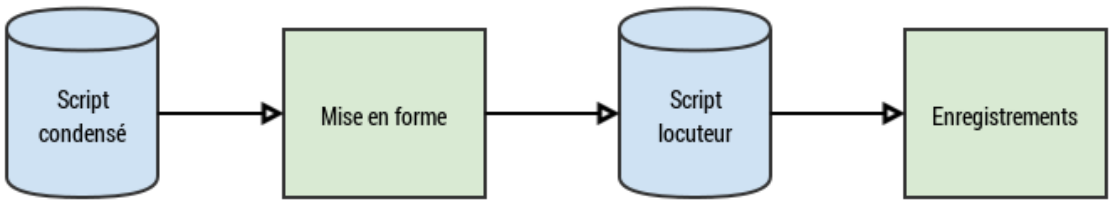


Figure 12 : schéma représentant le passage du script condensé au script locuteur.

Nous allons maintenant montrer en quoi les deux autres procédés de création de voix que nous avons expérimenté durant ce stage, Slot'N'Fill et Mavoa, différent du processus de création de voix complète.

Commençons par illustrer le procédé Slot'N'Fill ou voix contextuelles. Toujours dans un objectif de qualité, Voxygen met en place la création de voix contextuelles lorsque la demande s'étend sur un domaine restreint. Comme expliqué en partie 2_III_B, le principe du Slot'N'Fill (SNF) est de proposer une voix de très haute qualité en ne couvrant que les parties variables du corpus et en restituant les parties fixes telles qu'elles ont été enregistrées. Pour cela, comme pour la création d'une voix complète, il est nécessaire de construire un corpus, d'extraire des statistiques et de créer un script condensé. Cependant, la manière de passer par ces étapes est différente.

Première étape : la construction du corpus.

Lorsque dans un processus de création de voix complète, le corpus est construit à partir de sources extérieures et cherche à couvrir toute la langue, dans un processus de voix contextuelle, on cherche à couvrir seulement un domaine restreint, il est donc construit à partir de la demande du client et constitue un ensemble fini. A partir des phrases «commandées» par le client, on commence par définir les parties redondantes qui deviendront des parties fixes et on détermine ce qui constituera les parties variables. Comme évoqué en deuxième partie, les parties fixes sont répertoriées dans un fichier xml appelé PLS dans lequel on leur donne une transcription phonétique particulière (le sympole 7, qui permet au système de ne pas tenir compte de ces parties durant le parsing. Ce caractère a été choisi car il ne fait partie du jeu de phonèmes d'aucune langue, il ne provoque donc pas d'ambiguïté) ainsi qu'un attribut «meta» qui permet à chaque partie fixe d'avoir une identité propre. Voici un exemple de fichier PLS :

```
<lexeme>
    <grapheme>Votre vitesse moyenne sur le parcours est de</grapheme>
    <phoneme vox:meta="vitesse_moyenne">7##7##7##7##7##7##DEU"</phoneme>
</lexeme>
<lexeme>
    <grapheme>Votre consommation moyenne sur le parcours est de</grapheme>
    <phoneme
vox:meta="consommation_moyenne">7##7##7##7##7##7##DEU"</phoneme>
</lexeme>
<lexeme>
    <grapheme>Vous roulez à la vitesse de</grapheme>
    <phoneme vox:meta="vous_roulez_a">7##7##7##7##DEU"</phoneme>
```

</lexeme>
<lexeme>
<grapheme>mètres de votre destination</grapheme>
<phoneme vox:meta="metres_destination">M7##7##7##7</phoneme>
</lexeme>
<lexeme>
<grapheme>La distance restante est de</grapheme>
<phoneme vox:meta="distance_restante">7##7##7##7##DEU"</phoneme>
</lexeme>

On peut voir sur cet exemple en orange l'attribut «meta», différent pour chaque entrée et la transcription phonétique de chaque partie fixe (avec un /7/ par mot, séparés par /##/) en bleu avec le phonème de découpage, en violet, pour permettre la concaténation sur un phonème robuste.

Une fois que le PLS est complet, un script informatique en langagevpython est écrit. Il représente chaque phrase du corpus, et introduit des boucles sur toutes les parties variables pour permettre de générer tous les nombres souhaités. Un corpus représentant un ensemble fini de phrases est alors généré, c'est ce qui équivaut au corpus qui passera l'étape de parsing. Voici un exemple de script de génération de corpus.

```
for n in range(201):
  if n<2:
    print """Votre vitesse moyenne sur le parcours est de %d kilomètre par heure."""%n
  else:
    print """Votre vitesse moyenne sur le parcours est de %d kilomètres par heure."""%n
for n in range(1,20):
  if n<2:
    print """Votre consommation moyenne sur le parcours est de %d litre aux 100 kilomètres."""%n
  else:
    print """Votre consommation moyenne sur le parcours est de %d litres aux 100 kilomètres."""%n
      for m in range(1,10):
      print """Votre consommation moyenne sur le parcours est de %d,%d litres aux 100
```

```
kilomètres."""%(n,m)
for n in range(201):
  if n<2:
    print """Vous roulez à la vitesse de %d kilomètre par heure."""%n
  else:
    print """Vous roulez à la vitesse de %d kilomètres par heure."""%n
for n in range(50,1000,50):
  print """Vous êtes à %d mètres de votre destination."""%n
for n in range(2001):
  if n<2:
    print """La distance restante est de %d kilomètre."""%n
  else:
    print """La distance restante est de %d kilomètres."""%n
```

Ensuite, l'étape du parsing est la même que lors de la création d'une voix complète, à la différence près que l'étiquetage se fait uniquement sur les phonèmes des parties variables.

Exemple d'une phrase parsée : «*Vous roulez à la vitesse de 2 kilomètres par heure.*»

[u'# ', u'| [Vous]', u'7 6 72 4 5047d7b0', u'| [roulez]', u'7 6 72 4 5047d7b0', u'| [\xe0]', u'7 2 72 5 5047d7b0', u'| [la]', u'7 2 72 3 5047d7b0', u'| [vitesse]', u'7 2 72 3 5047d7b0', u'| [de]', u'D 2 3 5 5047d7b0', u'EU 2 3 5 5047d7b0', u'| [deux]', u'D 2 3 3 00000000', u'EU 2 3 3 00000000', u'| [kilom\xe8tres]', u'K 2 72 3 906578ac', u'7 2 72 3 906578ac', u'| [par]', u'7 5 72 2 906578ac', u'| [heure]', u'7 5 72 2 906578ac', u'# Vous roulez \xe0 la vitesse de 2 kilom\xe8tres par heure.', u'Vous roulez \xe0 la vitesse de 2 kilom\xe8tres par heure.', 0, [u'D_EU_D - 2NEU- -', u'D_EU_K - 2DIV- -'], [u'D_EU_D - 2NEU- -', u'EU_D_EU 2NEU- - 2DIV-', u'D_EU_K - 2DIV- -'], [u'# -', u'D -', u'D -', u'K -', u'# -'], [u'#_7 - -5047d7b0', u'7_D -5047d7b0 -', u'D_EU - 2NEU-', u'EU_D 2NEU- -', u'D_EU - 2DIV-', u'EU_K 2DIV- -', u'K_7 - -906578ac', u'7_# -906578ac -']]

La lecture de ce type de document n'étant pas importante pour la compréhension de la suite de ce travail, nous ne nous attarderons pas à l'expliquer dans le détail. On peut tout de même noter la transcription phonétique des sandwiches en bleu et de leur contexte en orange. Techniquement, les deux dernières étapes permettant de construire le script condensé sont également les mêmes que dans un processus de création de voix complète. La seule différence réside dans la couverture des unités. En effet, s'il est impossible de couvrir toute la langue et donc d'arriver à une couverture de

100% des unités dans la création d'une voix complète, c'est le but même de la création d'une voix contextuelle. La couverture des unités est donc très importante et les statistiques doivent toujours donner 100% de couverture pour pouvoir parler de SNF. De plus, dans le cas du SNF, le but premier étant la qualité, il est d'autant plus important d'avoir une couverture en sandwiches élevée là où elle est beaucoup moins importante dans une voix complète. En effet, en SNF, le plus important est de ne pas entendre de concaténations, l'utilisation du sandwich est donc de rigueur. C'est pourquoi, par exemple, un des critères pris en compte dans le calcul du score des groupes de souffle est la couverture des sandwiches. Les phrases les mieux notées sont donc celles contenant le plus de sandwiches. Dans un processus de voix complète, on attache beaucoup d'importance au diphone pour permettre au système de couvrir le maximum de contextes, il s'agit donc dans ce cas plutôt d'un compromis entre le diphone et le sandwich.

Enfin, le dernier processus de création de voix que nous avons pu mettre en place est celui destiné à l'application Mavoa. Dans ce cadre-ci, le procédé de création du corpus est complètement différent et on peut dire qu'il néglige quelques étapes pour plus de simplicité au vu de la qualité attendue. Ici, le but est de créer l'étonnement chez l'utilisateur en lui faisant enregistrer un minimum de phrases. Nous avons vu dans la partie 2_III_A de ce travail que le principe était de faire enregistrer des phrases à un utilisateur et, à partir de ces phrases, créer une phrase de synthèse avec la voix de l'utilisateur.

Dans le cadre de cette application, on ne différencie plus le corpus du script condensé puisqu'aucune étape intermédiaire n'est réalisée, nous parlerons donc seulement de corpus par la suite. Tout le travail de construction étant manuel, nous ne partons plus d'un corpus pour construire un script condensé mais de la cible (la phrase de synthèse) pour construire ce script. La différence avec les autres procédés automatiques est donc que la cible est restreinte (une seule phrase) et connue, il est alors possible de construire le corpus en prédisant les unités à sélectionner.

Comme expliqué dans la partie 2_III_A de ce travail, les phrases du corpus sont construites à la main, le travail consiste donc à trouver, dans trois ou quatre phrases, les bonnes unités qui seront concaténées pour former la phrase de synthèse. Ici, à la différence des autres procédés, le but n'est pas de chercher à couvrir tel ou tel type d'unité mais plutôt de chercher à couvrir le segment le plus long possible. En effet, les phrases étant peu nombreuses et enregistrées dans un environnement potentiellement bruité, il serait très difficile de ne sélectionner que de petites unités qui renverraient

une synthèse de mauvaise qualité. Ici, on cherche à créer un effet de surprise chez l'utilisateur tout en garantissant une qualité correcte. On choisit donc de créer des phrases desquelles on peut extraire des segments assez longs mais qui ne laissent pas l'utilisateur se douter de quelque chose avant même d'avoir entendu la synthèse. Il s'agit donc de trouver un juste milieu entre qualité et effet de surprise. Les segments sélectionnés doivent cependant toujours respecter les contraintes liées à la prosodie et aux concaténations, à l'image des autres procédés automatisés.

Dans les procédés développés plus haut, le corpus correspond à la «représentation» de ce que l'on cherche à couvrir dans la langue, au contraire, dans une application telle que Mavoa, la phrase de synthèse (la cible) «est» ce que l'on cherche à couvrir.

La notion d'entonnoir évoquée plus haut permet de donner un aspect imagé à ce que nous venons d'expliquer. En effet, plus le domaine est restreint, plus le corpus est restreint mais, au contraire, plus les unités sélectionnées sont potentiellement longues afin de garantir une synthèse de qualité.

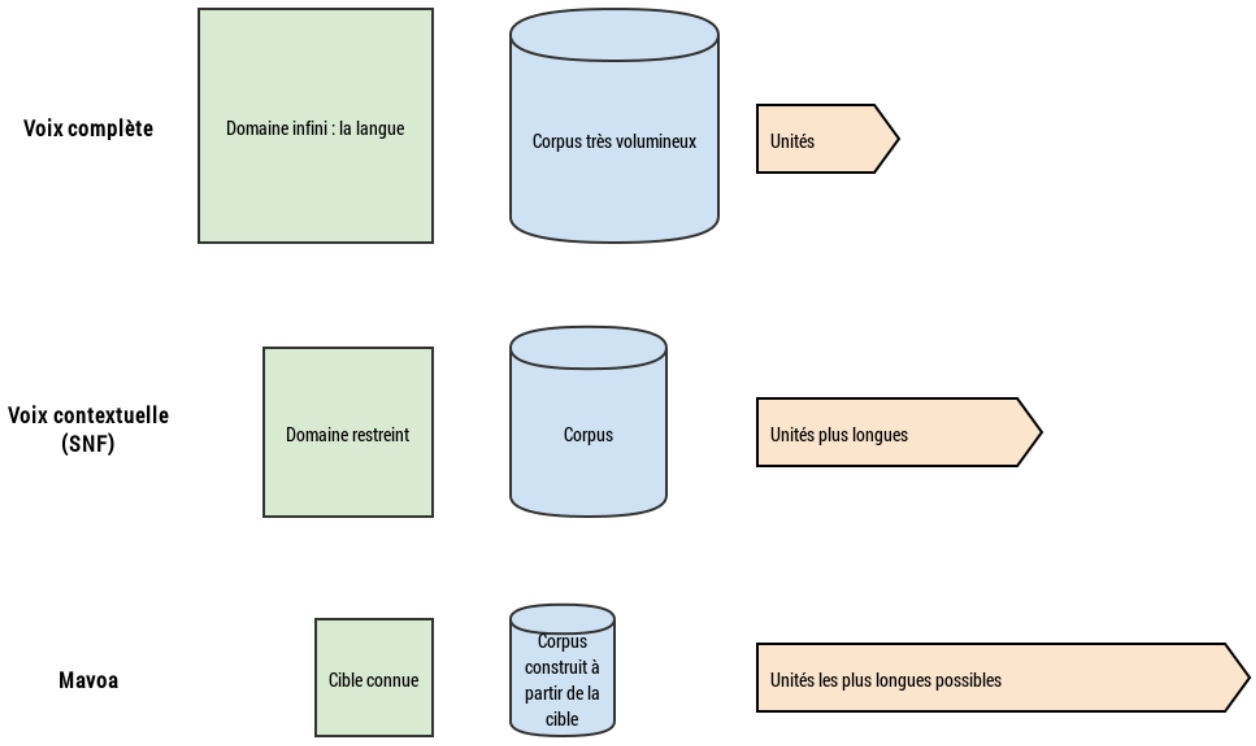


Figure 13 : représentation de la notion d'entonnoir dans la création des corpus et dans la sélection des unités

II. Une question de dimension

Au fil des parties développées jusqu'ici, nous avons pu évoquer deux dimensions dans la sélection des unités :

- La dimension phonétique
- La dimension prosodique

Nous allons donc voir dans cette partie comment, peu importe le procédé utilisé, ces deux dimensions sont dépendantes l'une de l'autre et comment, dans un processus manuel comme Mavoa, et dans un processus automatique comme la création de voix SNF, les difficultés liées à l'une ou l'autre de ces dimensions ont pu être résolues.

La création des scripts condensés se fait en fonction de paramètres phonétiques et prosodiques indispensables et indissociables pour pouvoir sélectionner les unités qui composeront une synthèse de bonne qualité. En effet, la sélection de la bonne unité est déterminée par la correspondance phonétique avec la cible ; l'unité peut donc contenir exactement les mêmes phonèmes que la cible ou bien être assimilable grâce aux regroupements permis. Mais la correspondance phonétique n'est pas suffisante pour définir la meilleure unité à sélectionner. En effet, entrent également en compte des paramètres prosodiques tels que la position syllabique, les marqueurs mélodiques, ou encore le contexte de chaque phonème, qui permettent de choisir telle ou telle unité. Ces deux dimensions sont bien sûr liées et un changement dans les paramètres de l'une ou de l'autre peut changer les coûts de sélection et ainsi donner une synthèse complètement différente.

Dans un processus manuel tel que Mavoa, on cherche avant tout, à l'écriture des scénarios, à couvrir les unités les plus longues possibles tout en gardant un effet de surprise pour l'utilisateur. Dans un premier temps, le but est de couvrir les unités d'un point de vue phonétique. On cherche à couvrir des unités correspondantes phonétiquement. Dans un deuxième temps, la dimension prosodique est prise en compte en faisant en sorte de respecter les positions syllabiques et donc les marqueurs mélodiques de la première et de la dernière unité d'un groupe de souffle. A ce stade, les cas de structures syllabiques, des marqueurs mélodiques, etc, en milieu de groupe de souffle sont relégués au second plan. En effet, ce n'est qu'au stade de test des enregistrements et de la synthèse que l'on peut vérifier les correspondances d'enr, une fois que le dictionnaire est créé et que l'on

peut voir les sorties du module «select» de Baratinoo. Ce module permet, entre autre, de voir toutes les valeurs d'enr de chaque phonème, et donc de vérifier que la sélection attendue est bien celle réalisée. Grâce à ces valeurs d'enr, on peut remarquer que la prise en compte de la prosodie influe sur la sélection des unités au même titre que la phonétique.

Par exemple, dans l'exemple donné en partie 2_III_A, la position syllabique et la structure syllabique de la cible attendue ne permettaient pas au système de sélectionner la bonne unité et entraînaient une coupure sur un phonème fragile, le /r/ : (en bleu l'unité sélectionnée par Baratinoo, en orange l'unité attendue pour le sandwich [T AA R T] en rose)

¿ Puedo imitar tu voz, oyes ?

Puedo imaginar tu voluntad.
¿ El hombre sabe imitar también, oyes ?
Tiene una hermosa voz, como Miguel.

Ici, on souhaite sélectionner directement le sandwich [T AA R T] contenu dans «imitar también» afin de ne pas couper sur une unité fragile comme dans «imaginar tu». Le système a néanmoins préféré choisir le diphone [R-T] contenu dans «imaginar tu». En fait, une règle espagnole rend la sélection de la syllabe initiale d'un mot de plusieurs syllabes (/t/ de «también») pour synthétiser un mot monosyllabique (/t/ de «tu») coûteuse. Le système a donc favorisé le diphone [R-T] seul et a donc coupé sur un phonème fragile, le [R]. Il a donc fallu régler ce problème en changeant la valeur de la structure syllabique du phonème [T] de «también» pour que le système le considère comme venant d'un mot monosylabique.

Dans l'exemple suivant, en français, c'est le marqueur mélodique du /a/ et du /v/ de «avec» dans la première phrase qui posait problème. En effet, en changeant leur valeur de 3 (montant) à 5 (neutre), l'unité attendue était sélectionnée.

Veux-tu venir te promener avec moi ?

Veut-il me promettre de rester avec moi ?
Et tu veux l'emmener avant de te préparer ?
Je venais finir ton travail.

D'autre part, nous avons pu nous rendre compte plus tard, qu'une règle linguistique stipulait qu'un mot fonctionnel («avec») ne devait pas être remplacé par un mot non fonctionnel et donc sémantiquement plein («avant»). C'est donc cette règle qui était à l'origine de «l'erreur» générée.

Par ailleurs, dans le cas d'un processus automatique et donc beaucoup plus lourd, l'objectif est différent. En effet, il s'agit plutôt de jouer sur les deux tableaux en même temps et donc de prendre en compte tous les critères. Dans un processus tel que la création de voix complète, l'intérêt est de définir des règles capables de sélectionner les meilleures unités, sans garantir qu'elles aient exactement les caractéristiques ciblées par la synthèse. On ne parle donc plus, dans ce cas, de «bonnes» unités puisque l'on n'attend pas réellement de cible spécifique, on cherche plutôt à ce que l'algorithme sélectionne la meilleure unité dans sa base de données, en considérant les paramètres phonétiques et acoustiques mais aussi les paramètres prosodiques et linguistiques. Il serait d'ailleurs impossible de vérifier à chaque fois et pour chaque unité les valeurs d'enr attribuées pour essayer de modifier la sélection dans un processus aussi important, c'est pourquoi on ne peut pas atteindre un taux de couverture de 100%. Par ailleurs, dans un processus tel que la création de voix contextuelles, le but est de renvoyer une synthèse de haute qualité, on cherche donc une couverture de 100% des unités. Ce taux de couverture confère au système la possibilité de sélectionner les «bonnes» unités et non pas seulement les meilleures. En effet, le domaine étant restreint et le corpus fini, les unités sont toutes sélectionnées et correspondent à la cible attendue.

On peut donc noter, encore une fois, que le processus de sélection, à la fois pour le script condensé mais aussi pour la synthèse, est tout à fait lié à la cible applicative visée, que ce soit d'un point de vue phonétique ou prosodique.

III. Des attentes différentes selon l'utilisation des corpus

Maintenant que nous avons vu comment les scripts condensés et les corpus étaient construits, nous pouvons nous interroger sur les raisons et la manière dont ces corpus sont choisis suivant l'utilisation qu'il en sera faite. En effet, nous pouvons évoquer trois façons de construire les corpus en fonction de trois utilisations différentes.

Dans un premier temps, intéressons nous au choix du corpus pour la création d'une voix complète. Nous avons eu l'occasion durant ce stage de travailler sur une voix complète un peu particulière : une « voix patient ». Les voix patient sont enregistrées dans le cadre du développement de l'application VoxMed, créée par Voxygen, en partenariat avec le Centre Hospitalier Universitaire Pontchaillou à Rennes. Le principe d'une telle application est de permettre à des patients atteints d'une maladie entraînant une perte partielle ou totale de leur voix, de pouvoir vocaliser n'importe quelle entrée textuelle avec leur propre voix enregistrée avant l'opération. Nous parlons ici d'une voix complète un peu particulière car, étant destinée à des patients dans un délai très court, il est nécessaire de construire un corpus qui soit le plus petit possible pour permettre des enregistrements rapides et sans soufrances.

Dans le cadre d'une voix complète donc, le corpus est défini de façon à couvrir (idéalement) toute la langue, le sens des phrases n'a donc pas vraiment d'importance. L'essentiel dans ce type de travail est d'avoir un pourcentage de couverture qui soit le plus élevé possible afin de pouvoir compter, au minimum, tous les diphones de la langue dans le script condensé. La particularité d'une voix patient réside donc dans le fait que le corpus est construit de façon à ce que le nombre de phrases soit assez réduit pour être enregistré en une seule journée. En effet, le laps de temps entre le moment où les patients sont mis au courant de la nécessité d'opérer et l'annonce de la date de l'opération qui leur privera de leur voix est très court. De plus les personnes étant malades et souvent âgées, elles se fatiguent très vite, il est donc nécessaire de constituer des scripts locuteurs assez courts.

La nécessité de construire des corpus et des scripts condensés courts oblige l'entreprise à revoir ses critères de sélection à la baisse. En effet, pour pouvoir réduire le nombre de phrases du script locuteur, il faut forcément attendre moins d'unités à couvrir. C'est en cela que nous avons parlé d'une voix complète «un peu particulière» car, à la différence d'une voix complète «normale» où l'on cherche à couvrir toute la langue en essayant d'obtenir un nombre de phrase correct (ni trop élevé ni trop faible) par rapport aux unités à couvrir, le but ici est de réduire au maximum le nombre de phrases du script locuteur tout en restant cohérent dans la couverture des unités.

Les concessions permettant la réduction du nombre de phrases dans le script condensé interviennent dans les deux dimensions évoquées dans la partie précédente : la dimension phonétique et la dimension prosodique. Pour permettre une telle réduction (1000 phrases pour une

voix patient contre plusieurs milliers selon les langues pour une voix complète «normale»), plusieurs méthodes sont mises en place :

- des fichiers de regroupements qui forcent l'algorithme à sélectionner les unités de couverture selon des règles prédéfinies : une réduction du nombre de phonèmes est réalisée en considérant comme assimilables certains phonèmes proches phonétiquement et prosodiquement. Par exemple : on considère /im/ et /um/ comme équivalents, il n'est donc pas nécessaire de les sélectionner tous les deux.
- On ne sélectionne que les diphones fréquents car on considère que les patients n'auront pas réellement besoin d'utiliser des diphones rares.
- Les sources utilisées pour constituer le corpus sont réduites, on utilise seulement le corpus « sous-titres de films » pour les voix patients (corpus conversationnel).

Deux nouveaux outils sont en cours de développement et nécessitent des tests :

- Un script permettant de remplacer une intonation descendante devant une pause par une intonation montante. Ceci permettrait de ne pas avoir à enregistrer tous les cas de figure mais engendrerait peut-être des unités manquantes.
- Un script permettant de sélectionner certaines unités dans une autre voix (appelée voix porteuse) pour remplacer d'éventuelles unités manquantes.

Ces petits ajustements permettent de réduire considérablement le nombre de phrases du script locuteur pour une voix patient. Le problème de ce genre de corpus pour des personnes non professionnelles et qui n'ont pas l'habitude d'enregistrer est le manque de cohérence entre le sens des phrases qu'ils lisent et ce qu'ils attendent de l'application (on peut comparer ceci à l'effet de surprise engendré par la synthèse de l'application Mavoa). En effet, il faudrait pouvoir adapter les corpus pour sélectionner des phrases plus cohérentes, contenant moins d'anglicismes, moins de termes techniques ou moins de termes parfois très familiers (ce que l'on peut rapprocher du processus de création de voix contextuelles, appliqué à un domaine restreint de la langue). Il est peut être compliqué, pour les patients, de comprendre ce que les phrases relativement insensées qu'ils lisent et enregistrent peuvent avoir à voir avec la synthèse de leur voix. De plus, et pour aller encore plus loin dans la cohérence avec les voix patient, il serait peut-être plus intéressant de constituer des corpus directement à partir des phrases entrées par les patients dans l'application. Puisque l'application qui leur est fournie intègre un historique des phrases synthétisées, il serait facile de les récolter afin d'en faire un corpus plus adapté (en SNF par exemple). En effet, à l'heure

actuelle, nous avons vu que les corpus étaient créés à partir des mêmes sources (néanmoins restreintes) que pour des voix complètes «normales», les scripts locuteurs contiennent donc énormément d'unités inutiles qui ne seront jamais utilisées au quotidien par les patients. La possibilité de récupérer les phrases déjà utilisées permettrait donc de construire des corpus beaucoup plus adaptés qui ne couvriraient que les unités nécessaires.

Dans le cadre de la création d'une voix contextuelle comme celles que nous avons pu réaliser pour l'automobile en français et en anglais américain, le choix du corpus qui constituera le script locuteur est complètement différent et très simplifié. En effet, nous sommes ici dans un domaine très restreint qui tend à n'utiliser la synthèse que dans le cadre de cette application précise, le script condensé est donc directement en rapport avec la cible. Ici, le script condensé est construit à partir du corpus lui-même issu de la description du domaine, aucune source extérieure n'est utilisée pour construire le corpus. Finalement, on peut dire ici que le script condensé constitue un sous-ensemble de la cible, le corpus est donc la cible puisque tout ce qui pourra être synthétisé se trouve dans le corpus de départ.

De plus, dans ce cas de figure, aucune question n'est à se poser du côté du sens des phrases puisqu'il est défini par le client qui souhaite intégrer telle ou telle tournure de phrase. Ici, le sens des phrases importe donc beaucoup mais n'est pas défini par l'entreprise.

Dans le cadre de l'application Mavoa, on se trouve plutôt dans un compromis entre les deux cas de figure précédents. En effet, ici la cible est très précise puisqu'il s'agit d'une seule phrase, il est donc nécessaire de la connaître pour construire le corpus. En partant de la phrase de synthèse, on construit le script locuteur (aussi corpus dans ce cas), le choix des phrases est donc relativement réduit puisqu'il faut pouvoir couvrir, à la main, les unités de la cible. On cherche donc simplement à remplir les caractéristiques phonétiques, linguistiques et prosodiques de la cible.

Cependant, il est également nécessaire de créer des phrases ayant un sens et n'étant pas simplement une suite de mots sans cohérence. On souhaite également créer un corpus agréable et plutôt fun. En effet, il s'agit d'une application destinée à un public très varié, pour une utilisation loisir, l'entreprise s'attache donc à garder un côté ludique mais néanmoins une certaine élégance dans la création de ces corpus.

Nous pouvons donc noter que la création des scripts condensés est fortement corrélée au type

d’application et au public visé. En effet, le type d’application entraîne un nombre plus ou moins important de phrases qui auront un impact sur la couverture des unités, et le public visé implique que l’on s’attache plus ou moins au sens des phrases.

Plurilinguisme et pluridisciplinarité

Nous avons vu dans les trois parties précédentes comment l'entreprise Voxygen a mis au point des outils et des techniques permettant d'obtenir une synthèse de meilleure qualité. Entre la sélection d'unités de longueurs variables et destinées à protéger les phonèmes fragiles des concaténations : les sandwiches, et la création de corpus dans un domaine restreint et fini : Slot'N'Fill, nous avons pu, durant ce stage, expérimenter et valider des procédés voués à augmenter la qualité de la synthèse. De plus, nous avons pu tester ses différentes techniques sur des langues de différentes natures et dans des contextes variés, c'est pourquoi nous avons choisi de parler ici de plurilinguisme et de pluridisciplinarité.

Dans cette dernière partie, donc, nous commencerons par exposer les différents travaux réalisés, du point de vue des langues, en évoquant les problèmes rencontrés et les solutions trouvées. Puis, dans un deuxième temps, d'un point de vue disciplinaire, nous présenterons les travaux variés que nous avons pu mener en décrivant toujours les problèmes rencontrés et les solutions trouvées. Enfin, nous essayerons de donner une conclusion personnelle à ce chapitre en faisant un bilan des travaux réalisés et en introduisant les prospectives envisagées.

I. Plurilinguisme

La notion de plurilinguisme est très présente chez Voxygen, elle-même une entreprise employant des personnalités souvent polyglottes et de nationalités diverses, puisque ses travaux de recherche et les services proposés sont accessibles dans de nombreuses langues. Nous avons donc pu, durant ce stage, réaliser plusieurs travaux dans différentes langues. Nous commencerons par exposer les travaux réalisés en français, puis ceux réalisés en anglais américain et enfin les travaux menés en espagnol. Pour chaque langue nous expliquerons les problèmes rencontrés ainsi que les solutions trouvées afin de montrer comment le fonctionnement de chaque langue peut engendrer des problèmes différents et donc des solutions divergentes.

A. Le français

Tout d'abord, nous avons travaillé sur le français dans deux applications différentes : les voix contextuelles (Slot'N'Fill) et Mavoa. Le travail réalisé pour l'automobile (SNF) en français a demandé plusieurs révisions à cause de nombreux problèmes rencontrés. Comme nous l'avons déjà évoqué, les travaux effectués pour l'automobile entraient dans un processus de création de voix contextuelles (SNF), il a donc fallu, dans un premier temps, créer un PLS et un script de génération de phrases. Comme expliqué en partie 3_I, le PLS répertorie toutes les parties fixes du corpus en donnant à chacune un attribut meta et une transcription phonétique matérialisée par le symbole /7/ pour chaque mot, séparés par /##/. Chaque partie fixe est découpée sur le premier ou le dernier phonème robuste rencontré (selon la place de la partie fixe dans la phrase). Ensuite, le programme de génération est construit en langage informatique python. Il liste toutes les phrases souhaitées avec leur partie fixe et une boucle numérique permettant de générer tous les nombres désirés.

- Un des premiers problèmes auxquels nous ayons dû faire face a été la prononciation des «e» muets dans certains nombres. En effet, le nombre «14» était phonétisé [K A T O R Z EU] par le système devant le mot «pour cent», or nous ne souhaitions pas la prononciation de ce /e/ muet en fin de mot dans ce contexte. Pour cela, nous avons ajouté une entrée dans le PLS pour chaque nombre se trouvant dans le même cas, en donnant la transcription phonétique souhaitée :

```
<lexeme>
    <grapheme>14 %</grapheme>
    <phoneme>KATORZ##POU"RSAN"</phoneme>
</lexeme>
```

Ainsi, le système est forcé de prononcer l'entrée telle qu'elle est phonétisée dans le PLS.

- Lors de l'écriture du PLS, la prise en compte du singulier et du pluriel séparément n'avait pas été évoquée. Or, à l'étape de validation, nous nous sommes rendu compte que ceci provoquait un manque de prise en compte de certaines parties fixes. En effet, la partie fixe au pluriel avait été entrée dans le PLS avec sa transcription phonétique mais dans le générateur de phrases, le pluriel ET le singulier étaient présents. Par conséquent, lors de la validation, le système renvoyait la partie fixe au singulier sans transcription

phonétique SNF (7##7...). Il a donc fallu, pour chaque phrase présentant une possibilité de pluriel et de singulier, entrer les deux variantes dans le PLS. De plus, en français, la prise en compte du pluriel n'entraînant pas de différence phonétique, nous avons pu entrer les deux graphèmes dans la même entrée, sous la même forme phonétique. (Nous verrons plus loin que cela n'a pas été possible pour l'américain).

```
<lexeme>
        <grapheme>kilomètres</grapheme>
        <grapheme>kilomètre</grapheme>
        <phoneme vox:meta="km">K7</phoneme>
</lexeme>
```

On peut donc remarquer que les deux graphèmes sont associés à la même prononciation (un seul attribut meta).

Sur le même principe, certains oublis de majuscule dans le script de génération ont provoqué un manque de prise en compte dans l'analyse dans l'outil de validation des phrases. Par exemple :

PLS :

```
<lexeme>
        <grapheme>l'heure d'arrivée escomptée est</grapheme>
        <phoneme vox:meta="l_heure_d_arrivee_escomptee">7##7##7TEI"##AI"</phoneme>
</lexeme>
```

Script informatique de génération des phrases :

```
for n in range(24):
   if n<2:
      print """L'heure d'arrivée escomptée est %d heure."""%n
```

On note la présence d'une majuscule dans le générateur mais pas dans le PLS, l'analyse ne peut donc pas considérer ces deux phrases comme identiques. Pour régler ces problèmes relatifs à la casse, nous avons ajouté une commande au PLS permettant au système d'être insensible à la casse.

```
<?xml version="1.0" encoding="utf-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
                http://www.w3.org/TR/pronunciation-lexicon/pls.xsd"
  xmlns:vox="http://www.voxygen.fr/tts"
  alphabet="x-voxygen" xml:lang="fr" vox:opt="i">
```

- Lors de l'écriture du PLS, une question s'est posée à propos d'un phénomène d'assimilation. En effet, certains phonèmes, selon leur environnement, peuvent engendrer des phénomènes d'assimilation, c'est à dire le *«transfert d'une caractéristique ou trait phonétique d'un son sur un son voisin. Quand deux sons de parole sont en contact, l'un communique à l'autre un de ses traits, partiellement ou totalement.»* [Snoeren, N. D., 2005]. Dans notre cas, par exemple dans la phrase «Attention, vous êtes 10 kilomètres par heure au-dessus de la limite autorisée.», «vous êtes» est une partie fixe, il faut donc la couper sur un phonème robuste, le [T]. Or, devant une consonne plosive ou fricative sonore ([D] de «dix»), il y a un phénomène d'assimilation et le [T] devient [D]. Nous nous sommes donc demandé s'il fallait alors entrer deux prononciations dans le PLS, [V OU Z AI T] et [V OU Z AI D]. Finalement, après vérification dans Baratinoo, nous avons pu voir que l'assimilation se faisait automatiquement en répertoriant seulement [V OU Z AI T].

- Par ailleurs, devant une voyelle, la liaison devrait se faire avec le [S], or, en coupant sur le [T] la liaison ne se fait pas. Nous avons cependant décidé de ne pas tenir compte de cette liaison, considérée comme facultative.

Finalement, les règles du français étant plutôt complètes, les problèmes que nous avons rencontrés venaient principalement de l'écriture du PLS qui n'était pas toujours optimale.

Dans un deuxième temps, intéressons-nous à l'utilisation du français dans l'application Mavoa. Dans ce contexte, que nous avons déjà évoqué plusieurs fois, les difficultés liées à la langue ont plutôt concerné le processus de sélection des unités à la synthèse. Le principal problème rencontré a touché la sélection d'un phonème non attendu à cause d'une règle linguistique peut-être un peu trop sévère. (L'exemple est détaillé en partie 3_II). En effet, une règle française précise que

pour deux candidats à un mot grammatical («avec»), l'un lexical («avant»), l'autre grammatical («avec»), le candidat lexical est éliminé par l'algorithme (et inversement).

Veux-tu venir te promener avec moi ?

Veut-il me promettre de rester avec moi ?
Et tu veux l'emmener avant de te préparer ?
Tu veux nager pour finir ta course.

Devant l'impossibilité de sélectionner le segment «av» du mot «avant», le système a sélectionné «mener» et «avec moi», coupant sur un phonème fragile, le [A]. Pour régler ce problème nous aurions pu, directement dans le fichier phn, exclure l'unité non désirée pour forcer le système à sélectionner l'autre candidat, mais nous avons préféré modifier certaines valeurs d'enr.

B. US english

Nous allons maintenant nous intéresser à l'américain, utilisé pour créer une voix contextuelle pour l'automobile (SNF), et montrer en quoi les problèmes rencontrés liés à la langue sont différents du français. Le processus de création, cependant, est exactement le même et le corpus créé passe par les mêmes étapes de parsing, calcul des statistiques et validation des phrases qu'en français.

- En américain, le PLS a été construit directement en distinguant le singulier du pluriel. Or, il ne suffit pas de les distinguer comme en français, il est également nécessaire de les répertorier dans deux entrées différentes puisque leur prononciation diffère. En effet, le [S] en américain, est prononcé. Le risque, si l'on ne distingue pas la moindre différence acoustique, est d'engendrer un manque de couverture des unités, et donc que le système puisse sélectionner un mot au pluriel pour synthétiser un mot écrit au singulier, et inversement.

```
<lexeme>
        <grapheme>minute</grapheme>
        <phoneme vox:meta="minute">m 7</phoneme>
</lexeme>
<lexeme>
        <grapheme>minutes</grapheme>
        <phoneme vox:meta="minutes">m 7</phoneme>
</lexeme>
```

- Par ailleurs, des problèmes liés à la lecture des chiffres en américain sont apparus. Par exemple, lorsque le système rencontre un nombre décimal, nous souhaitons qu'il prononce la virgule, or en américain, le système marquait une pause. Finalement, nous avons réalisé que les nombres décimaux, dans cette langue, étaient marqués par un point et non une virgule. Le problème a donc été résolu ainsi.
- De plus, en ce qui concerne la lecture des numéros de téléphone et des nombres d'au moins quatre chiffres, nous avons rencontré le même genre de problème. En effet, nous avions conservé l'écriture des numéros de téléphone français : 02, 56, 42. Or, en américain, la lecture des numéros de téléphone se fait chiffre par chiffre, on ne veut donc pas de pause. Nous avons donc finalement préféré la structure : 025, 458., qui pose moins de problèmes et qui permet de couvrir tous les cas de figure. Pour cela, il suffit, dans le programme de génération des phrases, de créer une boucle sur chaque chiffre pour permettre au système de tous les générer de 1 à 9.
- Concernant les nombres d'au moins quatre chiffres, le système lisait les chiffres un par un pour des nombres d'au moins cinq chiffres et deux par deux pour des nombres de quatre chiffres. Il a donc fallu modifier le générateur de phrases en mettant une virgule après le premier chiffre pour les nombres de quatre chiffres et après le deuxième chiffre pour les nombres de cinq chiffres.

- Durant la validation, nous sommes tombés sur une difficulté gênante et compliquée à régler puisqu'elle nous a demandé beaucoup de tests et donc de temps. La difficulté s'est présentée sur des phrases de la forme «It is 12:23 AM.» ou «It is 6 PM.». Concernant le «PM», le système renvoyait constamment «o'clock pm», ce qui était très gênant

puisqu'incorrect. Nous avons donc créé ce que l'on appelle un «alias» dans le PLS pour forcer la prononciation du «PM» en [pii_e_m]. Un alias est une balise du fichier PLS remplaçant la phonétique par défaut d'un graphème par celle souhaitée.

```
<lexeme>
    <grapheme>PM</grapheme>
    <alias>P M</alias>
</lexeme>
```

Concernant le «AM», plusieurs cas de figure se sont présentés :

- phonétisé «a_m» dans le cas des heures avec minutes «12:11 AM».
- Dans certains cas, phonétisé «@_k_l_o_k_|_ei_e_m_|_ei_e_m»
- Dans certains cas, phonétisé «@_k_l_o_k_|_ei_e_m»
- Dans certains cas, phonétisé «ei_e_m_e_m»

Alors qu'il aurait dû être phonétisé «ei_e_m». La solution trouvée a finalement été de créer, comme pour «PM», une entrée dans le PLS.

```
<lexeme vox:scope=«internal»>
    <grapheme>AM</grapheme>
    <phoneme>* ei . e m</phoneme>
</lexeme>
```

Le *vox:scope=«internal»* force le système à prononcer l'entrée ainsi seulement dans le cadre de ce script (internal), cette prononciation ne s'applique donc jamais ailleurs. «PM» n'a pas nécessité ce genre de précisions puisqu'il n'est prononçable par le système qu'en séparant les deux lettres «pii» «em». Le sigle «AM», lui, pouvait être prononcé comme l'auxiliaire : «am». Les problèmes de vocalisation du «AM» dans le cas des heures avec minutes venaient du fait que, au-delà de 12, le système ne considérait pas les nombres comme une expression de l'heure et donc ne phonétisait pas correctement.

Finalement, les difficultés rencontrées en américain étaient surtout liées à la traduction des phrases et à la «conversion» des unités. En effet, il n'a pas suffi simplement de traduire des phrases vers l'américain, il a fallu modifier ces phrases pour pouvoir les couper sur des phonèmes robustes.

Le fait de passer d'une langue à l'autre modifie le texte et les phonèmes robustes du français ne sont plus forcément les mêmes dans une autre langue. De plus, il a fallu convertir les unités de mesure qui sont différentes en américain et en français (exemple : km → mile, km/h → mph, litres aux 100 → miles per gallon), et le format des heures.

En définitive, les problèmes rencontrés dans les processus de création de voix contextuelles en français et en américain viennent de mauvaises prises en comptes phonétiques et linguistiques et de « bugs » présents dans les hauts-niveaux du système Baratinoo. Les solutions proposées pour régler les problèmes évoqués ci-dessus sont donc à envisager dans un contexte dans lequel il n'est pas possible, pour des questions de temps, de revenir sur des problèmes en profondeur, jusqu'aux hauts-niveaux. C'est donc pourquoi, ici, nous avons choisi de régler les problèmes rapidement et en surface plutôt que de les régler en amont. Précisons néanmoins que ces problèmes ont été réglés par la suite « à la racine ».

C. El español

Le travail réalisé en espagnol a porté sur l'écriture de scénarios pour l'application Mavoa. Comme déjà expliqué, l'objectif de cette application est de faire enregistrer des phrases à l'utilisateur afin de synthétiser une phrase qu'il n'a jamais prononcée, avec sa propre voix, et donc de créer un effet de surprise. Le processus de création des phrases à enregistrer étant manuel, il est très important de respecter certaines contraintes afin de rendre la synthèse possible et de qualité correcte. C'est donc le respect de ces contraintes qui a constitué la principale difficulté dans ce travail. Tout d'abord, le fait d'avoir à écrire des scénarios dans une langue étrangère constitue un premier écueil puisque, le vocabulaire manquant parfois, il est difficile de trouver, à la main, des expressions collant à la fois à la phonétique attendue mais également à des contraintes prosodiques. Dans un deuxième temps, il est nécessaire de préciser que toutes les langues présentent des caractéristiques linguistiques et phonétiques différentes dont il faut tenir compte dans la construction des phrases. En effet, les phonèmes fragiles et robustes ne sont pas toujours les mêmes, par exemple, en espagnol, le [R] et le [RR] sont des phonèmes fragiles, ce qui n'est pas le cas en français. Par ailleurs, l'espagnol comporte beaucoup de voyelles accentuées et donc allongées qui n'existent pas en français et dont il est primordial de tenir compte dans l'écriture des scénarios. En effet, si la phrase de synthèse comporte une voyelle longue, il est important d'utiliser une voyelle de même type dans l'unité à sélectionner. Le risque dans le cas contraire est que :

- le système choisisse l'unité attendue mais que la synthèse soit médiocre.
- le système ne sélectionne rien et que l'on se retrouve avec une unité manquante, ce qui serait très gênant dans ce type d'application qui synthétise en direct devant l'utilisateur.
- le système choisisse une autre unité considérée comme assimilable d'après la liste des phonèmes (dans la liste des phonèmes de la langue, certains sont notés comme étant remplaçables par un autre phonème au cas où le système ne trouverait pas l'unité attendue).

C'est cette dernière solution que le système a choisi sur un des scénarios écrits et qui nous a contraint à changer de phrase :

Pancha peca tuca peta pide picallo en pallanda

Una flecha pecadora toca Petra y toca tu corazón.
Pancho, en Tapiela, el hombre pide picadillos.
Das el callo en Páris.
Escapa llanuras de Irlanda.

Le /e/ de «pecadora» était attendu pour synthétiser celui de «peca». Or, le /e/ de «peca» étant dans l'avant dernière syllabe, il est accentué, et donc allongé, ce qui n'est pas le cas du /e/ de «pecadora» puisqu'il est dans la première syllabe d'un mot de quatre syllabes. Le système a donc bien sélectionné cette unité car, en cas de non présence du [EE], le système est autorisé à sélectionner un [E]. Notons que le système aurait pu sélectionner le [E], voyelle courte, de «Petra», cependant, il aurait été obligé de couper sur ce phonème fragile, le coût de concaténation était donc trop élevé. Finalement, le coût total engendré par la concaténation sur un [E] était plus élevé que le coût total engendré par la sélection du [EE] à la place du [E] puisque dans ce cas, la coupure se fait sur le /c/ (phonème robuste).

Par ailleurs, l'interface de l'application Mavoa, aujourd'hui n'est disponible qu'en français. En effet, seuls les scénarios sont présents dans plusieurs langues et toutes les consignes, qu'elles soient écrites ou orales, sont données en français. Il aurait donc peut-être été intéressant de traduire toute l'interface de l'application, pour chaque langue disponible dans les scénarios.

II. Pluridisciplinarité

La notion de pluridisciplinarité a été choisie pour qualifier ce stage en raison des nombreux travaux divers et variés qui ont pu y être menés. En effet, même si la plupart des tâches confiées avaient toutes un lien, que ce soit au niveau de la sélection des unités, des types d'unités ou des applications visées, les domaines d'application pouvaient être très différents. De plus, le fait d'être intervenue sur toutes les étapes des processus auxquels nous avons pris part a permis de renforcer cette impression de pluridisciplinarité.

Je tiens également à évoquer rapidement ici l'utilisation systématique du terminal dans un environnement Linux. En effet, Linux est un environnement que nous n'avons pas eu l'occasion d'utiliser durant la formation en Master, ceci a donc constitué une première difficulté à mon arrivée chez Voxygen. J'ai finalement dû apprendre à me familiariser avec cet environnement et à utiliser le plus possible les commandes dédiées. Cette façon de travailler, en définitive, a constitué un réel gain de temps et une véritable opportunité de progrès pour l'avenir.

Nous verrons, dans cette dernière partie, les différents travaux réalisés durant ce stage et montrerons en quoi la notion de pluridisciplinarité est omniprésente. Nous verrons dans un premier temps le processus complet de création de voix contextuelles (SNF), puis nous présenterons deux programmes écrits en langage python destinés à faciliter certaines étapes et enfin nous exposerons les différentes phases du fonctionnement de l'application Mavoa, de l'écriture des scénarios à leur affichage dans l'application. Pour chaque tâche, nous exposerons les problèmes, relatifs aux processus menés et non plus seulement à la langue, que nous avons rencontrés ainsi que les solutions trouvées.

A. Le Slot'N'Fill, du PLS à la sélection des phrases

Comme déjà détaillé dans la partie 3_I de ce travail, la création de voix contextuelles est un processus en plusieurs étapes. Tout d'abord, il s'agit, comme dans tout processus de création de voix, de créer un corpus et un script condensé afin de déterminer les unités à couvrir. Dans le cadre du Slot'N'Fill, le corpus est construit à partir du PLS et du programme de génération de phrases,

d’après les recommandations du client. Ces deux étapes n’ayant pas réellement posé de problèmes, ayant déjà été détaillées en partie 3_I et les fichiers étant déjà prêts avant le début du stage, nous ne les présenterons pas ici. Les étapes de la création d’une voix contextuelle sont donc les suivantes :

1. Écriture du PLS
2. Écriture du générateur de phrases (génère un fichier texte contenant toutes les phrases : c’est notre corpus)
3. Parsing à partir du corpus (on obtient la pioche)
4. Calcul des statistiques sur la pioche
5. Validation des phrases sur la pioche (on obtient le script condensé)
6. Mise en forme du script condensé pour former le script locuteur
7. Enregistrements
8. Segmentation
9. Sélection des phrases

Les quatre premières étapes du processus ayant déjà été détaillées dans la partie 3_I de ce travail, nous ne les présenterons pas ici. De plus, le PLS et le générateur de phrases ayant déjà été écrits avant le début du stage, nous n’avons eu qu’à les modifier afin de les adapter à nos besoins.

L’étape de validation des phrases peut être réalisée soit par un opérateur, soit automatiquement. En cas de validation automatique, une limite du nombre de phrases à couvrir doit être donnée pour éviter que le système ne couvre toutes les phrases. La validation manuelle, que nous avons le plus souvent mise en oeuvre que ce soit pour le français ou l’américain, se fait sur un outil dédié :

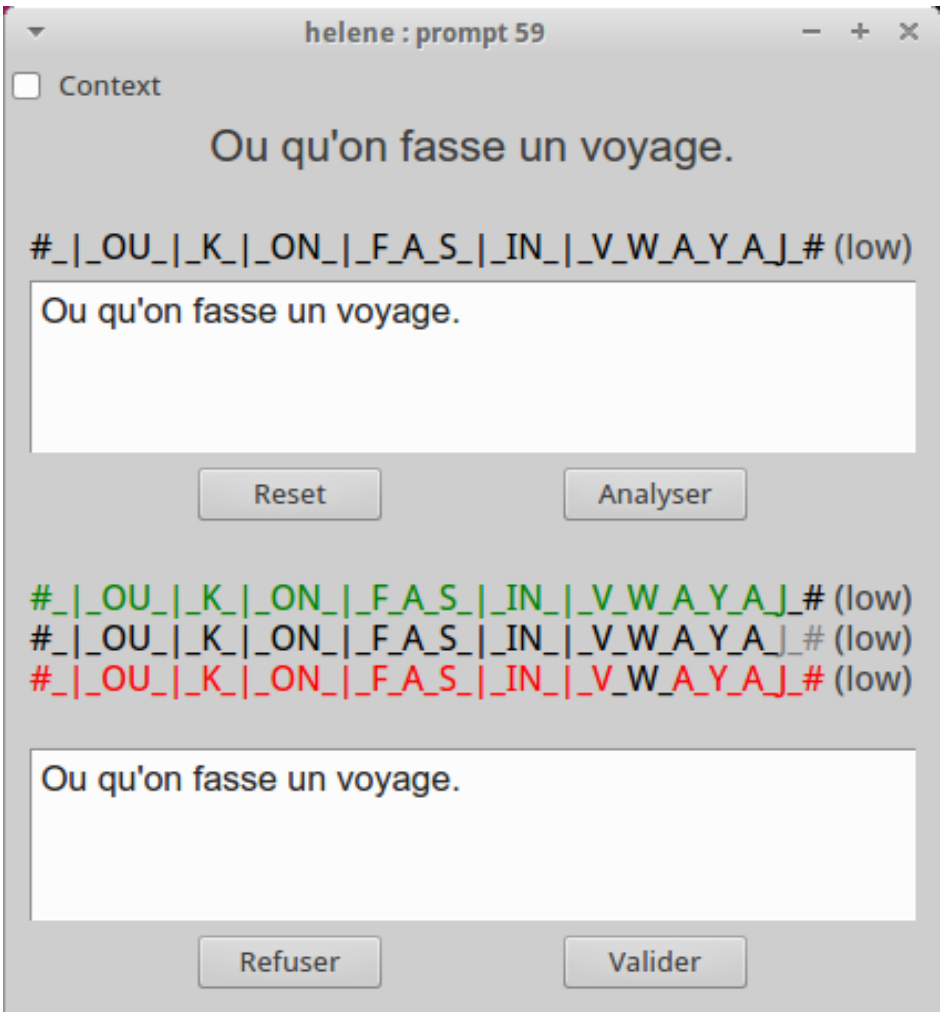


Figure 14 : outil de validation des phrases

Les phrases se présentent une par une avec leur transcription phonétique avec, en haut l'analyse du parsing et en bas l'analyse du système. L'opérateur doit vérifier la concordance des deux analyses et surveiller la couverture des unités. Durant la session, il peut accepter ou refuser des phrases qui ne lui semblent pas pertinentes ou contenant des erreurs quelconques. Les unités de couleur verte représentent les diphones qui apportent une augmentation de la couverture des unités, au même titre que les unités de couleur noire représentant les clusters consonnantiques et les unités de couleur rouge représentant les sandwiches. Les unités n'apportant rien à la couverture deviennent noires. A la fin de la session de validation, toutes (ou presque) les unités sont noires, la couverture est donc complète (ou presque) et le script condensé est prêt.

Ensuite, à partir de ce script condensé, le script locuteur est créé grâce à une simple commande. Le script est en fait simplement mis en forme avec la phrase à lire, sa transcription phonétique ainsi que d'éventuelles annotations ajoutées lors de la validation.

Vous avez un appel du 07, ^^^ #_7_7_7_7_D_U_Z_EI_R_AU_S_AI_T_#

Voici une phrase extraite du script locuteur français pour l'automobile. On peut noter que ce script SNF contient les fameux symboles /7/ présents dans le PLS qui caractérisent les parties fixes. En effet, la prononciation de parties fixes important peu, le locuteur est libre de les dire telles qu'il le souhaite, on ne lui donne donc pas de phonétique. Pour information, les ^^^ représentent une intonation montante, c'est ce qui permet d'informer le locuteur sur l'intonation souhaitée et donc de fixer la prosodie du groupe de souffle.

Arrive alors l'étape des enregistrements. Cette étape est souvent longue puisque l'on y consacre une journée entière pour des voix SNF et, suivant les langues, entre une et deux semaines pour une voix complète. Cette étape incluant des procédés réalisés personnellement, j'utiliserai la première personne pour en parler. Durant ce stage, j'ai pu observer et co-diriger les enregistrements de l'américain par Paul Gouin, technicien R&D chez Voxygen et franco-américain, et moi-même enregistrer le français. Concernant l'américain, les enregistrements se sont déroulés en deux temps. En effet, le travail de création de voix contextuelle étant nouveau dans cette langue, nous avons voulu comparer deux regroupements différents, l'un pauvre, l'autre riche. Le regroupement riche entraîne donc un nombre de phrases plus élevé que le regroupement pauvre qui rassemble un maximum d'unités. Nous avons donc enregistré un premier script de 495 phrases en regroupement pauvre, puis un deuxième de 564 phrases en regroupement riche.

La session d'enregistrement se déroule dans une pièce insonorisée contenant une cabine «sourde» qui ne laisse passer aucun son extérieur. Le locuteur est donc placé dans cette cabine et dispose d'un écran sur lequel il peut lire les phrases à enregistrer, d'un micro et d'un casque lui permettant d'entendre les commentaires des opérateurs en dehors de la cabine. Les opérateurs en dehors de la cabine, eux, sont chargés d'envoyer les phrases sur l'écran du locuteur, de lancer l'enregistrement et de contrôler la bonne prononciation, l'intonation attendue, le rythme souhaité... des phrases prononcées par le locuteur.

Les enregistrements, bien que constituant une tâche agréable, ont néanmoins posé quelques difficultés d'ordres acoustique et linguistique :

- Les bruits de bouche constituent une difficulté très compliquée à éviter mais entachent néanmoins la qualité du signal.

- le fait de parvenir à garder un ton homogène au fur et à mesure de l'enregistrement est également très difficile pour le locuteur, il faut donc régulièrement écouter les phrases précédemment enregistrées pour se rendre compte des changements éventuels.

Au niveau linguistique, nous avons été confrontés à deux obstacles :

- En américain, la prononciation du /d^/ (appelé /d/ flap). En effet, ce phonème situé entre le /d/ et le /t/ n'était pas toujours prononcé de la même façon par le locuteur et cela posait problème dans la correspondance entre la phonétique et le signal. Nous avons finalement décidé de laisser le locuteur les prononcer comme il l'entendait afin de garder une prononciation naturelle et non forcée.
- les numéros de téléphone en français : étant rassemblés trois par trois, séparés par une virgule, nous avions normalement prévu de prononcer le premier groupe avec une intonation montante puis le deuxième avec une intonation descendante. Cependant, pour plus de modularité à la synthèse, nous avons décidé de faire prononcer les deux groupes de la même façon afin de pouvoir les réutiliser dans n'importe quel contexte.

En ce qui concerne le français, les recommandations sont les mêmes, il faut contrôler la bonne prononciation du locuteur, faire attention aux bruits de bouche, etc. En ce qui me concerne, enregistrer a été une certaine satisfaction puisque je participais réellement à la création de la voix. Nous avons enregistré 368 phrases, dans les mêmes conditions que pour l'américain. Les difficultés que j'ai pu rencontrer sont plutôt personnelles puisqu'il a fallu que je me sente assez à l'aise pour effectuer différentes intonations sur les parties fixes. En effet, l'objectif était d'enregistrer les parties fixes avec des tonalités variées et enjouées. De plus, les bruits de bouche, surtout après un grand nombre de phrases enregistrées, sont difficiles à contrôler. Par ailleurs, concernant les numéros de téléphone, dans lesquels les nombres sont séparés par des virgules, nous nous sommes demandé s'il était préférable, comme défini par le script, de les enregistrer séparément ou ensemble jusqu'au point final. Nous avons finalement décidé de réaliser les deux, dans un premier temps seuls puis en entier. En effet, le fait de s'arrêter sur les virgules pourrait engendrer une intonation plus allongée et moins naturelle. Nous verrons plus loin qu'un test a montré que les enregistrements des numéros entiers étaient plus appréciés pour leur naturel.

Les enregistrements constituent finalement une tâche extrêmement importante dans le processus de création de voix pour la synthèse puisque c'est l'étape sur laquelle repose la sélection des unités. En effet, c'est à partir des enregistrements du locuteur que les unités sont sélectionnées

pour construire la synthèse. Par conséquent, si les enregistrements se passent mal, la synthèse sera de très mauvaise qualité. Il peut donc se poser de nombreux problèmes qui engendreraient un manque de couverture, une mauvaise intelligibilité, etc. Les enregistrements représentent donc une étape fondamentale dans la création d'une voix.

Après les enregistrements, vient la phase de segmentation. Une segmentation automatique est d'abord effectuée sur les phrases enregistrées afin de délimiter tous les phonèmes présents sur le signal. Cette étape permet ensuite à l'algorithme de sélection, lors de la synthèse, de sélectionner les bonnes unités, d'un point de vue acoustique. Toutefois, même si la segmentation ne comporte pas énormément d'erreurs, il est nécessaire d'effectuer une vérification manuelle. Ce travail est très minutieux et demande beaucoup de concentration et de temps. Il s'agit, dans cette phase, de vérifier les bonnes frontières de phonèmes en écoutant les unités et en faisant attention à n'avoir que le phonème lui-même entre deux frontières et pas un morceau du suivant. Il faut également placer une marque de milieu de phonème. En effet, la synthèse ayant pour unité minimale le diphone, chaque phonème est concaténé sur sa partie stable, au milieu. Sur certains phonèmes comme les voyelles, la marque se trouve vraiment au milieu du phonème, cependant sur des phonèmes plosifs, le /t/ par exemple, caractérisés par une occlusion puis une plosion, la marque de concaténation doit être placée dans l'occlusion, juste avant la plosion. Les problèmes rencontrés durant cette phase sont facilement contournables. En effet, ils sont souvent liés à des oublis de mots ou de pause ou des erreurs de prononciation qui sont passées à la trappe lors des enregistrements.
Par exemple, en français, la phrase «*L'heure d'arrivée escomptée est 18 heures 1 »* a été enregistrée «*L'heure d'arrivée escomptée est 18 heures une»* alors que l'on attendait «*L'heure d'arrivée escomptée est 18 heures un»*, il a donc fallu changer l'étiquette du phonème [UN] pour la remplacer par les étiquettes [U] et [N]. La phonétique doit donc toujours correspondre au signal de parole, même si une erreur de prononciation a été commise à l'enregistrement, il faut changer la phonétique pour que les deux concordent. Par ailleurs, s'il est impossible de trouver la bonne phonétique, soit parce que l'enregistrement est inaudible, si un bruit gênant se fait entendre, etc, il est possible d'exclure des unités en ajoutant un point d'exclamation avant ou après, selon la moitié de phonème que l'on souhaite exclure.

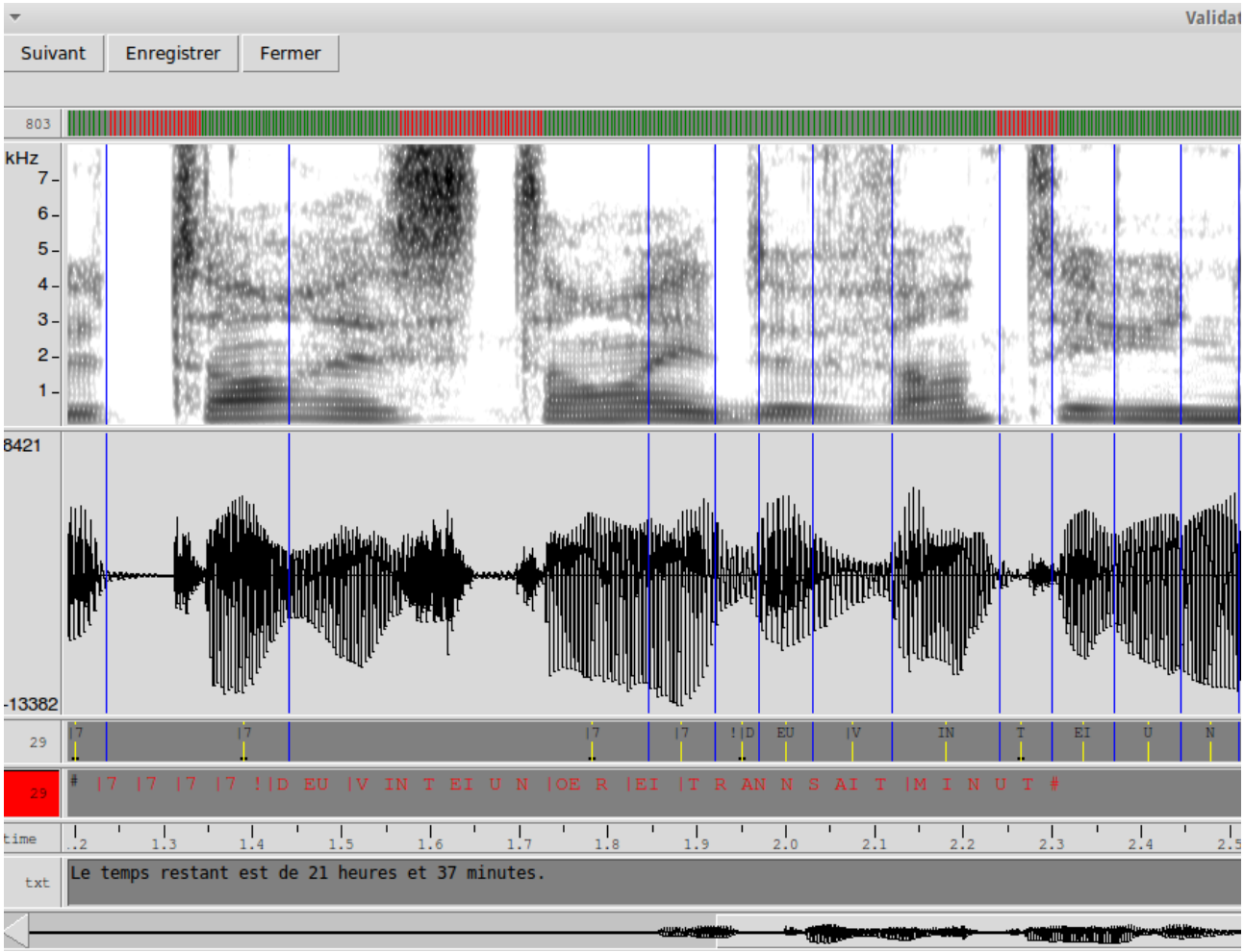


Figure 15 : outil de segmentation

Enfin, après la phase de segmentation, on réalise une phase de sélection des parties fixes. Cette étape consiste à écouter toutes les phrases pour ne garder que celles qui paraissent enjouées et agréables. En effet, les parties fixes étant restituées telles quelles, il est nécessaire d'éliminer les moins bonnes. La sélection est donc très subjective et ne se fait qu'à l'oreille.

Nous avons évoqué plus haut que les enregistrements en américain avaient été réalisés selon deux types de regroupements et les enregistrements des numéros de téléphone en français selon deux structures, deux dictionnaires différents ont donc été générés pour chaque langue : Paul_A regroupements riches, Paul_B regroupements pauvres, Camille_A numéros seuls, Camille_B numéros entiers. Des tests ont donc été menés au sein de l'entreprise pour savoir si les différences étaient très audibles, et surtout si elles étaient significatives et entraînaient une différence de qualité. Nous avons sélectionné douze phrases en français, et dix phrases contenant des nombres et dix autres contenant des numéros de téléphone en américain, et ce, dans chaque dictionnaire. Nous

avons ensuite envoyé toutes ces phrases à neuf personnes afin qu'elles disent, pour chaque phrase, laquelle semblait la plus agréable et la plus naturelle. Les tests ont finalement révélé une préférence pour le dictionnaire Paul_A en ce qui concerne les nombres longs comme ceux sélectionnés dans les phrases avec miles, mais une préférence pour le dictionnaire Paul_B en ce qui concerne les nombres courts, avec plusieurs groupes de souffles (virgule) dans les phrases avec des numéros de téléphone. On peut donc penser que les phrases avec plusieurs groupes de souffles, relativement courts, ne nécessitent pas forcément de regroupements très riches contrairement aux phrases contenant des nombres longs et avec un seul groupe de souffle. D'autre part, en ce qui concerne les tests sur les dictionnaires Camille_A et Camille_B, on remarque que la majorité des auditeurs préfèrent le dictionnaire B, dans lequel les numéros ont été enregistrés en entiers, sans couper à chaque groupe de souffle. Le fait de ne pas couper les enregistrements donne peut-être un effet plus fluide, plus homogène, bien que plusieurs auditeurs aient fait la réflexion que les phrases étaient peut-être un peu monotones. Au contraire, les phrases du dictionnaire A, dans lequel les numéros ont été enregistrés séparément, apparaissent plus vivantes mais moins naturelles à cause des différences d'intensité et de hauteur de voix.

Dans un autre contexte, j'ai pu assister à l'enregistrement de la voix d'une patiente allant subir une laryngectomie qui l'empêcherait de parler. Comme déjà expliqué, l'application de synthèse vocale VoxMed, proposée par Voxygen représente, pour les patients, une façon de garder leur voix naturelle, qu'ils la perdent totalement ou seulement partiellement. Le fait de pouvoir entendre et faire entendre leur propre voix rassure les patients, mais aussi leurs proches.

La personne en charge d'enregistrer le patient est autorisée par l'hôpital à utiliser une pièce d'audiométrie (isolation phonique) ainsi qu'une table et des chaises. Elle doit cependant apporter tout le matériel nécessaire à l'enregistrement :

- Machine VoxMed : VoxMed est le nom de l'interface qui sert à enregistrer les patients
- Ecran locuteur : cet écran renvoie les phrases que le patient doit lire, accompagnées d'un signal d'enregistrement.
- Carte son.
- Micro pour le locuteur.
- Enceinte externe pour permettre au locuteur d'entendre un modèle avant de prononcer une phrase.

Le patient attend que la phrase à enregistrer et le signal d'enregistrement (cadre rouge autour de la phrase) apparaissent sur son écran, puis peut parler en essayant de respecter au mieux les consignes qu'on lui a données :

- respecter l'intonation montante (dans le cas d'une fin de groupe de souffle avec une virgule ou un point d'interrogation) ou descendante (dans le cas d'une fin de groupe de souffle avec un point). Une flèche rouge présente juste après la phrase donne l'information au patient.
- respecter le texte et ne pas dire autre chose que ce qui est écrit.
- ne pas commencer à parler avant le signal d'enregistrement.

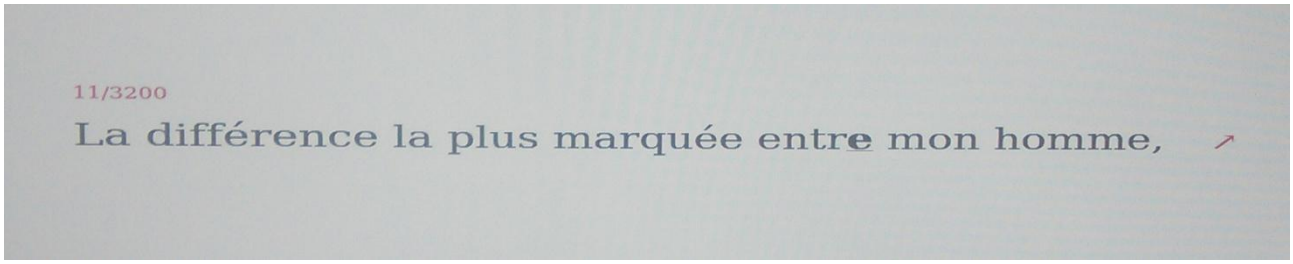


Figure 16 : interface script locuteur

Lorsque ces consignes ne sont pas respectées, l'opérateur doit savoir juger s'il est nécessaire ou non de réenregistrer la phrase. La difficulté de ce jugement réside dans le fait que les patients sont souvent des personnes âgées, malades, qui ont tendance à fatiguer rapidement. Il est donc parfois délicat de demander à une personne de répéter plusieurs fois la même phrase pour un détail qui peut lui sembler insignifiant, l'intonation par exemple. Cependant, pour que la synthèse soit de bonne qualité et intelligible, il est nécessaire de pouvoir couvrir tout ce dont on a besoin. Si, donc, par exemple, il manque une intonation montante pour tel ou tel son, on risque de se retrouver avec une unité manquante. La difficulté de la mission de l'opérateur se situe donc dans la décision de reprendre un enregistrement, qui fatiguerait potentiellement le patient, ou de laisser passer une variante non attendue, qui entraînerait potentiellement un manque de couverture.

Si nous nous recentrons sur le cas de l'enregistrement que j'ai pu suivre, le corpus était composé de 1000 phrases, extraites de différentes bases de données. Nous avons eu la chance d'avoir eu affaire à une patiente avec un bon niveau de lecture, elle n'a donc eu aucun mal à lire les phrases de façon naturelle, sans hésiter ou "buter" sur les mots. Nous avons toutefois rencontré deux types de problèmes :

- des “erreurs” d’intonation,
- des “erreurs” liées aux annotations nouvelles du script, qui ne lui ont pas été expliquées.

Nous avons tenté de donner à la patiente une sorte d’automatisme pour la pousser à effectuer les intonations attendues en la reprenant et en lui expliquant pourquoi nous devions réenregistrer. Malheureusement, après quelques phrases, les consignes étaient oubliées… Nous avons donc décidé de ne plus la reprendre et de considérer que nous trouverions les unités manquantes à d’autres endroits dans le corpus. En ce qui concerne les “erreurs” liées aux annotations, la patiente les a certainement faites par manque d’information. Ces annotations sont présentes pour forcer le locuteur à effectuer la prononciation attendue. Par exemple, pour la suite “quatorz**e** minutes” écrite ainsi, on attend du locuteur qu’il prononce le “e” muet. Pour une suite telle que “les grands -z- enfants”, on demande au locuteur de faire la liaison /z/ entre “grands” et “enfants”. On peut également solliciter le locuteur à ne pas effectuer une liaison en apposant le signe // comme dans la suite “ne pas parler// à haute voix”, etc. Pour gérer ces problèmes de prononciation, nous n’avons pas établi de règles, nous avons décidé de traiter au cas par cas, selon les unités déjà enregistrées et l’état de fatigue de la patiente.

Au-delà des difficultés “linguistiques”, nous avons pu remarquer que la patiente n’était pas très réceptive à nos commentaires et qu’elle ne semblait pas réellement se rendre compte de la manière dont sa voix allait être restituée en synthèse. En effet, il peut être très compliqué de comprendre, pour des patients qui sont souvent âgés, ce que les phrases relativement insensées qu’ils lisent et enregistrent peuvent avoir à voir avec la synthèse de leur voix. Nous avons tout de même pu présenter l’application VoxMed à la patiente sur la tablette de l’entreprise mais elle n’a pas montré l’envie de l’utiliser.

Finalement, un des grands enjeux de la synthèse est de trouver le juste milieu entre expressivité et naturel. En effet, nous avons pu le constater en faisant les tests sur les différents dictionnaires de voix contextuelles et avec la patiente, il ne suffit pas d'être simplement naturel ou expressif, il s'agit plutôt de réussir à concilier les deux pour faire une synthèse qui ne soit pas seulement expressive mais inintelligible ou naturelle mais monotone.

B. Programmation

Durant ce stage, j'ai pu apprendre, non seulement à lire, mais aussi à écrire des programmes informatiques dans un langage inconnu : python. En effet, les programmes de génération de phrases des voix SNF étant écrits en python, j'ai dû apprendre à les déchiffrer pour pouvoir les compléter. De plus, j'ai eu l'occasion, avec l'aide de Laure Charonnat, ingénieur R&D, d'écrire deux programme différents en python :

- class_wav.py
- make_alias.py

Class_wav.py a été écrit pour faciliter la sélection des parties fixes. En effet, pour pouvoir être efficace, il faudrait pouvoir écouter les phrases contenant les mêmes parties fixes les unes à la suite des autres. Or, elles ne sont pas classées dans cet ordre mais dans l'ordre dans lequel elles ont été enregistrées. Chercher, donc, à chaque fois les phrases avec les mêmes parties fixes est une perte de temps. Class_wav.py a donc été construit pour permettre, à partir des attributs meta du PLS qui confèrent à chaque partie fixe une identité propre, de classer les phrases par partie fixe, dans des répertoires séparés. L'algorithme adopté est donc le suivant :

- *lecture du PLS*
- *extraction des attributs graphèmes et des attributs meta de chaque entrée*
- *création d'un répertoire portant le nom du meta*
- *recherche dans la BDS des fichiers texte (script locuteur) qui contiennent l'attribut graphème correspondant à son attribut meta*
- *création d'un lien symbolique sur le fichier wav correspondant.*

Le deuxième script écrit, toujours avec l'aide de Laure, a été fait pour faciliter le passage d'un fichier PLS «normal», c'est-à-dire avec pour chaque graphème une transcription phonétique dans une balise <phonème>, à un PLS «alias». Le PLS «alias» permet d'avoir, pour chaque graphème, la prononciation exacte attendue dans une balise <alias>.

<lexeme>

<grapheme>Vous êtes à l'arrêt</grapheme>

<phoneme vox:meta="Vous_etes_a_larret">7##7##7##7</phoneme>

</lexeme>

<lexeme>

<grapheme>Vous êtes à l'arrêt</grapheme>

<alias vox:meta="Vous_etes_a_larret" prefer="true">Vous êtes à l'arrêt</alias>

</lexeme>

Pour ce faire, l'algorithme adopté est le suivant :

- *Lecture du PLS*
- *Écriture des lignes du PLS dans le fichier de sortie*
- *Pour chaque lexème, écriture de l'élément et de ses attributs*
- *Pour chaque fils de lexème, écriture du graphème*
- *Pour chaque phonème avec un attribut, écriture de l'alias avec l'attribut meta*

Les difficultés liées à ce programme ont amené à réfléchir sur la construction du PLS lui-même. En effet, nous nous sommes rendu compte de la nécessité de créer une entrée propre à chaque graphème dès lors que l'on a la moindre différence acoustique. Par exemple, le cas des pluriels et singuliers en anglais entraîne forcément la création de deux entrées dans le PLS. En effet, lorsque l'on a deux graphèmes acoustiquement différents dans un lexème, on ne peut pas demander au système de choisir l'une ou l'autre des prononciations, il faut alors créer deux entrées. Nous avons finalement décidé de traiter trois cas de figure dans notre programme :

- cas n°1 : un élément <alias> dans le PLS : on veut réécrire tout simplement cette ligne telle quelle
- cas n°2 : plusieurs éléments <grapheme> pour un élément <phoneme> : on veut dupliquer l'élément <lexeme> pour avoir une entrée pour chaque graphème avec les attributs de l'élément <phoneme> et le texte de graphème dans l'élément <alias>. Il faut considérer que la moindre différence acoustique entre deux graphèmes entraîne systématiquement la création de deux lexèmes différents.
- cas n°3 : pas d'élément <phoneme> : on considère qu'on ne peut pas traiter ce cas (qui a peu de chances de se présenter), on renvoie donc simplement un message.

C. *Mavoa, de l'écriture des scénarios à l'utilisation des scripts d'affichage*

En ce qui concerne l'application Mavoa, nous avons pu prendre part à toutes les étapes du processus, de l'écriture des scénarios que nous avons déjà détaillée, à l'utilisation de scripts python permettant d'afficher quels segments des phrases enregistrées ont été sélectionnés pour fabriquer la phrase de synthèse. Cet affichage permet à l'utilisateur de mieux comprendre le principe de la synthèse vocale. Le processus d'écriture des scénarios ayant déjà été présenté nous ne reviendrons pas dessus ici, nous nous pencherons donc plus particulièrement sur la manière dont sont construits les scripts d'affichage, les problèmes rencontrés et les solutions, ainsi que sur la mise à jour d'une procédure d'ajout des scénarios à l'application.

Comme présentée en partie 2_III_A de ce travail, l'application Mavoa consiste à provoquer un effet de surprise chez l'utilisateur qui enregistre des phrases et entend en retour une phrase non prononcée, avec sa propre voix. Sur l'écran final, la phrase de synthèse et les phrases enregistrées apparaissent enrichies de segments de couleur représentant les unités sélectionnées pour construire la phrase de synthèse.

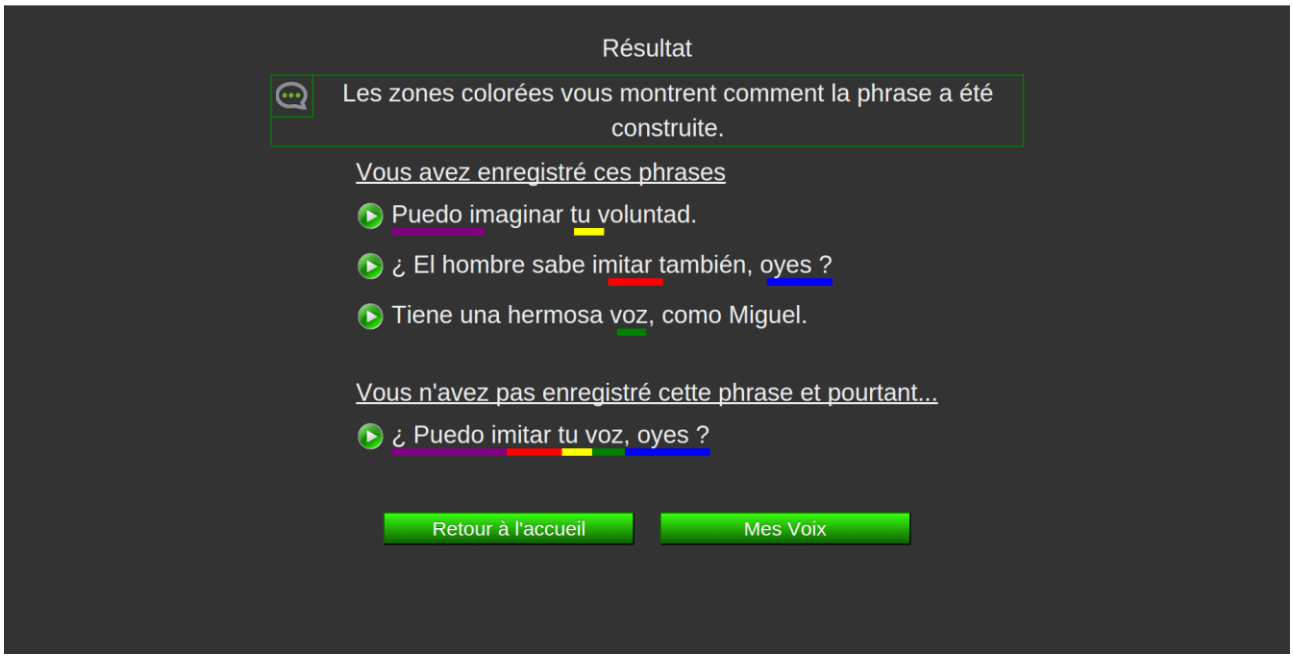


Figure 17 : Affichage des segments de couleurs représentant la sélection des unités pour la création de la phrase de synthèse.

Une fois les scénarios écrits et les unités à sélectionner «prédites», l'objectif est de compléter le script d'affichage des couleurs que nous allons maintenant détailler (il n'est pas nécessaire de lire et de comprendre le script entier ci-dessous, toutes les étapes seront ensuite explicitées) :

scenario["markupinput"] = [{'num':1,'txt':u"""<m></m>Dices que <m class="p2">v</m>ive<m>s</m> en <m class="p4">P</m>en<m>s</m>acola ?<m></m>"""},

{'num':2,'txt':u"""<m class="p1"></m>Quien <m>v</m>iaje <m class="p3">s</m>in <m>p</m>ren<m class="p5">s</m>a<m>r,</m> <m class="7">p</m>uede <m>d</m>escuvrir.<m></m>"""},

{'num':3,'txt':u"""<m></m>Sin informaciones <m class="p6">n</m>o <m>p</m>uedo pre<m class="p8">d</m>ecir que vive.<m></m>"""},]

scenario["output"] = u"Quien vive sin pensar, no puede decir que vive."

scenario["markupoutput"] = u"""<m></m>Quien <m>v</m>ive <m>s</m>in <m>p</m>e<m>n</m><m>s</m>a<m>r,</m> <m>n</m>o <m>p</m>uede <m>d</m>ecir que vive.<m></m>"""

scenario["parts"] = (7,6,5,4,6,7,7,14)

La première étape concerne les trois premiers blocs :

scenario["markupinput"] = [{'num':1,'txt':u"""<m></m>Dices que <m class="p2">v</m>ive<m>s</m> en <m class="p4">P</m>en<m>s</m>acola ?<m></m>"""},

{'num':2,'txt':u"""<m class="p1"></m>Quien <m>v</m>iaje <m class="p3">s</m>in <m>p</m>ren<m class="p5">s</m>a<m>r,</m> <m class="7">p</m>uede <m>d</m>escuvrir.<m></m>"""},

{'num':3,'txt':u"""<m></m>Sin informaciones <m class="p6">n</m>o <m>p</m>uedo pre<m class="p8">d</m>ecir que vive.<m></m>"""},]

Ici, le but est de définir le début et la fin de chaque segment de couleur pour chaque phrase à enregistrer. Chaque segment est coupé sur un milieu de phonème pour bien rendre compte de la zone stable de chaque phonème, c'est pourquoi chaque balise est nommée <m> qui signifie le milieu. Prenons par exemple la première phrase :

¿Dices que vices en Pensacola ?

{'num':1,'txt':u"""<m></m>¿ Dices que <m class="p2">v</m>ive<m>s</m> en <m class="p4">P</m>en<m>s</m>acola ?<m></m>"""}

En orange, on peut voir que l'unité sélectionnée s'étend du milieu du /v/ au milieu du /s/ qui sont tous les deux notés entre les balises <m> et </m>. Pour chaque début de segment, l'argument «class="pX"» est attribué afin de déterminer, dans la phrase de synthèse, en quelle position ce segment apparaîtra. Dans cet exemple, le segment «vives» apparaîtra en deuxième position dans la phrase de synthèse.

La deuxième étape concerne les deuxième et troisième blocs du script :

scenario["output"] = u"Quien vive sin pensar, no puede decir que vive."

scenario["markupoutput"] = u"""<m></m>Quien <m>v</m>ive <m>s</m>in <m>p</m>e<m>n</m><m>s</m>a<m>r,</m> <m>n</m>o <m>p</m>uede <m>d</m>ecir que vive.<m></m>"""

Ici, il s'agit simplement de donner la phrase de synthèse (en bleu) telle quelle puis dans un deuxième temps, de définir, toujours grâce aux balises <m> et </m> les segments à souligner. Il s'agit donc là de reprendre les segments choisis dans les phrases du corpus et de les reproduire dans la phrase de synthèse. Par exemple, on peut voir, en orange, le segment «vives» repris ici, s'étendant du milieu du premier /v/ de «vive» au milieu du /s/ de «sin».

Enfin, la troisième étape concerne le dernier bloc :

scenario["parts"] = (7,6,5,4,6,7,7,14)

Le but ici est de déterminer le nombre de phonèmes contenu dans chaque segment de la phrase de synthèse. Il suffit donc de prendre la phonétique de la phrase et de découper selon les parties choisies puis de compter les phonèmes et de les reporter entre les parenthèses.

 # K J E M B
B II BB E S
S I M P
P E N S
S AA R # #
 # # N OO P
P W EE DD E DD
D E Z II R K E BB II BB E #

Ici, par exemple, on retrouve, en orange, le segment «vive s» apparaissant en deuxième position et composé de six phonèmes : [B], [II], [BB], [E], un espace, [S].

Toutefois, certains cas ont posé quelques difficultés qu'il a fallu contourner. Certains scénarios demandent l'utilisation d'un même segment plusieurs fois dans la même phrase de synthèse, il faut donc donner deux emplacements dans les class (exemple : class= «p1 p2»).

{'num':2,'txt':u"""<m class="p7"></m>Sotto <m>l</m><m class="p2 p8">a</m>pan<m>c</m>hina in <m class="p5">c</m>am<m>p</m>agna a <m class="p4 p10">p</m>ra<m>c</m>chia<m></m>"""},

Cependant, d'autres segments sont utilisés une fois entièrement et une autre fois seulement partiellement. Il faut donc faire un découpage du segment pour faire apparaître chaque «morceau» seul. Exemple d'un scénario italien :

Sopra la panca la capra campa, sotto la panca la capra crepa.

Sopra la palanca la capricciosa muffa.
Sotto la panchina in campagna a Pracchia.
Produci consuma crepa.
Quando ti arriva la vampa,

Le segment «Sotto la panca» contenu dans la deuxième phrase du corpus (en rose) est utilisé une fois en entier (en orange dans la phrase de synthèse) et une autre fois partiellement avec seulement «la panc» (en bleu dans la phrase de synthèse). Il faut donc séparer le segment en deux parties dans le script d'affichage, un avec «Sotto l» et un autre avec «la panc».

scenario["markupinput"] = [{'num':1,'txt':u"""<m class="p1"></m>Sopra <m>l</m>a palan<m class="p3 p9">c</m>a la ca<m>p</m>ricciosa muffa.<m></m>"""},

{'num':2,'txt':u"""<m class="p7"></m>Sotto <m>l</m><m class="p2 p8">a</m>pan<m>c</m>hina in <m class="p5">c</m>am<m>p</m>agna a <m class="p4 p10">p</m>ra<m>c</m>chia<m></m>"""},

{'num':3,'txt':u"""<m></m>Produci consuma <m class="p11">c</m>repa.<m></m>"""},

{'num':4,'txt':u"""<m></m>Quando ti arriva la vam<m class="p6">p</m>a,<m></m>"""},]

scenario["output"] = u"Sopra la panca la capra campa, sotto la panca la capra crepa."

scenario["markupoutput"] = u"""<m></m>Sopra <m>l</m>a pan<m>c</m>a la ca<m>p</m>ra<m>c</m>am<m>p</m>a,<m></m> sotto <m>l</m>a pan<m>c</m>a la ca<m>p</m>ra <m>c</m>repa.<m></m>"""

Pour ce genre de segments apparaissant plusieurs fois, une fois en entier et une fois avec seulement une partie (ex : Sotto la panc), il y avait en fait deux solutions :

- mettre deux class sur le segment entier. Résultat : au passage de la souris sur les segments de la phrase de synthèse, le segment entier se colore, dans le cas de son apparition totale ET partielle. Inconvénient : dans le cas de l'apparition partielle,

l'utilisateur ne voit pas ce qui a EXACTEMENT été sélectionné.

- couper le sandwich «Sotto la panc» en «Sotto l» et «a panc». Résultat : au passage de la souris sur les segments de la phrase de synthèse, les segments se colorent selon leur correspondance. Inconvénient : comme on ne peut pas attribuer à la fois une fin de segment et une class (début) à une même lettre, il y a un petit espace vide dans le découpage.

Ce travail a été réalisé dans plusieurs langues : anglais américain, italien et espagnol.

D'autre part, l'intégration de ces scénarios à l'application et la création d'une nouvelle langue n'étant pas réellement définies, j'ai mis à jour une procédure d'ajout que vous pourrez trouver en annexe (3. Procédure d'ajout d'une nouvelle langue et de nouveaux scénarios dans Mavoa) .

III. Perspectives

Certaines tâches ont été évoquées durant le stage mais n'ont pas pu être réalisées faute de temps. En effet, il a été question de travailler sur une voix contextuelle en Wolof, langue du Sénégal, afin de construire une calculatrice « parlante ». Dans ce cas, les parties variables auraient pu être tous les nombres possibles à entrer dans la calculatrice, et les parties fixes à répertorier dans le PLS ; les différentes opérations possibles (addition, multiplication, division, etc).

D'autre part, il a également été question de construire une interface pour tablette sur android permettant aux commerciaux de pouvoir faire une démonstration « propre » des voix contextuelles pour l'automobile en français et en américain. Le but d'une telle interface est de pouvoir donner aux clients la possibilité de tester en temps réel les phrases qu'ils souhaitent entendre, tout cela sur un support mobile et simple d'utilisation.

***

Finalement, on peut se rendre compte ici de la notion de pluridisciplinarité qui a rythmé mon stage pendant six mois. En effet, j'ai eu la chance de participer aux tests et à l'amélioration de différentes applications, dans différentes langues, ce qui m'a permis d'être à l'aise dans plusieurs domaines. En définitive, cette dernière partie montre la dynamique de ce stage et dans quelles mesures sa variété m'a permis de me rendre compte des différents processus de création de voix, qu'ils diffèrent par leur langue ou par leur domaine d'application. Je pense que la richesse de cette expérience réside dans cette notion même de pluridsciplinarité. Le fait d'avoir pu, pour chaque tâche accomplie, expérimenter les processus dans leur ensemble et prendre part à chacune des étapes, a permis d'ordonner des démarches encore floues.

Aussi, pour conclure ce chapitre polyglote et pluridisciplinaire sur un bilan personnel, je peux dire que le fait d'avoir pu travailler sur des langues diverses m'a permis d'apprendre à ne pas me contenter de faire ce que je qualifierais de «synthèse de surface» qui ne rendrait pas compte de la diversité des langues. En effet, nous avons pu voir ici que, selon les langues sur lesquelles nous travaillions, les règles et les procédés différaient et, par conséquent, les problèmes rencontrés et les solutions trouvées ne pouvaient pas simplement être réutilisées d'une langue à l'autre. J'ai donc appris ici à entrer réellement dans les processus et à chercher à comprendre, avant tout, le fonctionnement des langues pour travailler.

Par ailleurs, le fait d'avoir utilisé plusieurs langues dans des processus et des applications différentes a été pour moi un moyen de varier les plaisirs et de ne pas tomber dans une routine pesante. Il ne s'agit donc pas simplement de traduire la langue elle-même mais également tout ce qui l'entoure. De plus, la diversité des nationalités et des cultures des employés de Voxygen permet à tout le monde de pouvoir discuter et travailler en équipe plus facilement et dans des conditions plus optimales. En effet, les travaux qui m'étaient confiés ont toujours été menés en relation avec une partie de l'équipe, ce qui a souvent entraîné des discutions intéressantes sur la façon de régler les problèmes. Le fait de toujours travailler avec quelqu'un empêche de s'isoler et permet d'avoir l'avis d'un œil extérieur en cas de problème. Aussi, l'équipe, dans sa totalité, a toujours été ouverte à répondre aux différentes interrogations qui se sont présentées à moi et a toujours pris le temps de m'expliquer et de me ré-expliquer les choses.

Conclusion

Nous nous sommes interrogés dans ce mémoire, dans un premier temps, sur la pertinence de l'utilisation d'une unité de longueur variable : le sandwich, puis dans un deuxième temps, sur l'importance du choix des corpus utilisés pour sélectionner les unités à couvrir, pour la construction d'une synthèse de haute qualité.

Pour répondre à ces interrogations, après une présentation de la synthèse vocale et de la technologie de sytnhèse vocale par sélection d'unités utilisée par Voxygen, nous avons choisi de présenter et de définir précisément le sandwich puisqu'il constitue un des points centraux de ce travail. A travers les travaux d'écriture de scénarios pour une application de synthèse vocale grand public, nous avons pu montrer la pertinence de l'utilisation de cette unité qui permet de protéger les phonèmes fragiles des concaténations.

La sélection du sandwich vocalique n'étant pas suffisante pour garantir une synthèse de qualité, nous avons tenté de démontrer dans quelles mesures le choix des corpus de départ, à partir desquels la sélection des unités à couvrir est définie, constitue une étape déterminante selon l'objectif visé. La création de voix contextuelles (Slot'N'Fill), mission principale de ce stage, a servi de point de départ à la démonstration de l'importance du choix des corpus selon les domaines d'application.

Nous avons dans un dernier temps choisi d'illustrer ces procédés par les différents travaux effectués tout au long du stage qui ont montré l'intérêt de suivre un même fil conducteur, constamment dans un but qualitatif : Mavoa, Slot'N'Fill, Voix patient.

Dans une autre dimension, les questions posées s'appliquent ici aux différentes langues sur lesquelles nous avons pu travailler. Nous avons donc pu voir comment les variations phonétiques, linguistiques et prosodiques peuvent engendrer des différences dans la prise en compte de certains phénomènes.

En nous appuyant donc sur les recherches effectuées en parallèle de ce travail et sur les

nombreuses discussions avec les membres de l'équipe, notamment Chiara et Laure, nous avons pu mener à bien les tâches confiées et l'écriture de ce travail.

Cette dernière remarque me permets de rebondir sur le bilan personnel que je souhaite faire de ce stage. La mission qui m'a été confiée n'étant pas clairement définie, je n'ai pas pu être totalement autonome dans toutes les tâches que j'ai menées. Cependant, ce manque d'autonomie m'a permis de pouvoir en apprendre d'avantage en travaillant fréquemment avec des membres de l'équipe. Finalement, le stage que j'ai effectué durant ces six mois m'a permis de m'ouvrir à un domaine peu exploré de façon pratique dans la formation en Master Industries De la Langue, et de me rendre compte des connaissances que je pouvais mettre en pratique dans un environnement professionnel. J'ai également pu apprendre beaucoup au contact de chacun des membres de l'équipe de par leur culture et parcours variés. En définitive, je pense que le point fort de mon stage a été la diversité des tâches effectuées tant au niveau des langues que des domaines applicatifs. J'ai pu observer des phénomènes différents d'un point de vue linguistique mais également technique du fait de la quantité de traitements nécessaires à la construction d'une voix de synthèse. En effet, même si mon stage n'a pas constitué une tâche de « fond », j'ai eu la chance de prendre part à toutes les étapes de chaque processus auquel j'ai participé, ce qui m'a permis d'avoir une vue d'ensemble de la synthèse vocale et de pouvoir mieux appréhender la réalité professionnelle.

Références

I. Bibliographie

Boëffard, O., Emerard, F. (1997). *Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm*. (Eurospeech). Repéré à http://www.mirlab.org/conference_papers/International_Conference/Eurospeech%201997/pdf/th4c/a0056.pdf

Bozkurt, B., Dutoit, T., Pagel, V. (2002). *Synthèse vocale par sélection d'unité : une méthode pour la redéfinition de la courbe intonative*. (Journées d'Étude sur la parole, Nancy). Repéré à http://tcts.fpms.ac.be/publications/papers/2002/jep2002_bbtdvp_fr.pdf

Cadic, D., Boidin, C., d'Alessandro, C. (2010). *Towards optimal TTS corpora*. (*Text, speech and dialogue : 17th International Conference on Language Resources and Evaluation*, Valetta, Malte). Repéré à http://hnk.ffzg.hr/bibl/lrec2010/pdf/608_Paper.pdf

Cadic, D., Boidin, C., d'Alessandro, C. (2009). *Vocalic sandwich, a unit designed for unit selection TTS*. (Interspeech, Brighton). Repéré à http://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_2079.pdf

Cadic, D. (2011). *Optimisation du procédé de création de voix en synthèse par sélection*. (Thèse de doctorat non publiée, Université Paris-Sud 11).

Falaise, A. (2005). *Constitution d'un corpus de français tchaté*. (Récital 2005, Dourdan). Repéré à http://www.atala.org/taln_archives/RECITAL/RECITAL-2005/recital-2005-long-010.pdf

Feugère, L. (2013). *Synthèse par règles de la voix chantée contrôlée par le geste et application musicales*. (Thèse de doctorat, Université Pierre et Marie Curie, Paris.) Repéré à https://tel.archives-ouvertes.fr/file/index/docid/926980/filename/these_Lionel_Feugere.pdf

Florent, X. (2011). *Synthèse Vocale : intégration du français au système Mary Text-To-Speech.* (Mémoire de master 2, Université Pierre et Marie Curie, Paris). Repéré à ftp://ftp.ircam.fr/pub/IRCAM/equipes/repmus/Atiam/Florent.pdf

François, H. (2002). *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*. (Thèse de doctorat, Université Rennes 1). Repéré à ftp://ftp.irisa.fr/techreports/theses/2002/francois.pdf

Krul, A., Damnati, G., Moudenc, T., Yvon, F. (2006). *Constitution d'un corpus textuel basé sur la*

divergence de Kullback-Leiber pour la synthèse par corpus. (Journées d'Étude sur la Parole). Repéré à http://jep2006.irisa.fr/openconf/author/final/final-129.pdf

Lonchamp, F., (2010). La transcription phonétique du français. (Université Nancy 2). Repéré à http://francois.lonchamp.free.fr/TranscriptionPhonetique/transcriptionPhonetique.pdf

Lutz, M. (2013). *Learning python : Powerful Object-Oriented Programming*. (5ème édition). O'Reilly Media.

Moulines, E., Cappé, O. *Synthèse de la parole à partir du texte*.

Snoeren, N. D. (2005). *Variations phonologiques en production et perception de la parol : le phénomène de l'assimilation*. (Thèse de doctorat, Université Paris 5 René Descartes). Repéré à http://www.afcp-parole.org/doc/theses/theseNS05.pdf

II. Documentation Voxygen

Charonnat, L. (2015). *Spécification des Bases de Données de Signal Segmenté*.

Emerard, F. (2003). *Spécifications de la sélection des unités pour la synthèse par corpus*.

HRA (2013). *Manuel technique pour la validation phonétique et la vérification de la segmentation des bases de données*.

PYJ (2015). *BARATINOO exceptions lexicon manual*.

VoxMed Mode d'emploi.

VoxMed Installation du matériel.

(2014). *Mavoa_Demo : documentation*.

III. Sitographie

Bagshaw, P., Burnett, D. C., Carter, J., Scahill, F. (2008). *Pronunciation Lexicon Specification (PLS) Version 1.0*. Repéré à http://www.w3.org/TR/pronunciation-lexicon/ Date de la dernière consultation : 03/09/2015

Mudry, A. (2015). *L'oreille, ses maladies et ses traitements. Examen de l'oreille : l'audiométrie*.

Repéré à http://www.oreillemudry.ch/l%E2%80%99audiometrie/ Date de la dernière consultation : 03/09/2015

Prudon, R., d'Alessandro, C., Boula de Mareüil, P. (2003). S*ynthèse par sélection, prosodie et qualité vocale*. Repéré à http://www.limsi.fr/RS2003FF/CHM2003/PERSI2003/PERSI8/persi8.html Date de la dernière consultation : 03/09/2015

Rohart, M. N. (2006). *Qu'est-ce que la synthèse vocale ?* (Université Paris X). Repéré à http://www.technolangue.net/article.php3?id_article=275 Date de la dernière consultation : 03/09/2015

Python. https://www.python.org/ Date de dernière consultation : 03/09/2015

OpenClassroom. (2015). *Apprenez à programmer en python*. Repéré à https://openclassrooms.com/courses/apprenez-a-programmer-en-python Date de la dernière consultation : 03/09/2015

Ruter, W. (2005). *The International Phonetic Alphabet.* Repéré à http://westonruter.github.io/ipa-chart/keyboard/ Date de la dernière consultation : 03/09/2015

Voxygen. https://www.voxygen.fr/ Date de la dernière consultation : 03/09/2015

Voxygen. *Wiki Voxygen*. Date de dernière consultation : 03/09/2015
- *Création des voix*
- *Slot'N'Fill (voix contextuelles)*
- *Hauts-niveaux - Polyglot : lancement de Baratinoo et lecture des sorties*

Wikipedia (2015). *Help:IPA for Spanish.* Repéré à https://en.wikipedia.org/wiki/Help:IPA_for_Spanish Date de la dernière consultation : 03/09/2015

Wikipédia (2015). *Transcripción fonética del español con el AFI.* Repéré à https://es.wikipedia.org/wiki/Transcripci%C3%B3n_fon%C3%A9tica_del_espa%C3%B1ol_con_el_AFI#cite_note-N-20 Date de dernière la consultation : 03/09/2015

Wikipédia (2015). *Fonología del español.* Repéré à https://es.wikipedia.org/wiki/Fonolog%C3%ADa_del_espa%C3%B1ol Date de la dernière consultation : 03/09/2015

Wikipédia (2014). *Anexo:Comparación de los inventarios fonéticos latino y español*. Repéré à https://es.wikipedia.org/wiki/Anexo:Comparaci%C3%B3n_de_los_inventarios_fon%C3%A9ticos_latino_y_espa%C3%B1ol Date de la dernière consultation : 03/09/2015

Williamson, G. (2014). *Speech and langage therapy information : syllabic consonants.* Repéré à http://www.sltinfo.com/syllabic-consonants/ Date de la dernière consultation : 03/09/2015

Table des illustrations

Table des annexes

Annexes

A. Scénarios espagnol pour l'application Mavoa

En italique, la phrase de synthèse. En couleur, les segments sélectionnés pour construire la phrase de synthèse.

¿ Puedo imitar tu voz, oyes ?

Puedo imaginar tu voluntad.
¿ El hombre sabe imitar también, oyes ?
Tiene una hermosa voz, como Miguel.

Digo cosas que nunca has dicho.

Digo con mis palabras lo que has dicho.
Hay cosas que no sabes de mi.
Se que nunca has distribuido los papeles.

Ser o no ser, ésa es la cuestión.

¿ Sero estar, estar o ser, como saber ?
Postre casero no sigue nuestra sugestión.
Esa es la cuestación para la iglesia.

A buen hambre, no hay pan malo.

El alambre, no se rompe.
A buen ámbito, no hay problema.
Coliflor con pan me parece malo.

Una pintura es una fotografía hecha a mano.

Una caligrafía hecha por un artista.
Una pintura es una fuerte obra trabajada a mano.
Es un fotograma hecho a una fecha amistosa.

Quien vive sin pensar, no puede decir que vive.

Dices que vives en Pensacola ?
Quien viaje sin prensar, puede descubrir.
Sin informaciones, no puedo predecir que vive.

Lleva tiempo llegar a ser jóven.

Lleva tu pancarta.
Que el tiempo llegara sobre el sol.
Llevarse para ser jóven.

B. Scénarios patient

En italique, la phrase de synthèse. En couleur, les segments sélectionnés pour construire la phrase de synthèse.

Je voudrais parler au médecin.

Je viendrai pour ton dessin.
Je vous demande de choisir la médecine.
Il veut parler au ministre.

Veux-tu venir te promener avec moi ?

Veut-il me promettre de rester avec moi ?
Et tu veux l'emmener avant de te préparer ?
Je venais finir ton travail.

C. Procédure d'ajout d'une nouvelle langue et de nouveaux scénarios dans Mavoa

L'objectif de ce document est mettre en place une procédure permettant d'intégrer des scénarios ou d'ajouter une nouvelle langue à l'application Mavoa, sans avoir à passer par preparationMavoa pour

créer les BDS.

Pré-requis indépendants de Mavoa :
- Avoir un baratinoo à jour (module python de baratinoo installé sur Mavoa)
- Avoir un sound2dico à jour
- Avoir un fichier de config de voix vide valide

✖ *Ajout d'une nouvelle langue :*

- Ajouter l'image (drapeau de la langue) dans le répertoire /var/www/Mavoa_Demo/WebService/assets/images.

- Dans /var/www/Mavoa_Demo/WebService/templates/ paramétrer le fichier choix.html : copier le div d'une langue déjà existante et le coller à la suite. Modifier l'id et le chemin de l'image.

- Dans /var/www/Mavoa_Demo/WebService/assets/css/ paramétrer le fichier choix.css : copier le bloc d'une langue déjà existante et le coller à la suite. Modifier simplement l'id de la langue.
  Pour configurer la largeur des boutons, ligne 11, ajouter : , #<langue> input (<langue> au format : espagnol)

- Dans /var/www/Mavoa_Demo/WebService/Scenarii/, créer le répertoire de la langue (au format : es-ES)

- Créer le répertoire cfg et y ajouter :
- un fichier de config baratinoo.templ.cfg (le copier depuis une langue déjà existante)
- un fichier xml de voix (voiceconfig.templ.xml.xml) : copier le fichier xml d'une voix existante et modifier les champs propres à la langue à ajouter d'après le fichier xml d'une voix locuteur, par exemple Marta pour l'espagnol. Il est nécessaire de modifier les champs à la main.
- le même fichier xml crypté. Dans ~/svn/tts/baratinoo/build/ utiliser la commande ./crypt.py. Cette commande va générer un fichier voiceconfig.templ.xml.lmx qu'il faut renommer en voiceconfig.xml

✖ *Ajout de nouveaux scénarios :*

Pour chaque scénario :

- les fichiers txt des phrases à enregistrer
- un fichier __init__.py permettant l'affichage des couleurs dans l'application
- un fichier text_to_record.txt contenant les phrases à enregistrer
- les fichiers wav correspondant aux modèles

- une BDS conforme :
- enr, phn, ref, txt

✖ *Création de la BDS : (BDS_REF)*

Pour créer la BDS_REF pour chaque scénario, il faut faire appel à des outils sound2dico indispensables pour n'importe quel processus de création de voix.

- Dans le répertoire de la langue /var/www/Mavoa_Demo/WebService/Scenarii/<langue>, créer le répertoire du scénario (en lui donnant le nom qui sera affiché sur le site, exemple : Almodovar).

- Créer un répertoire txt dans lequel copier les fichiers txt des phrases à enregistrer. Chaque fichier doit être nommé phraseXXXXX.txt, où XXXXX correspond au numéro de la phrase en 5 chiffres, et être numérotés de 1 à <taille_corpus>

- Créer un répertoire wav dans lequel copier les fichiers wav modèles enregistrés. Chaque fichier doit être nommé phraseXXXXX.txt, où XXXXX correspond au numéro de la phrase en 5 chiffres, et être numérotés de 1 à <taille_corpus>

- Copier le fichier __init__.py dans var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus>

- Copier le fichier text_to_record.txt dans var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus>

Calculer les enr à partir des txt :

- Dans var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus> créer un répertoire enr

- Ajouter le fichier enr.info dans le repertoire enr (le copier depuis la voix vide de la langue /mnt/nas/Synthese/BDS/vide/videes.1.0/BDS/enr/ ou depuis un scénario déjà existant dans la même langue)

- lancer la commande ~/svn/tts/ACV/sound2dico/sound2dico.py --lang <langue> txt2enr (<langue> au format : ES) (ajouter -f après txt2enr pour écraser des fichiers déjà existants)

Calculer les phn à partir des txt :

- S'assurer que la voix vers laquelle pointe baratinoo est bien configurée à showEnr=1.

- Dans var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus> lancer la commande ~/svn/tts/ACV/sound2dico/sound2dico.py --lang <langue> txt2phn (ajouter -f après txt2phn pour écraser des fichiers déjà existants)

Création du répertoire ref :

- Dans var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus> créer un répertoire ref

- Ajouter le fichier listpho.ref dans /var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus>/ref (le copier depuis la voix vide de la langue mnt/nas/Synthese/BDS/vide/videes.1.0/BDS/ref/ ou depuis un scénario déjà existant dans la même langue)

- Ajouter le fichier voix.ref dans /var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus>/ref (le copier depuis la voix vide de la langue mnt/nas/Synthese/BDS/vide/videes.1.0/BDS/ref/ ou depuis un scénario déjà existant dans la même langue)

- Dans /var/www/Mavoa_Demo/s2d/ générer le fichier list_all_id.ref en lançant la commande : python GenListId.py /var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus>/enr/ /var/www/Mavoa_Demo/WebService/Scenarii/<langue>/<corpus>/ref/list_all_id.ref

D. Résultats des tests sur les différents dictionnaires utilisés en français et en américain (SNF automobile)

Auditeur 1 :

US :

Résultats très partagés, il n'y a pas de différence réellement audible entre les deux procédés.
Les critères sur lesquels l'auditeur s'est basé pour pouvoir choisir une des deux phrases sont plutôt la prosodie, l'intonation, la vitesse, etc. Il semble qu'il n'y ait pas de différences audibles sur le choix des unités, ce qui peut nous permettre par la suite de se contenter d'un regroupement en 17 classes.

FR :

Différence flagrante dans les réponses, CamilleB (dictionnaire construit avec les numéros enregistrés en triplette et non séparément), semble bien meilleur d'un point de vue prosodique. La synthèse est donc plus naturelle et il vaut peut-être mieux enregistrer les numéros en triplette pour un meilleur rendu final.

Auditeur 2 :

US :

Les résultats montrent une préférence assez nette pour le dictionnaire A (regroupements riches) dans le cas des phrases «miles». La différence est relativement importante également sur les nombres contenus dans les phrases «numéros de téléphone», mais avec cette fois une préférence pour le dictionnaire B.

FR :

Sur le français les résultats également plutôt révélateur. Le dictionnaire B, contenant les numéros

enregistrés en triplette, est favorisé.

Auditeur 3 :

US :

La différence n'est pas évidente, sur les miles la tendance est plus sur le A tandis que sur les numéros, la tendance est sur le B. Cependant, sur les miles l'auditeur a choisi de ne pas se prononcer sur deux phrases et a seulement donné sa préférence à deux phrases A de plus que B. Même chose sur les numéros, trois réponses sont restées sans opinions, trois autres tendent vers le A et quatre vers le B.

FR :

Nette préférence pour le dictionnaire A. Aucune réponse n'est restée sans opinion.

Les commentaires de l'auditeur sur ce test montrent qu'il a préféré, en générale, le dictionnaire A, quelle que soit la langue *«il me semble que les versions 'a' sont plus "plates", plus monotones. Du coup la version 'b' apparait plus naturelle lorsqu'elle est bien rendue, par contre lorsqu'il y a des "hachures" entre les syllabes, ca choque aussi beaucoup plus. L'idéal serait un compromis entre les deux, peut-être ?»*. Cependant, on peut penser qu'il a donné sa préférence au dictionnaire A plutôt par défaut puisque les «hachures» du dictionnaire B le gênaient beaucoup plus que le manque de vivacité du dictionnaire A.

Auditeur 4 :

US :

Résultats quasiment exactement contraires entre les miles et les numéros. Pour les miles, l'auditeur préfère plutôt le dictionnaire A, tandis que sur les numéros de téléphone, il préfère le dictionnaire B.

FR :

Nette préférence pour le dictionnaire B. Aucune réponse n'est restée sans opinion.

Auditeur 5 :

US :

Les réponses de l'auditeur tendent tantôt vers le dictionnaire A (miles) tantôt vers le dictionnaire B (numéros) sans énorme différence néanmoins. Ses commentaires montrent cependant plusieurs choses intéressantes :

Les choix pour le dictionnaire A ont souvent été guidés par un manque d'intelligibilité ou un défaut (saut de picth) de la phrase B. Il semble donc que le dictionnaire B soit meilleur au niveau prosodique et naturel mais que certain défauts de pitch poussent à choisir le dictionnaire A.

FR :

L'auditeur a trouvé le dictionnaire A plus monotone que le dictionnaire B selon ses commentaires, mais a quand même donné sa préférence à plus de phrases A que de B.

Auditeur 6 :

US :

Pas de réelle différence marquée. Selon les commentaires de l'auditeur, tantôt le A est plus naturel et plus vivant tantôt c'est le B. Mais on note quand même encore une fois quelques défauts de

coupures, hachures dans le dictionnaire B.
Par contre, pour ce qui est des numéros, la différence est marquée avec une préférence pour le dictionnaire B. Selon ses commentaires, le dictionnaire B est plus naturel dans quasiment tous les cas, plus fluide.

FR :
Préférence marquée pour le dictionnaire B. Le B est encore une fois perçu plus naturel, plus vivant, mais présente quand même quelques coupures désagréables.

Auditeur 7 :
US :
Pas de grosse différence sur les miles mais une préférence plutôt marquée pour le dictionnaire B pour les numéros.

FR :
Légère préférence pour le dictionnaire B mais pas flagrante. Selon les commentaires, la perception du naturel varie en fonction des phrases et n'appartient jamais à un dictionnaire en particulier. Quelques défauts dans le dictionnaire A poussent à choisir le dictionnaire B et inversement.

Auditeur 8 :
US :
Pas de différence très marquée mais tout de même une préférence pour le dictionnaire A pour les miles et les numéro, à la différence des autres qui préfèrent tous le dictionnaire B pour les numéros.

FR :
Une différence plus marquée sur le français avec une préférence pour le dictionnaire B.

Auditeur 9 :
US :
Pas de différence flagrante entre les dictionnaires même si on remarque une préférence pour le A en ce qui concerne les miles et pour le B en ce qui concerne les numéros, à l'image des autres auditeurs. Cependant on peut noter ici que l'auditeur a laissé beaucoup de réponse sans opinion, précisant qu'il ne faisait pas vraiment de différence. Selon ses commentaires, la différence joue sur des phénomènes d'emphase, le dictionnaire A montre plus souvent des emphases trop marquées dans le cas des miles. Au contraire, pour les numéros, le dictionnaire B l'emporte sur des critères d'intonation de fin de phrase.

FR :
L'auditeur préfère le dictionnaire A pour le cas des numéros de téléphone en français. En effet, la différence est assez importante, même si on note encore un nombre relativement important de réponses sans opinion.

En définitive, il semble très difficile de se prononcer sur les différences entre les deux dictionnaires de chaque langue puisque personne ne semble complètement d'accord. Les phrases contenant des numéros de téléphone semble faire plus consensus que les autres mais sans préférence évidente chez tout le monde. Nous allons maintenant répertorier les réponse de tous les auditeurs dans un tableau

pour se rendre mieux compte de la perception des différents dictionnaires.

Tableau de résultats :

	Paul A	Paul B	sans opinion	total
Miles	**50**	30	10	90
Numéros	27	**47**	16	90

	Camille A	Camille B	sans opinion	total
Numéros	29	**69**	10	108

Table des matières

Résumé

Les technologies vocales se développant toujours plus dans notre environnement, la recherche de la qualité et du naturel sont devenus les enjeux majeurs de la synthèse vocale. Ce mémoire met donc en évidence les techniques développées par Voxygen pour garantir une synthèse de qualité, naturelle et expressive : l'utilisation d'une unité particulière ; le sandwich, et la création de voix contextuelles (Slot'N'Fill). Il appuie également sur l'application des ces technologies à des langues diverses. Ce travail fait suite à un stage de six mois effectué chez Voxygen durant lequel un processus de création de voix contextuelles a été validé sur deux langues et au cours duquel de nouvelles langues ont été intégrées à une application de synthèse vocale grand public. Les deux missions confiées ont été réalisées dans l'optique de fabriquer une synthèse de qualité, selon des procédés différents. Ce travail présente donc l'importance de l'utilisation du sandwich et l'impact du choix des corpus dans la création de voix naturelles et expressives. Le stage ayant permis de prendre part à toutes les étapes des processus de création de voix dans des langues diverses, nous avons choisi de présenter ces différentes phases sous les notions de plurilinguisme et de pludsciplinarité.

Mots-clés : synthèse vocale, sandwich vocalique, Slot'N'Fill

Abstract

With speech technology playing an increasing role in our everyday lives, the quest for quality and naturalness in speech synthesis has become a major objective. This report highlights two techniques developed by Voxygen to ensure high-quality, natural and expressive TTS: the use of a particular unit, the vocalic sandwich; and the creation of Slot'N'Fill voices. It also details how these techniques can be applied to various languages. This paper sums up a six-months internship at Voxygen, during which the Slot'N'Fill voice creation process was refined in two languages, and new languages incorporated into mainstream voice-synthesis application. Both of these tasks aimed at producing high-quality synthesis, each in their own way. Therefore, this report evaluates the relevance of the sandwich as a sound-database unit, as well as the impact of corpus choice in creating natural and expressive voices. Since the internship allowed the author to participate in all stages of the voice-creation process in several languages, notions will be expounded from a multilingual and trans-disciplinary perspective.

Keywords : Text-to-speech, vocalic sandwich, Slot'N'Fill