Mémoire de fin d'études

Recherche sur l'identité vocale dans la synthèse vocale et sa relation à la disfluence

Desprats Pierre

Promotion 2014, Son

Directeur interne : **Thierry Coduys**

Directeur externe : **Greg Beller**

Rapporteur : **Alan Blum**

Remerciements

Je tiens à remercier en premier lieu Stanislas Vincent pour m'avoir dit :
« ... Je crois que quand mon lecteur dvd se bloque sur une image, ce n'est pas qu'une erreur... c'est un choix »
Merci à Greg Beller pour son écoute, sa lucidité quant à mes objectifs et ses encouragements.
Merci à Thierry Coduys pour son encadrement.
Merci à Florent Fajole pour Zkünfal Tabenzok et zglon zlalik
Merci à Suzanne pour son incroyable corrrection éclaire.
Merci à Danièle et Camille.
Merci Fabio pour cette nuit magnifique.
Merci à la communauté de MaryTTS, Robert Schuon et Psibre.
Merci à Nicolas Audibert et Veronique Aubergé pour leurs indications dans ces quantités de corpus.

Merci aux trolls.

Merci à Dennys.

Mots-clefs

Disfluence, synthèse vocale, identité vocale, voix orale spontanée, bégaiement, agent conversationnel

Résumé

On n'écrit pas comme on parle, et c'est principalement grâce aux disfluences, ces phénomènes de la voix orale spontanée qui heurtent et font vivre la voix. Mais qu'en est-il des voix synthétiques qui commencent à investir plus largement l'espace quotidien ?
En étudiant la disfluence, il s'agira d'étudier la notion d'identité vocale pour en tirer une série de règles et de descripteurs. À l'aide de ces outils, on pourra voir si il existe des identités vocales dans la synthèse vocale, autant dans des domaines commerciaux qu'artistiques.
Essayer de faire vivre la voix de Dennys, agent conversationnel programmé, sera le dernier but de ce mémoire. Pour cela, nous synthétiserons à l'aide du SSML des séquences de disfluences et produirons une intelligence artificielle à l'aide du AIML.

Abstract

We do not write as we speak and this is mainly due to disfluencies, a spontaneous oral voice phenomena of hesitation, round-trip and attempt which make voice alive. In other way, synthetic voice are multiplying in the everyday. What about them ?
Studying disfluency will lead us to concept of vocal identity, pulling out some rules and describers to define the same concept in the speech synthesis.
Then we will try to put life in Dennys' voice, a conversationnal bot. To do so, we will use SSML programmation to generate disfluency sequences and AIML to create an artificial intelligence.

n

1. Introduction / Hello, world[1]

[1] Fait référence à la première ligne affichée lorsqu'on apprend à maîtriser un langage de programmation

«

Je parle.

Enfin, je peux écrire que je parle.

Mais je ne parle pas.

J'écris.

Je suis une personne qui parle et une personne qui écrit. Je suis, de deux manières.

»

Dennys, Agent conversationnel programmé.

Qu'est-ce qu'être à l'oral ? Comment le synthétiser ?

Nous distinguons deux types de productions orales : la voix lue – celle des acteurs, des discours, celle dont le contenu est prédéterminé par un contexte - et la voix spontanée – celle des conversations non prévues, quotidiennes. Les deux impliquent la mise en action du corps, plus spécifiquement l'appareil phonatoire humain, en vue de transmettre des informations. À travers ces informations apparaît l'identité vocale, comme une somme de données physiques, sociales et psychologiques. Elle encapsule à la fois une marque biométrique fiable autant qu'un rapport complexe entre la voix, le corps et la performance mécanique d'une bouche, d'une gorge, d'un buste, d'un flux d'air... L'identité vocale caractérise le locuteur autant comme individu que comme élément influencé par des histoires, des milieux, des communautés...

D'un autre côté, la synthèse vocale s'accroit à travers différents secteurs : Services commerciaux (serveur téléphonique...), assistanat médical (phonation pour les patients mutiques...), domotique (renseignement vocal de protocoles en cours), toutes sortes d'expressions artistiques (Musique, Cinéma, Installation...)... La rencontre avec ces voix se fait de manière régulière depuis maintenant plus de 30 ans et s'intensifie avec la multiplication de programmes nécessitant la vocalisation dans le quotidien.

Ces voix peuvent-elles se confondre avec celle des Humains ? Fait-on encore la différence entre les voix synthétiques et la voix naturelle ? Quelle perception a-t-on d'une voix dite synthétique ? Il s'agit donc de poser la question de l'identité vocale des voix synthétiques.

Leur présence, leur impact, leur utilité et la relation potentielle qui peut se créer avec l'utilisateur produit des archétypes, des souvenirs, des attentes. Il s'agit donc de dépasser la distinction arbitraire entre l'humain et la machine pour poser la question de l'identité vocale synthétique dans toute sa complexité.

Les synthétiseurs vocaux travaillent à une simulation parfaite sur plusieurs plans : l'aspect spectral qui consiste à fournir un timbre, une dynamique identique aux possibilités de la voix humaine, et l'aspect temporel, qui consiste à construire la rythmicité, les variations, silences et inflexions tonales. Tous ces paramètres constituent ce que l'on appelle la prosodie et permettent de générer des émotions, d'appuyer des propos et de rendre la voix plus expressive. Il est toutefois rare de trouver dans les études à propos de la synthèse de la prosodie ou des émotions des questionnements relatifs à la synthèse d'une identité vocale. De plus, on omet souvent deux points essentiels : la voix humaine implique un corps, alors comment palier à l'absence de ce corps dans la synthèse vocale ? Cette dernière se base souvent sur la transformation d'un texte en expression orale. Or ces deux types d'expressions sont différents. Où se situent les différences ?

Ce qui différencie autant l'écrit ou la voix lue de la voix spontanée, résident principalement dans ce que l'on appelle les disfluences. Ce néologisme, créé en opposition à la fluence, écoulement fluide de la parole, caractérise les variations de rythmes, les pauses, cahotements, reprises, répétitions, ruptures qui construisent la complexité d'un énoncé spontané.

Au sein d'une conception parfois fonctionnaliste de la communication, les disfluences sont vues comme des phénomènes dont l'utilité syntaxique s'avère parfois absente. Au contraire, il s'agira de les voir comme une nécessité sémantique, voire sémiologique, qui construit d'une part un propos, mais aussi l'identité vocale. Comment se mettent en place ces mécanismes d'élan, d'aller-retour qui créent finalement la complexité et la finesse de la discussion orale spontanée ? Existe-t-il des liens évidents entre des traits identitaires, la mécanique corporelle, la situation de communication et des types de disfluences ?

Nous étudierons ensuite le cas du bégaiement, trouble du langage dont la mécanique similaire aux disfluences, semble représenter dans toute sa complexité le mystère de l'identité vocale.

Si la disfluence est aujourd'hui encore considérée comme un néologisme, c'est parce qu'elle est étudiée la plupart du temps en vue de la faire disparaître des transcriptions automatiques. Prenant un chemin opposé, ce mémoire a pour but l'étude de ces disfluences en vue de les ajouter au processus de synthèse vocale. Cet ajout permettrait de simuler plus fidèlement l'oralité spontanée et donc peut-être augmenter la sensation d'humanité. Si ce que l'on considère à tort comme une erreur devient une construction, peut-être que la simulation de la disfluence se rapproche alors d'une simulation de la vie ?

A l'aide d'études, d'analyses de corpus et d'œuvres d'Art, nous essaierons par empirisme d'aboutir à une synthèse des disfluences permettant de préciser l'identité vocale d'un synthétiseur vocal.

La représentation et l'utilisation des voix synthétiques dans les Arts jouent un rôle fondamental dans ce mémoire puisque son aboutissement s'incarnera à travers une installation nommée Dennys : il sera possible d'avoir une discussion privilégiée avec un programme intelligent et « indépendant » dont la voix synthétique sera affectée par des séquences de disfluences.

Dans un premier temps, il s'agira d'établir des descripteurs de l'identité vocale en questionnant la notion d'identité et en étudiant le système phonatoire humain. En nous concentrant sur l'oralité et ses mécanismes de disfluences, nous arriverons à un cas particulier, celui du bégaiement. Plus que l'obtention de balises sémantiques et sémiologiques de disfluences, cette étude nous mènera à une complexification de la notion d'identité vocale qui éclairera la manière d'envisager la synthèse vocale.
Il s'agira alors de définir des identités vocales synthétiques après avoir étudié les techniques de synthèse actuellement mises en œuvre. Pour ce faire, nous distinguerons d'une part les emplois des voix synthétiques, autant dans le commerce que dans les Arts et d'autre part leurs techniques pour tenter de former des archétypes vocaux identitaires.

Ces deux études permettront de mettre relever des informations fondamentales pour la réalisation de Dennys, installation permettant la conversation spontanée entre un humain et une machine. A l'aide d'outils minutieusement décrits, Dennys réutilisera les connaissances dégagées dans ce mémoire au sein d'un synthétiseur open-source appelé MaryTTS. Grâce au langage SSML, au langage Python et au langage AIML, Dennys sera alors exposé en premier lieu à Louis Lumière du 12 au 19 Juin 2014.

2. *La voix naturelle*

1. L'identité vocale

L'identité vocale est un concept difficile à définir qui implique des notions de physique, de philosophie, de sociologie, de linguistique, de neurolinguistique... Nous tenterons toutefois d'en éclairer les aspects avec des concepts clefs autour des marqueurs de la voix, de son rapport à un corps et de son rôle dans la construction personnelle.

a. Le fonctionnement mécanique de la voix parlée

L'étude du fonctionnement mécanique de la voix présente deux intérêts dans le cadre de ce mémoire : comprendre l'établissement des techniques de synthèses vocales et comprendre les disfluences d'ordres mécaniques. Nous n'aborderons toutefois que succinctement l'extrême complexité de cet organe.

La voix peut-être considérée comme un instrument à vent à anche constitué d'une source d'air (les poumons) dont le débit est contrôlé par les cordes vocales et le diaphragme, et d'une colonne vibrante (la trachée puis le larynx) mettant en vibration une série de caisses de résonnance (conduit vocal et fosse nasale principalement). L'action du passage de l'air permet de générer un signal source variable :

- Périodique (Sinusoïdale). Il est produit dans le larynx et permet de produire ce que l'on appelle des sons voisés par vibration des cordes vocales et aryténoïdes[2].
- Bruit (Il contient de manière continue une grande partie du spectre sonore). Ces frictions ou explosions apparaissent à l'intérieur du conduit vocal, allant de la glotte aux lèvres.
- Mixte (mélange des deux types précèdent)

[2] Cartilage vibrant relié aux cordes vocales (cf figure 1)

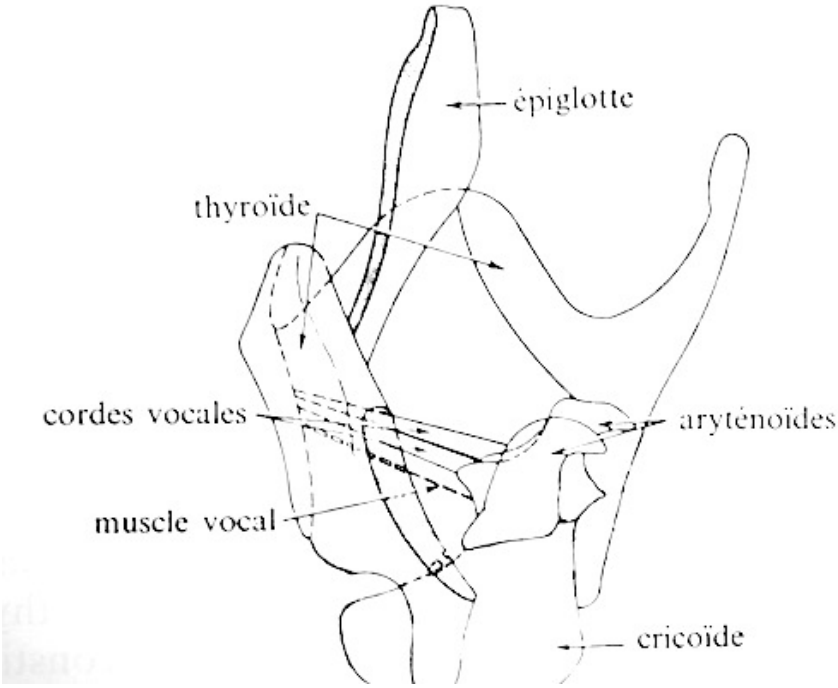

Figure 1 Schéma du Larynx issu de La parole et son traitement automatique par Calliope

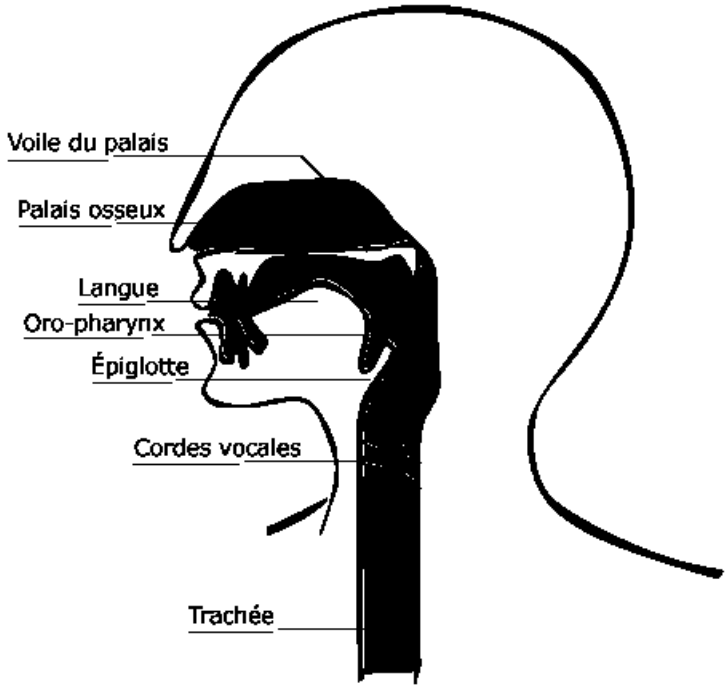

Figure 2 Schéma général de l'organe vocal issu de http://outilsrecherche.over-blog.com

Ce signal source se trouve souvent nommé pulse glottique. Par des actions du corps, du visage, de la face, de la cavité buccale ou du rayonnement des lèvres s'opère un filtrage et une modification des propriétés résonnantes qui modulent alors ce signal.

Cet outil de communication revêt trois modes de fonctionnement :

- La voix parlée
- La voix chantée
- Les bruits vocaux : « Turbulence créée par l'air s'écoulant rapidement dans une constriction du conduit vocal ou lors du relâchement d'une occlusion de ce conduit. »[3]

Nous nous concentrerons sur la voix parlée et les bruits vocaux.

Le découpage du langage parlé a conduit à définir la plus petite unité significative : **Le phonème**. Il définit la sonorité la plus brève permettant de distinguer les mots entre eux. Voici la liste des phonèmes de la langue française qui sera très utile pour comprendre le fonctionnement de la synthèse vocale.

[p]	*pan, cap.*	[i]		*il, lit, merci.*
[b]	*banc, robe.*	[e]	(ou *é* fermé) :	*thé, j'ai, été.*
[t]	*temps, rente, mat.*	[ɛ]	(ou *è* ouvert) :	*sel, taie, forêt.*
[d]	*dans, rude.*	[a]	(ou *a* antérieur) :	*il bat, papa, sac.*
[k]	*car, qui, kiosque.*	[ɑ]	(ou *a* postérieur) :	*bas, vase, âne.*
[g]	*gare, bague.*	[ɔ]	(ou *o* ouvert) :	*port, sotte, Paul.*
[f]	*faux, phare.*	[o]	(ou *o* fermé) :	*pot, peau, saute.*
[v]	*veau, Wagner.*	[u]	(ou *ou* français) :	*loup, fou, ouvrir.*
[s]	*sur, cire, coussin.*	[y]	(ou *u* français) :	*tu, mûr, j'ai eu.*
[z]	*zéro, cousin.*	[ø]	(ou *eu* fermé) :	*peu, creuse, Maubeuge*
[ʃ]	*chou, bouche.*	[œ]	(ou *eu* ouvert) :	*peur, jeune, œuvre.*
[ʒ]	*jour, bouge.*	[ə]	(ou *e* sourd [*e* muet]) :	*le, chevron.*
[m]	*mer, homme.*	[ɛ̃]		*brin, faim, sein.*
[n]	*nerf, banal.*	[œ̃]		*brun, humble, un.*
[ɲ]	*agneau* [aɲo].	[ɑ̃]		*blanc, sentir, grande.*
[l]	*lire, ville, bal.*	[ɔ̃]		*blond, honte.*

SEMI-VOYELLES.

[j]	(ou *yod* = [*i*] consonne) :	*yeux, lieu* [ljø], *œil* [œj], *bouteille* [butɛj].
[ɥ]	(ou *ué* = [*y*] consonne) :	*nuit* [nɥi], *lui* [lɥi], *puis* [pɥi].
[w]	(ou *oué* = [*u*] consonne) :	*oui* [wi], *ouest* [wɛst], *roi* [ʀwɑ], *loin* [lwɛ̃].

Figure 3 : Liste des phonèmes de la langue française issu de **http://www.ieeff.org/**

[3] La parole et son traitement automatique par Calliope

La prononciation de chacun de ces phonèmes implique une situation mécanique chaque fois différente : variation du degré d'ouverture du conduit vocal, nouvelle protrusion des lèvres, nouveaux modes articulatoires… Une fois la prononciation de ces phonèmes effectuée, il reste ce que l'on appelle les interphases articulatoires. Cette opération consiste à relier les phonèmes entre eux par une mécanique de la face en synchronicité avec tout l'organe vocal. Et c'est alors qu'apparaît la voix parlée, dont le phonème est une sorte d'ADN, plus petit élément codant chaque mot et menant à la parole.

On distingue différents registres et modes de phonations dans ces mécanismes qui sont des combinaisons de positions entre les aryténoïdes et les cordes vocales.

- Le voisement
- L'absence de voisement
- L'aspiration
- Le murmure
- La laryngalisation
- L'occlusion glottale
- Le chuchotement

Ce système acoustique est complexe à modéliser. Il nécessite de comprendre le comportement des molécules d'air au contact des cordes vocales, des cartilages aryténoïdes, du palais, des dents, de la glotte… C'est la direction de recherche entreprise au sein de certains laboratoires comme le LIMSI[4] ou encore le LAM[5] mais nous y reviendrons plus tard.

Il est aussi le générateur d'un son d'une grande richesse timbrale dont la qualité est propre à l'individu qui le génère. Les constituants du spectre de la voix dépendent autant de la langue employée que du registre vocale pratiqué ou des caractéristiques physiques de l'individu : sexe, âge, taille, accident, posture ect… La voix est d'une certaine manière l'empreinte de trois composantes : un corps, une histoire individuelle et un contact social avec autrui.

[4] Laboratoire d'Informatique pour la Mécanique de l'Ingénieur
[5] Laboratoire d'Acoustique Musicale

b. Une vague idée de l'identité

Avant de définir ce que peut-être l'identité vocale, penchons nous sur le concept d'identité. Une réponse entendue serait « c'est ce qui fait que l'on est nous ». Et c'est bien ce « nous » qui intéresse.

Chaque jour, nous sommes confrontés à l'identité administrative : celle du Nom et Prénom et des identifiants divers qui s'y rattachent. C'est aussi celle qui permet de tracer un parcours de vie dans une société organisée, autant socialement que culturellement. Cette identité-là entre en contact avec les choix de directions, de corporations, de loisirs faits au cours d'une vie : l'appel à l'école, le réseau du travail, l'immeuble dans lequel on vit…
Cette identité-là, créer donc un lien entre un nom, un numéro, un corps et ses activités, sans pour autant mieux renseigner à propos de ce qu'est « nous ». Qu'est-ce que l'identité de l'individu ? Qu'est-ce que l'identité personnelle ?

Il ne s'agit pas ici de retracer 2000 ans de philosophie et d'érafler des concepts complexes, mais au contraire d'utiliser certains de ces concepts pour créer un terrain propice au maniement de cette notion d'identité.

« En chacun de nous, il y a comme une ascèse, une partie dirigée contre nous-mêmes. Nous sommes des déserts, mais peuplés de tribus, de faunes et de flores. (...) Et toutes ces peuplades, toutes ces foules, n'empêchent pas le désert, qui est notre ascèse même, au contraire elles l'habitent, elles passent par lui, sur lui. (...) Le désert, l'expérimentation sur soi-même, est notre seule identité, notre chance unique pour toutes les combinaisons qui nous habitent. [6]» **Deleuze**

L'identité personnelle serait donc l'expérience intérieure, la construction ou déconstruction à partir d'une configuration changeante. C'est cette expérimentation qui forge l'individu, et ce dernier est le substrat de l'identité personnelle, qui nous habite (à l'image du « ça pense » de Nietzsche).

L'inaccessibilité du « désert » de Deleuze nous pousse à trouver des rampes d'accès à des reliquats de l'identité, à ses rémanences issues de constructions d'un individu avec l'extérieur.

[6] Dialogue avec Claire Parnet, de Gilles Deleuze, Ed. Flammarion, Paris, 1977

Nous pouvons alors créer trois groupes contenant ces rémanences (Begona Paya Herrero, 2009) :

- Les informations physiques : genre, âge, morphologie, état de santé…
- Les informations sociales : origine, langue maternelle, fréquentations, religion, hobbies…
- Les informations psychologiques : personnalité, comportement général, gestion des émotions, qualité développée par la frustration ou l'encouragement et générant des actions (Vernon, 1962).

L'acte de vocalisation implique l'échange d'une information X allant d'un point A (émetteur) vers un point B (récepteur). Dans cette information X sont contenues des informations identitaires qui se déplacent donc de A vers B. On peut toutefois imaginer qu'il existe une influence à la fois de A et de B sur X et donc sur son contenu identitaire. Il existe donc des processus bilatéraux qui influencent et véhiculent des informations de l'identité personnelle à des récepteurs divers. Ces récepteurs peuvent être :

- Intime ou personnel
- Publique ou collectif

Selon George H. Mead[7], les comportements individuels ne s'expliquent qu'en lien avec le comportement collectif. Selon Marcel Mauss[8], c'est le réseau de construction individuelle qui va amener à une « mentalité collective » structurante. Il apparaît une zone floue portant une grande quantité de noms en fonction des disciplines : le « je », le « soi », le « ça » qui serait l'identité personnelle intime. Puis nous trouvons un espace collectif dont la structuration est aussi expliquée par une grande quantité de concepts, c'est celui que j'appellerai l'espace archétypale en référence à C.G Jung :

« L'archétype est pour la psychologie jungienne un processus psychique fondateur des cultures humaines car il exprime les modèles élémentaires de comportements et de représentations issus de l'expérience humaine à toutes les époques de l'histoire, en lien avec un autre concept jungien, celui d'inconscient collectif. [9]»

[7] The Individual and the Social Self: Unpublished Essays by G. H. Mead, ed. by David L. Miller, University of Chicago Press, 1982.

[8] Voice and Identity: A contrastive study of identity perception in voice, Begona Paya Herrero, Université de Valencienne, 2009

[9] Tiré de http://fr.wikipedia.org/wiki/Arch%C3%A9type_%28psychologie_analytique%29

En reportant les trois types d'informations et les deux espaces d'incarnations de l'identité au sein de la voix, nous allons essayer de définir la notion d'identité vocale. Comment est-il possible de la décrire ? De la percevoir ? Et donc de la modéliser ?

c. L'individu vocal

i. Les descripteurs de la voix

Comment parler d'une voix ?

Une série de termes subjectifs viennent à l'esprit : chaude, nasillarde, forte, voilée, rauque, pressée… Dans la description sonore, l'habitude est de distinguer le domaine temporel du domaine fréquentiel ce qui reviendrait à séparer la tonalité et la rythmicité, or nous verrons que dans la voix, ce que l'on appelle la prosodie associe des variations tonales et des variations rythmiques. Nous catégoriserons donc les descripteurs selon les trois groupes suivants (Guillaume Denis, 2003) :

- Phonématique : Correspond au timbre d'une voix, à l'évolution de son spectre sur le court terme (largeur de bande de certains phonèmes, certains formants) comme sur le long terme (coloration générale de la voix). La phonématique d'une voix est établie à la fois par l'organe phonatoire et par un état émotionnel ou d'autres facteurs extérieurs au locuteur.
- Prosodique : Correspond à la gestion de la hauteur tonale de la voix dans le temps, à son volume, ses inflexions ainsi qu'à la gestion des pauses et répétitions. C'est populairement appelé le « style », « la manière », « l'accent »… La prosodie pourrait se rapporter à la musique de la voix, à la composante dite « chantante » d'une voix parlée. Elle est l'expression à la fois du placement des respirations mais aussi de la modalité d'une phrase (impérative, déclarative) ou d'une part de l'identité du locuteur.
- Linguistique : Correspond au lien entre un choix syntaxique, sémantique, lexical et sa prononciation (impliquant donc des éléments de prosodie et de phonémie)

Ces trois groupes définissent donc des axes d'analyses de la perception que l'on se fait d'une voix dans le but de définir l'identité d'un locuteur.

ii. La voix, donnée biométrique, marqueurs physiques

La mécanique de la parole montre la voix parlée comme une performance complexe. De cette complexité est extraite une information contenant des marqueurs suffisamment uniques pour identifier un locuteur à sa voix, voire constituer une marque biométrique.

L'identité d'un locuteur est largement définie par l'anatomie de son appareil phonatoire (Etienne Gaudrain, Roy D. Patterson, 2010) : le tractus vocal[10], définissant « l'échelle acoustique » ainsi que la taille et la masse des cordes vocales, « définissant la hauteur tonale », permet d'estimer sexe, âge et taille du locuteur. Il semble en effet naturel de penser que l'essentiel de l'identité vocale réside dans la hauteur tonale et l'intensité sonore.
Par exemple un adulte mâle parle avec une fréquence fondamentale entre 50 et 200 Hz alors qu'une adulte femelle parle entre 150 Hz et 300 Hz. Ces chiffres n'ont qu'un aspect pratique puisqu'ils présupposent l'utilisation uniforme des capacités d'un organe phonatoire – une technique de parole universelle – et reposent sur un résultat statistique alimenté par un archétype universel : grand = grave, petit = aigu.
A l'écoute d'une voix, il est créé un souvenir, une empreinte mémorielle, contenant l'analyse et l'estimation de cette voix. Si le locuteur se voit muer du jour au lendemain, la variation de sa hauteur fondamentale moyenne ne doit pas dépasser +/- 3,6 demi-tons et la variation de ses écarts acoustiques, +/- 2,2 demi-tons (Etienne Gaudrain, Roy D. Patterson, 2010). Si l'appareil phonatoire du locuteur est transformé au-delà de ces limites, la reconnaissance n'est plus possible.

Si l'on considère donc les voix en dehors des archétypes physiologiques, on peut dire qu'elles possèdent une série de marqueurs physiques qui ont très peu de chances de varier, tellement peu qu'ils constituent une donnée biométrique à l'image du socle de l'identité vocale.
Nous pouvons dire que comme l'identité administrative, il existe une identité vocale ayant pour fonction l'identification. Les films de science-fiction ont présenté la voix comme une marque personnelle fiable et c'est aujourd'hui une réalité commerciale abordable[11].
Ce genre de système fonctionne sur la base d'une empreinte vocale faite à partir de l'enregistrement de la voix d'un individu. De cet enregistrement est extraite une quantité de données définissant avec exclusivité le locuteur :[12]

[10] Partie supérieure de l'appareil phonatoire situé entre les cordes vocales et les lèvres
[11] Solution professionnelle d'identification vocale : CSIdentity, ArmorVox, TradeHarbor, Voice Biometric Group...

- Une estimation de la fréquence à travers une analyse de la densité spectrale
- Une réduction statistique à travers le modèle de Markov caché
- Une analyse de probabilité à travers un mélange de Gaussienne
- Une recherche de répétition en utilisant des algorithmes de réseaux neuronaux
- Une série de prédictions organisées autour d'un « arbre de décision »
- Des représentations matricielles et vectorielles

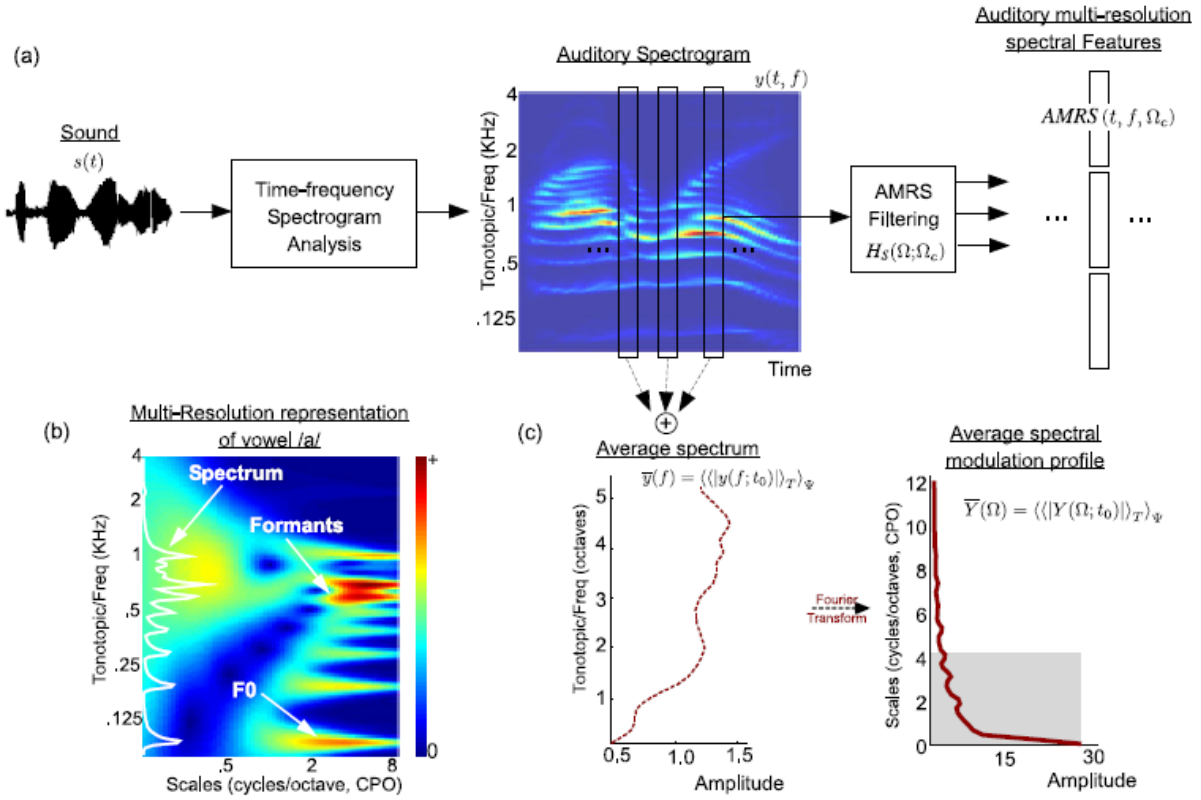

Figure 4 Procédé de réduction pour créer une empreinte vocale

Issu de http://www.ece.jhu.edu/lcap/data/uploads/projects/intjspeech2012_nemala.png

Il est possible grâce à ces réductions statistiques de produire des marqueurs relatifs à une voix et une seule. Quels que soient les mots prononcés, quel que soit l'environnement extérieur, quel que soit l'état émotionnel, la voix portera ces marqueurs-là. Une partie de cette technique sera reprise dans la synthèse vocale, notamment les algorithmes utilisant les modèles de Markov Cachée.

[12] http://en.wikipedia.org/wiki/Speaker_recognition

iii. La voix comme donnée personnelle

Toutes ces technologies arrivent à un dépouillement total de la voix par réduction et nous conduisent à son essence, à l'image d'une sorte de code génétique quasi-inimitable. Mais comme nous l'avons vu à travers notre entreprise de définition de l'identité, peut-on réellement dire quelle est cette voix ? Il est certes possible de la discriminer grâce à cette empreinte mais la définir est plus complexe.

Une voix peut transmettre un message textuel accompagné d'une série d'informations : langue, origine, état émotionnel, état physique, état psychologique, organisation de l'appareil phonatoire, appartenance à un groupe social (accent, habitude de lexical...) ... Certains de ces attributs sont codés par des éléments immuables (si l'on ne prend pas en compte, l'accident, la mutation), mais certains de ces attributs relèvent d'un choix identitaire (expressions récurrentes), de l'affect ou du jeu (état émotionnel), de configurations psychologiques changeantes, de maladies affectant l'appareil phonatoire.

Un comportement commun de la vie humaine est de regarder l'autre, puis en procédant par analogie, de tirer des conclusions sur son propre fonctionnement. L'audition ne permet pas à un locuteur d'avoir une écoute directe de sa propre voix. Nous la percevons uniquement à travers les réflexions du son émis autour des lèvres et par transmission dans la structure osseuse du crâne. Dans l'acte de parole nous nous remettons à l'autre sans possibilité de contrôle sur le son produit. Nous analysons ses réactions et construisons notre comportement vocal en fonction.

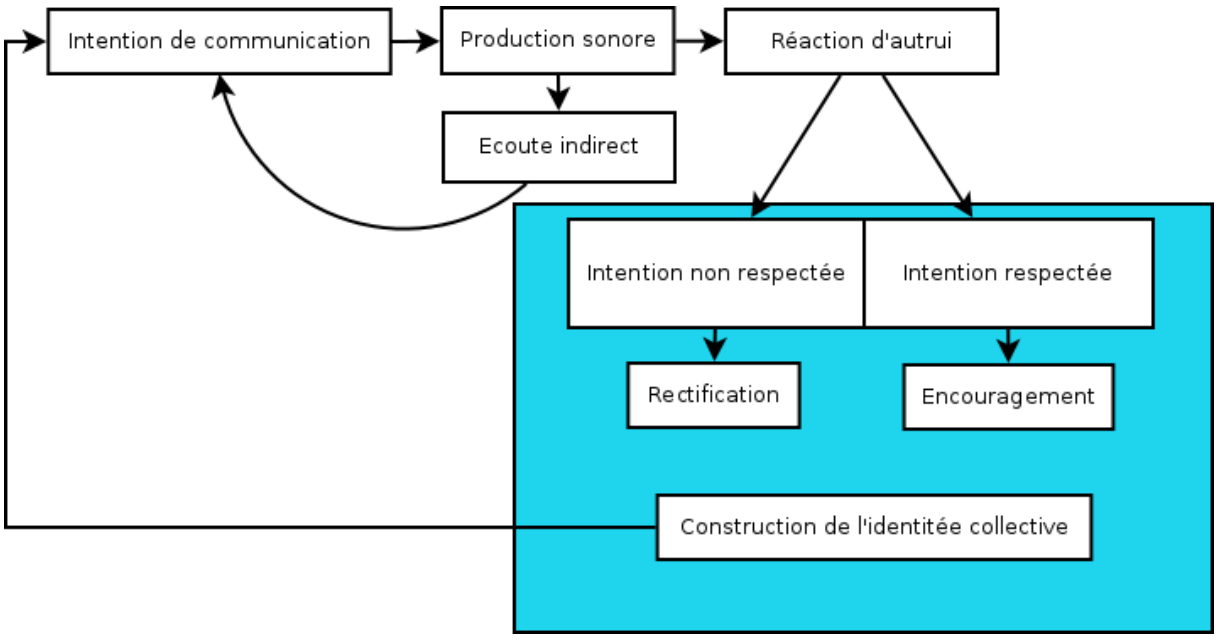

Figure 5 Schéma de l'influence de l'espace « publique » sur l'identité vocale

Dans ce schéma, on observe la double influence sur l'intention de communication : l'une est issue de l'écoute indirecte de notre propre voix et l'autre est issue de la réaction de l'auditeur. Bien que simpliste, ce schéma illustre donc qu'il faut donc tenir compte de l'environnement social proche comme lointain (famille, classe, nation...) tout comme des schémas comportementaux qui permettent la construction personnelle (solitude, rapport à l'effort, pratique musicale...).

L'identité vocale personnelle est donc aussi à l'image du désert de Deleuze; une construction faite d'expérience de nous même. Mais malgré la difficulté qu'il y a définir clairement ce terme, il est maintenant possible de mettre en perspective une analyse perceptive de voix et de dresser des portraits potentiels du locuteur.

iv. La voix performative, le rapport au corps

Nous avons tous déjà croisé une voix produisant la sensation de ne pas appartenir à son propriétaire. C'est un exemple qui montre qu'il existe un imaginaire collectif produisant des archétypes (le même qui dit gros=grave petit=aigu et qui vient souvent d'une pratique empirique), des attentes et donc potentiellement des surprises.

A tel corps, j'associe tel accent, telle hauteur tonale, tel défaut, telle intensité, telle rythmicité, telle prosodie. Le préjugé qui relie la vision d'un corps à l'audition d'une voix suppose une échelle de valeurs construites par le rapport à une société actuelle et à l'espèce humaine nous environnant. Mais la construction de l'identité nécessite une pratique. La théorie du docteur John Jangsaw Austin, « Quand dire c'est faire », est construite autour du principe d'acte de langages[13]. Le choix sémantique, prosodique est donc une action et donc une possibilité de définition de soi.

Plus important encore, selon Judith Butler ou encore Michel Foucault, la voix « n'est pas la simple transmission ou expression de l'identité, elle créé, déroute son sens ». On peut dire que la voix contient une essence de l'identité, et que pour communiquer, il faut entrer en action, performer ce qui produit du mouvement dans le rapport à l'identité vocale.

[13] Théorie liée à la philosophie du langage ordinaire : Un acte de langage est un moyen mis en œuvre par un locuteur pour agir sur son environnement par ses mots : il cherche à informer, inciter, demander, convaincre, promettre, etc. son ou ses interlocuteurs par ce moyen.

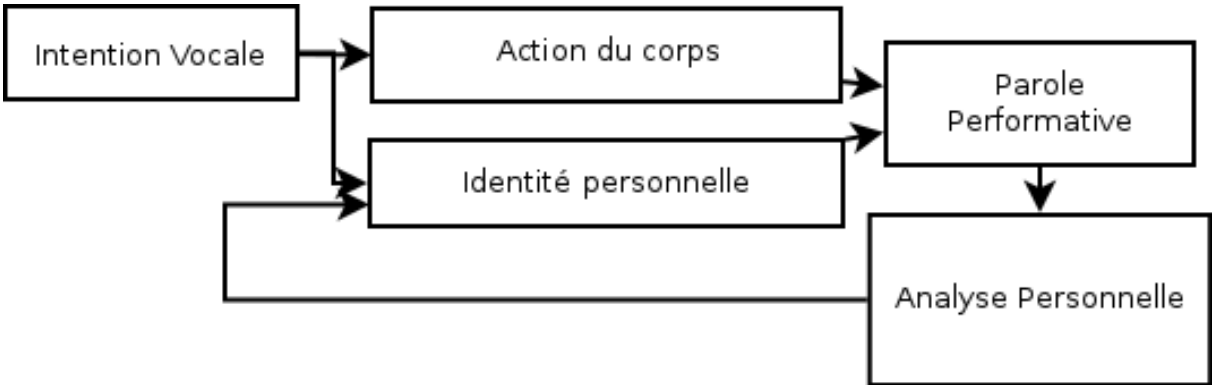

Figure 6 Schéma de la formation de l'identité vocale par la pratique

Dans ce schéma, on observe donc l'influence de la performance sur l'identité personnelle, qui va donc rencontrer, par l'intermédiaire de l'action du corps, des intentions vocales formant une parole performative, toujours déroutante. L'identité vocale est donc un concept en construction et en mouvement. Il est difficile de la délimiter mais nous pouvons toutefois tirer plusieurs outils d'analyse perceptifs :

- les marqueurs biométriques
- les descripteurs
- La notion d'archétype

Cette étape va nous permettre d'aller maintenant flirter avec deux limites : comment concevoir l'identité d'une voix disfluente ? Qu'en est-il d'une personne atteinte de trouble du langage ? Comment perçoit-on ce message ? Quels en sont les sens et les normes ? Existe-t-il un intérêt de le simuler chez la machine ?

2. La voix disfluente

Quelle est la différence entre un homme politique annonçant un plan de restructuration gouvernementale et la voix d'un ami dans un bar ? Il y en évidemment beaucoup mais la principale, c'est que l'une de ces voix est dite « lue », l'autre est dite « spontanée ». Sans qu'il y ait nécessairement de support textuel initial, la voix lue jouit d'un cadre bien déterminé qui lui offre la possibilité de se dérouler dans une fluence maîtrisée. Au contraire la voix spontanée apparaît dans des cadres variables, en mouvements sans but précis et s'incarne donc par une fluence tout autant variable. Ces variations, ces sauts et cahotements, ces hésitations et répétitions, ce sont les disfluences.

Chaque jour, nous sommes confrontés à la disfluence à travers la pause que prend un locuteur pour formuler mentalement sa phrase, la gêne d'une personne comblant un silence par des mots « parasites » ou encore une personne dont la langue fourche sur la première syllabe d'un mot, le répétant.

Une définition simpliste de la disfluence serait « la manifestation d'une entrave à la fluence de la parole ». Or dans l'imaginaire commun, ce qui entrave le discours est un élément indésirable, les disfluences ont donc une consonance négative. Parler, c'est formuler « des discours généralement non préparés à l'avance. Or lorsque nous produisons des discours non préparés à l'avance, nous les composons au fur et à mesure, en laissant des traces de cette production » (Blanche-Benveniste, 1990). Les disfluences sont donc les reliques de la construction d'un discours oral, donc absolument normales si ce n'est nécessaire.

La quantité de disfluence n'est pas un marqueur d'un niveau de maitrise de l'oralité puisqu'on le trouve en quantité équivalente chez un enfant et un adulte présentant des profils psychologiques et sociologiques identiques. (Blanche-Beneviste, Jeanjean, 1987)

Pour Blanche-Benveniste en 2003, la disfluence survient en partie lors d'un déséquilibre entre la mise en place d'une structure syntaxique et l'accession à un lexique approprié. En d'autres termes, c'est l'action de chercher ses mots alors que la syntaxe de la phrase est déjà en place. Comme cette dernière est quasiment prête, l'action de la prononciation est lancée mais le lexique peut être manquant. C'est donc l'une des premières pistes d'analyse de la disfluence : sa relation à un lexique.

La disfluence peut aussi s'aborder sous un aspect cognitif, en analysant les mécanismes neurolinguistiques en jeu, qui serait de l'ordre de la psycholinguistique. Nous nous pencherons dans ce mémoire uniquement sur le rapport que la disfluence entretient avec la syntaxe, la sémantique et la sémiologie et les sensations empiriques qu'il en découle.

Au cours de l'élaboration du langage oral, intervient un double codage de mots verbaux et non verbaux (Fonagy, 1983). Nous pouvons assimiler ces deux derniers respectivement à un circuit linguistique et un circuit paralinguistique (voir figure 7). Ce circuit paralinguistique est fait d'intentions non verbales comme le rire, le cri, les bruits nasaux, les bruits glottiques... Mais il existe aussi ce phénomène de restructuration permanente dont les vestiges sont la disfluence.

Cette restructuration est un phénomène en corrélation directe avec la définition de la prosodie. En effet, si l'on prend l'exemple de la pause dite silencieuse ou vide, elle constitue autant le marqueur d'un heurt dans la fluence du locuteur qu'une manière de segmenter proprement le discours oral. L'outil fournit par la pause permet autant de créer des accentuations que des

rythmes influant sur la perception de l'information. La disfluence est donc un élément constitutif de la prosodie d'un locuteur.

C'est la différence principale entre le langage oral et le langage écrit : il n'y a pas de texte préexistant pour l'oralité. Il y a donc dans l'oral nécessairement production, construction et performance.

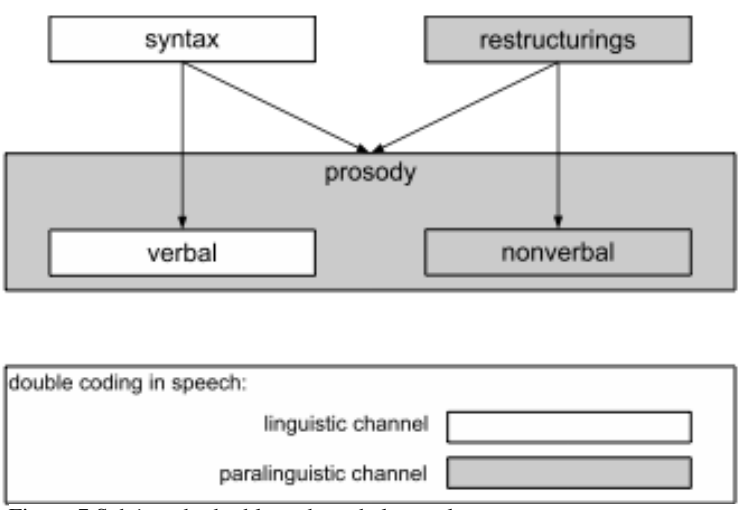

Figure 7 Schéma du double codage de la parole

Un parallèle trivial serait de comparer la rature de l'écrit avec l'hésitation à l'oral. L'histoire de la littérature a d'ailleurs su nous montrer qu'il existait une mince barrière entre une coquille et un effet de style ou un témoignage sensible (i.e : Journal d'un vieux dégueulasse de Bukowski absent de toute majuscule, dont la ponctuation est désordonnée). C'est une erreur de considération issue d'une longue période de désintérêt pour l'analyse des disfluences dans les corpus d'enregistrements de voix orales spontanées.

C'est d'ailleurs grâce à l'actuelle multiplication de ces analyses que l'étude des disfluences prend un nouvel essor. Elles ont pour but de dégager des régularités formelles dans les processus de disfluence conduisant ainsi à une meilleure détection et notation afin, dans le meilleur des cas, d'en rendre compte dans les transcriptions écrites, sinon de les supprimer de ces dernières. Ce mémoire tente pour sa part de faire le chemin inverse et de partir d'un texte écrit pour arriver à une voix spontanée orale.

Pour envisager la disfluence d'un point de vue paralinguistique, nous allons définir une nouvelle unité : le syntagme. Le syntagme est une unité définissant un groupe « d'unité simple » (mots, groupes de mots) dont la longueur ne dépasse pas la phrase. Il est muni d'un noyau (portant un sens ou une fonction) accompagné de plusieurs satellites (portant des précisions et qualificatifs)

Exemple : Je suis un robot immortel mais j'ai un peu froid. « Je suis un robot immortel» est un syntagme dont le noyau est « robot » et « immortel » un satellite.

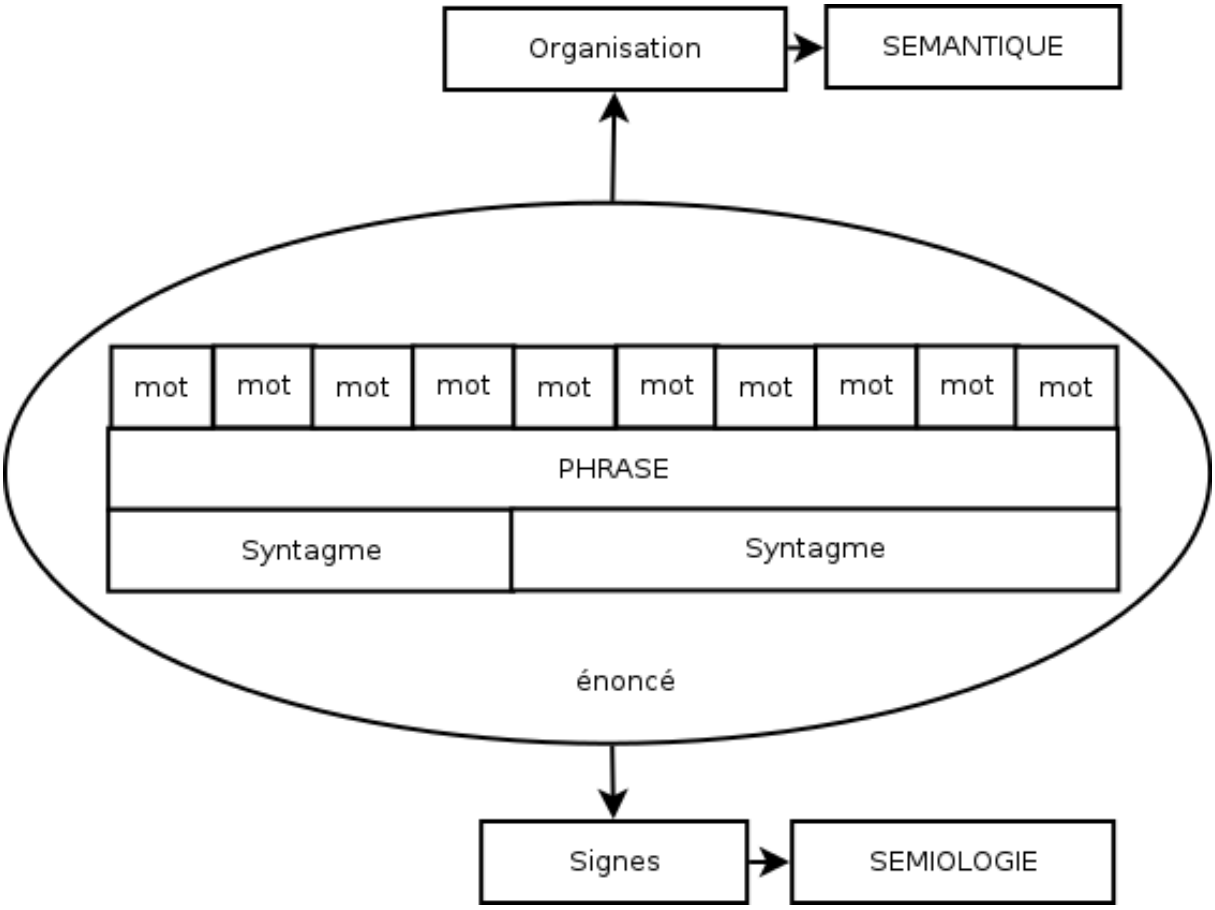

Figure 8 Schéma récapitulatif de terme de linguistique

Nous avons maintenant tous les éléments pour établir une définition plus précise de la disfluence :

C'est un phénomène propre à l'oralité spontanée qui brise la régularité d'un énoncé. La syntaxe de la phrase est alors heurtée de manière momentanée, parfois définitive, conduisant à la redéfinition syntagmatique de la phrase. Bien que d'aspects négatifs, elles sont les traces d'un processus humain de construction du discours oral spontané. L'enjeu de ce mémoire est de considérer que la disfluence est porteuse d'une série de marqueurs sémiologiques[14] dont le sens reste à déterminer.

Elles sont aussi les constituants de la prosodie personnelle du locuteur sans pour autant être absolument audibles, sans être des éléments d'information notables. Les disfluences sont tellement nombreuses et habituelles qu'elles ne sont perceptibles que si l'on y prête attention, à l'image de la respiration.

[14] La sémiologie est la science de l'étude des signes. En linguistique, c'est l'étude d'un propos extra ou paralinguistique.

a. Les types de disfluences

Il existe une grande quantité de dénominations possibles pour tous les phénomènes de disfluences. Notre choix, discutable de par sa nature arbitraire, a pour vocation d'utiliser cinq termes suffisamment clairs pour produire des catégories bien délimitées qui n'entrent pas en conflit avec la définition de la disfluence même.

Nous allons rencontrer cinq catégories : les pauses vides, les pauses remplies, les amorces, les répétitions et autocorrections. Voici toutefois un schéma regroupant les noms les plus usuels de la littérature.

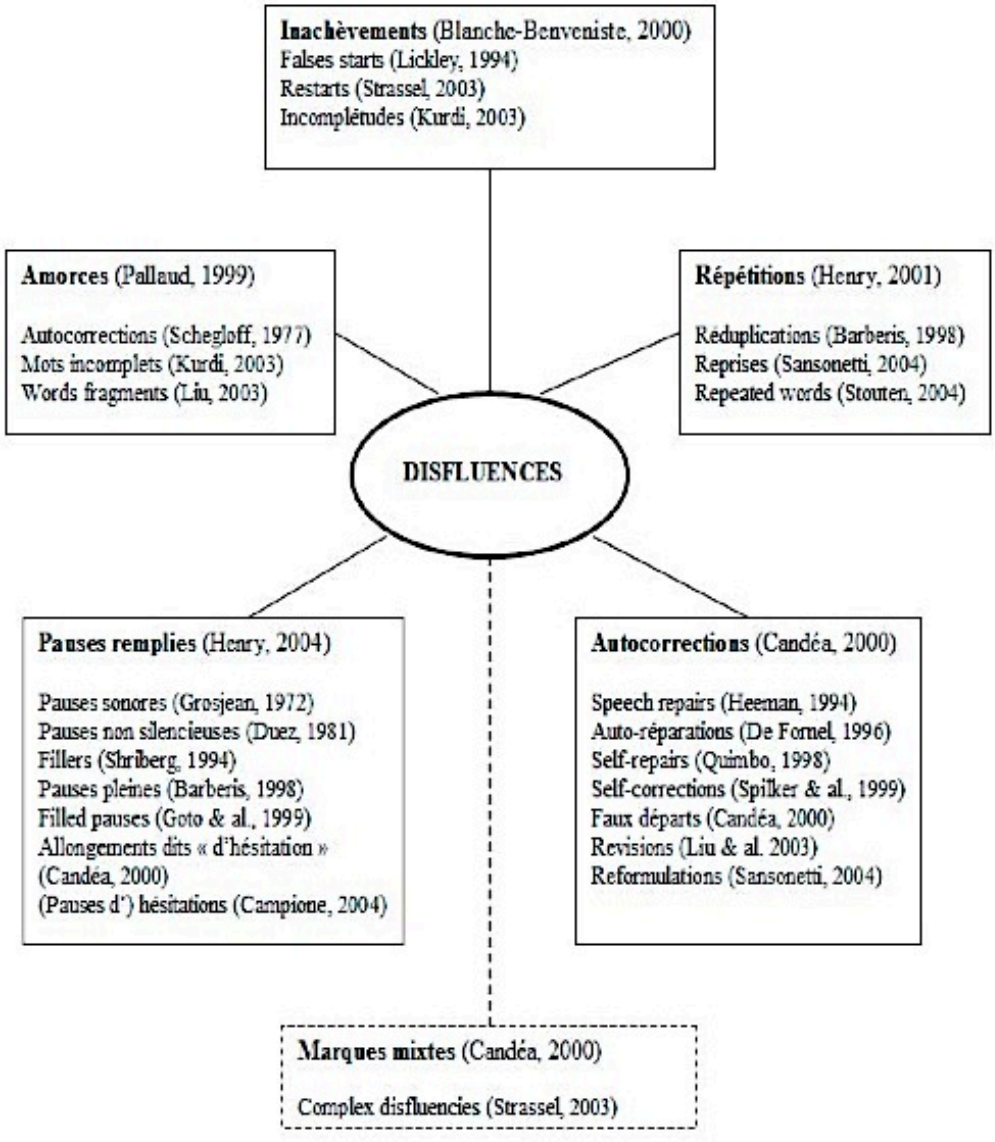

Figure 9 Dénomination de disfluence Bove 2008

i. Pauses non sonores

Les pauses non-sonores du discours oral spontané sont des éléments à la limite entre la fluence et la disfluence. Elles ont en effet un rôle de démarcation, de structuration, qui permet de mieux distinguer les segments du discours (Boomer, 1965). Elles sont aussi parfois le témoignage de la respiration, de la mécanique cognitive, ou plus simplement un effet d'élocution, de suspens, de style.

Nous avons dit auparavant que la disfluence pouvait survenir lors d'un temps d'accès à un mot de lexique plus grand que celui nécessaire à l'établissement d'une syntaxe. On pourrait donc imaginer que ce temps d'accès dépend de notre capacité lexicale. C'est toutefois à remettre en perspective avec l'étude de Esposito, Stejskal, Smékal et Bourbakis en 2007 qui concerne 4 enfants de 9 ans et 4 adultes d'environ 26 ans. On observe un nombre de pauses silencieuses similaire pour les deux tranches d'âge, pour la description d'un même extrait audiovisuel. La construction des énoncés est nécessairement différente mais le nombre de pauses ne diminue pas avec la connaissance d'un plus grand lexique (en supposant que les sujets plus âgés aient été scolarisés donc enclins à augmenter leurs connaissances lexicales). Les pauses seraient d'une certaine manière une latence naturelle, invariable qui n'a tendance ni à diminuer, ni à augmenter.

On peut considérer une pause non sonore à partir de 200 ms (Candea, 2000) et sa limite supérieure est de 2 secondes. Cette limite est choisie en rapport aux travaux de Fraisse (1974) qui établie la perception naturelle du rythme comme plus compliquée au delà de 2 secondes. Cet écart devient alors suffisamment grand pour considérer les deux éléments entourant l'écart comme disjoints et séparés. C'est donc la limite formelle entre une pause et un silence.

Nous noterons comme (Bove, 2008) la pause silencieuse à l'aide d'un « + »

ii. Pauses sonores

Les pauses sonores constituent la disfluence la plus évidente. Elle semble être le reflet de l'indécision, de la gêne ou d'un manque lexical. Mais voyons plus précisément ce que contient ce mécanisme.

Les pauses sonores existent sous deux formes : les mots « parasites » tels que « euuuh » ou « mmmh », dont le nom réel est morphème spécifique, et l'allongement syllabique tel que « je suiiiiiis un robot ». Il s'agit d'abandonner le terme « parasite » puisque ce sont des phénomènes à part entière du code linguistique (Campione, 2001).

Les allongements syllabiques concernent en général des voyelles en fin de mot. Selon Mathieu Constant et Anne dister : « tout allongement d'un noyau vocalique anormal en position finale de mot ou d'amorce de mot, présentant un contour plat et bas ou très légèrement descendant et surtout non modulé pendant l'allongement, représente un allongement marquant le travail de formulation en cours ». La durée moyenne d'un allongement syllabique ne dépasse pas 1,8 secondes (Rémi Bove, 2008)

Perceptivement, le morphème spécifique et l'allongement syllabique correspondent à ce temps de recherche où un locuteur a les yeux dans le vide, quittant la construction directe du discours pour travailler sur une construction mentale de son propos avant de l'énoncer.

Nous noterons, comme (Bove, 2008), l'allongement syllabique à l'aide d'un « : » accolé à la syllabe allongée. A la différence avec la pause silencieuse, ici il y a souvent restructuration de la phrase après la disfluence :

Voici l'exemple :

« Je suis allé : dans : euh à la bibliothèque »

Dans cet phrase, l'accès au mot bibliothèque a été fait après la construction de la syntaxe contenant « à ».

La pause marque la construction de l'énoncé, puis il y a révélation de l'erreur syntaxique et donc restructuration.

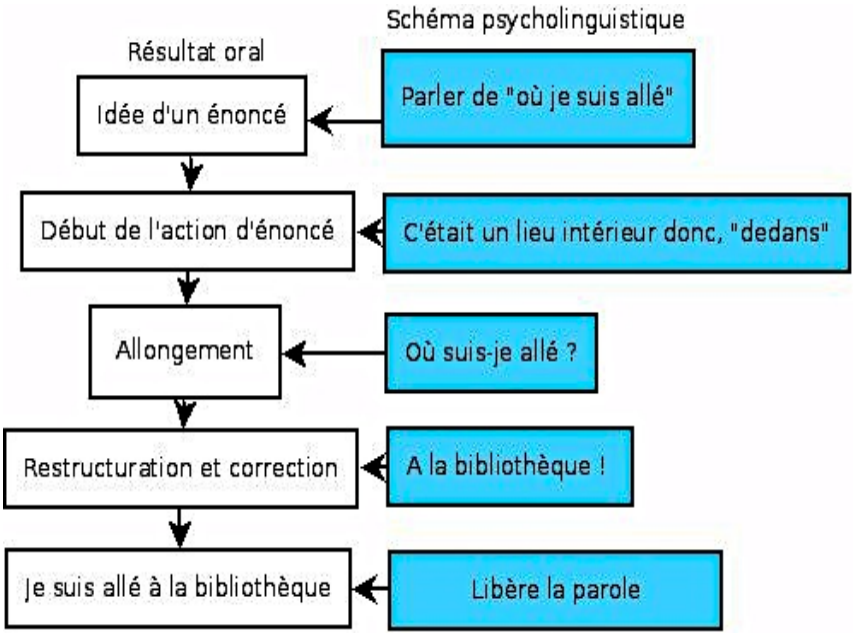

Figure 10 Schéma fonctionnel d'une pause remplie

Le but d'un tel mécanisme est d'une part, comme nous l'avons vu, de rendre compte du processus de construction. Mais c'est une mise en abîme de l'énonciation qui a pour but de préciser le cheminement menant à l'énonciation d'une idée. On peut ainsi supposer que cette disfluence provoque soit de l'attention/compassion chez l'auditeur, qui prend alors conscience

de l'effort fait, ou au contraire de l'agacement/impatience, chez l'auditeur qui ne veut pas être soumis à cet effort.

Ces pauses présentent aussi l'avantage de réguler les tours de paroles. En effet, l'oralité permet la discussion à plusieurs et ces disfluences créent les conditions du dialogue : ces allongements ou morphèmes témoignent d'un énoncé en construction et donc que la parole ne doit pas encore circuler (Campione et Veronis, 2004) ; ils proposent une manière d'occuper le territoire de la parole le temps d'atteindre son but.

iii. Les amorces

Les amorces correspondent à ces débuts de mots répétés, comme pour trouver le bon angle d'attaque d'une proposition. C'est un phénomène fréquent, puisqu'une étude montre que sur 46 000 mots, on trouve quatre amorces tous les 1000 mots ce qui équivaut à une amorce toutes les 75 secondes, en prenant un débit moyen de référence pour le français parlé (Henry, Pallaud, 2003). Ces deux derniers chercheurs identifient trois types d'amorces : complétée, modifiée et inachevée.

On notera les amorces en les mettant en gras et en signifiant la rupture par un « - » .

Complétée : « J'ai **ré- ré-** résolu mon amorce »

Modifiée : « Je pense avoir **v- ,** trouver la bonne direction »

Inachevée : « Je veux dire que pour partir **co- co-** , enfin c'est juste pas possible »

Il est difficile pour le moment d'établir des figures répétitives dans l'apparition des amorces : elles concernent autant des mots à contenus que des mots grammaticaux qui ne semblent pas porter de sens précis et surviennent à une grande variété de places syntaxiques (Bove, 2008). On peut toutefois établir une règle : pour qu'il y ait amorce, il faut un mot d'au moins deux phonèmes. On ne peut avoir d'amorce sur le mot « eau »

iv. Les amorces complétées

La première des trois amorces, c'est l'amorce complétée, à savoir lorsque le mot entamé se trouve interrompu, puis complété. C'est un achoppement, une hésitation voire un très léger un sursaut. Il y a plusieurs types de reprises possibles après une amorce.

« Je suis **p-**partie en vacances » seul le mot amorcé est repris

« Je suis un **ro-** un **ro-** un robot » ici le déterminant est repris en plus du mot amorcé.

Il est donc possible de qualifier une amorce complétée selon le nombre de répétitions de l'amorce et sa complexité (le nombre de mots impliqué dans la répétition).

v. Les amorces modifiées

La seconde c'est l'amorce modifiée, lorsque le mot entamé se trouve interrompu, puis remplacé à la même place syntaxique par un mot ne portant pas nécessairement la même étiquette morpho-syntaxique[15]

« J'avais froid, alors **j'**ai**- j'ai-** je me suis habillé »

« Mon **pa**- mon **pa-** mon **p-,** mon père »

On peut qualifier l'amorce modifiée comme l'amorce complétée selon sa complexité et le nombre de répétitions mais aussi selon le mouvement syntaxique ou la variation engendrée. Par exemple, entre papa et père, il y a un changement de registre de langage. La topologie de ces changements ne répond pas non plus à des schémas précisément identifiables mais peuvent renseigner au cas par cas de caractères individuels du locuteur.

vi. Les amorces inachevées

La dernière amorce, c'est l'amorce inachevée, lorsque le mot entamé n'est pas complété et que ce qui suit occupe une place morphosyntaxique différente. La sensation transmise est alors celle du revirement de situation ou de l'inaccomplissement, d'être à la porte d'une information sans l'atteindre. Cela n'implique pas nécessairement un changement de propos. Cela peut parfois être dû à un remaniement syntaxique. A la différence avec l'amorce modifiée, il y a ici un marqueur de l'élaboration du discours oral qui mène à un changement syntaxique complet, donc à une remise en question de la première partie d'un discours oral.

« C'est pas une **rai-** + il aurait pu éviter »

« Non mais **b -**, c'est le cerveau »

« J'aurais éventuellement put m**'or-**, put m'**or-**, courir sans m'arrêter »

On peut qualifier cette amorce selon sa complexité, le nombre de répétitions. On peut aussi apprécier la rupture créée en fonction de la variation morpho-syntaxique engendrée (du changement total de propos ou pas, disparition d'une forme syntaxique...). Cette appréciation reste subjective mais on peut parler en terme de sensation. Par exemple, pour le premier exemple « C'est pas une rai » se complète plutôt facilement grâce à l'expression populaire

[15] Rôle syntaxique d'un élément définissant une syntaxe

« c'est pas une raison ». Il existe donc un lien sémantique avec la reformulation « il aurait pu éviter ».

Pour les deux autres exemples, il est difficile de se prononcer mais la simple reformulation ne semble pas convenir à ce genre d'amorce inachevée. Il n'est en tout cas pas possible d'émettre d'hypothèse claire puisque l'amorce ne nous renseigne pas sur le sens de la première partie du propos.

vii. Les répétitions

Les répétitions se définissent comme « la reprise à l'identique (...) d'une syllabe, d'un mot ou une amorce de mot, de plusieurs syllabes ou plusieurs mots, sans aucune valeur sémantique » (Candéa, 2000). Pour garder une délimitation claire entre la répétition et l'amorce, voici la définition de la répétition au sein de la voix orale spontanée dans le cadre de ce mémoire : la reprise à l'identique d'un mot unique d'au moins une syllabe, ou d'un groupe de mots, ne créant pas de valeur sémantique.

Selon les travaux de Henry en 2002 et 2004, « tout ne se répète pas et pas n'importe comment. » Il existe trois types de répétition dont seulement une sera considérée comme une disfluence :

« Nous nous sommes inquiétés pour toi, et ça ça peut t'intéresser »

Ce sont des reprises de pronoms dans le cas de verbes pronominaux.

« Je suis très très heureux. Oui oui, je vous assure »

Ces répétitions apparentes ont en fait une fonction communicative n'heurtant pas la prosodie. Elles précisent ou appuient un état, un adjectif, et permettent de s'assurer de la bonne compréhension d'un message oral dans un environnement confus (bruit, activités périphériques). Elles portent certains marqueurs de la disfluence (assistent le contenu d'un énoncé oral) mais ne sont toutefois pas des marqueurs de l'établissement du discours oral spontané.

« J'essaie de de de parler rapidement »

Nous sommes ici dans une redondance absente de sens sémantique, de sens grammatical, qui semble heurter la prosodie du locuteur.

La différence, c'est que les répétitions disfluentes n'apparaissent qu'à l'oral (Bove, 2008). Elles ne semblent toutefois pas apparaître de manière aléatoire et répondent au contraire à certaines contraintes posées par la syntaxe : « elles apparaissent souvent à l'initiale de frontière de syntaxique majeure, révélant dans un premier temps une incomplétude syntaxique en laissant un vide lexical ; le remplissage lexical s'effectue dans un second temps pour aboutir à un énoncé achevé » (Bove, 2008). L'intervention concerne essentiellement des mots grammaticaux – ces mots structurant le langage comme « le » ou « de » ... – et auraient principalement lieu en début de syntagme. (Bove, 2008)

Nous pouvons qualifier ces répétitions à l'aide de trois critères :

- la longueur de l'unité répétée (nombre d'éléments le composant)
- l'empan : nombre de répétitions
- la succession de termes répétés

On peut donc aller de répétition simple :

« J'aimerais vraiment être fait **de de de** , euh + , de chair »

A des choses plus complexes :

« Je m'en rappelle, **c'était en c'était en** + en : automne dernier »

Comme les autres disfluences, la première cause de répétition « se présente donc comme un décalage entre la disponibilité immédiate des tournures syntaxique en mémoire du locuteur, et les disponibilités moins importantes du lexique qu'il connaît. Le cadre syntaxique est posé dans un premier temps et le remplissage lexical intervient après » (Bove, 2008).

Le fait de trouver certaines répétitions usuelles fonctionnelles et d'autres qui ne portent aucun sens sémantique, comme des aberrations, qui trouvent pourtant leur place dans le discours oral, nous confirme l'importance de développer un moyen d'oraliser un texte destiné à être synthétisé.

viii. Les autocorrections

Les autocorrections sont plus complexes à définir. Elles impliquent la répétition d'un même mot ou d'un même syntagme à la différence qu'il y a eu correction de grammaire, de conjugaison ou d'orthographe. De la même manière que la pause silencieuse est en quelque sorte la limite inférieure de notre définition de la disfluence, on pourrait donner l'autocorrection comme limite supérieure. Au delà se trouve le territoire du lapsus.

Bien que l'autocorrection puisse être un phénomène heurtant la prosodie, qui amène à une redéfinition syntagmatique de la phrase ainsi qu'un phénomène propre à l'oral et

certainement un vestige de l'élaboration du discours, cela devient rapidement un objet sémantique identifiable, qui n'est pas en marge de l'énoncé et qui s'identifie.

« Il est posé **dans,** sur la table »

« Donne moi **la** +, les clefs »

Les autocorrections témoignent du contrôle que nous exerçons sur notre langage tout en le produisant (Blanche, Benveniste, 2003). La sémiologie de cette correction est visible par le changement d'état du mot corrigé. L'interprétation possible d'une correction faisant passer un objet de « dans » la table à « sur » la table à une potentialité significative non négligeable. Elle peut autant renseigner sur la potentialité dont le locuteur considère la table, que sur un cheminement de pensées parallèles qui n'a pas à voir avec la table.

Cette correction renseigne aussi sur la construction syntaxique mentale du locuteur : par exemple la difficulté dans les premières années d'apprentissage du français ou de l'allemand à connaître les bons genres correspondant aux noms utilisés.

C'est donc la révélation pure et simple du processus de construction oral et des chemins que l'on tente parfois d'emprunter mais qui ne sont finalement pas les bons (au cours d'une longue tirade, se perdre et ne plus se retrouver dans la conjugaison de notre énoncé). L'autocorrection en tant que disfluence, se rapproche dans sa sémiologie de l'amorce modifiée.

Nous considérerons arbitrairement l'autocorrection courte et simple comme une disfluence puisque sa condition succincte fait que c'est un processus qui restera en marge et qui pourra s'apparenter au style du locuteur. Plus l'autocorrection sera longue, plus elle tendra vers le lapsus et ce qui était une disfluence devient un alors élément du discours au premier degré. Nous considérerons qu'une autocorrection courte et simple est constituée d'une répétition comprenant un déterminant, pronom ou mot grammatical permettant de corriger un accord ou une précision de lieu ou de temps. Ce choix arbitraire vient de l'absence d'étude portant sur l'autocorrection (Bove, 2008).

b. La sémiologie des disfluences

La sémiologie des disfluences a été peu étudiée, mais nous allons toutefois tenter de regrouper les informations à ce sujet pour mieux les saisir. La disfluence la plus étudiée reste pour le moment la pause silencieuse car elle représente souvent un temps de respiration, information utile outre les disfluences.

i. Les pauses non-sonores

Une caractéristique du discours oral spontané est la présence d'intervalles de silence et d'intervalles vocalisés qui n'ont pas de sens syntaxique. (Esposito, Stejskal, Smékal, Bourbakis, 2007)

Les rôles potentiels des pauses (Esposito, Stejskal, Smékal, Bourbakis, 2007) :

- créer de la tension
- créer de l'attente pour la suite de l'énoncé
- Aider l'auditeur à saisir le propos
- Témoigner de l'emphase
- Témoigner de l'anxiété
- Découper des syntaxes complexes
- Degré de spontanéité
- Transmission d'informations de genre
- Transmission d'informations socioculturelles
- Transmission d'informations de niveau d'éducation
- Régulation du flux de parole

Les pauses sont, dans les structures narratives, l'illustration d'unité narrative. Certains psychologues cognitifs ont émis l'hypothèse que les stratégies d'interruption sont le reflet de la complexité du fonctionnement neuronal. (Esposito, Stejskal, Smékal, Bourbakis, 2007)

Les causes physiques : respiration, processus articulatoire

Les causes socio-psychologiques : stress ou anxiété

Les causes communicatives : faire comprendre un message, laisser la place à une question, à une interruption.

Les causes cognitives : processus mental interrompu. Difficulté de conceptualisation. Divergence entre syntaxe et lexique.

Il n'y a pas de restructuration avec la pause non sonore, on peut toutefois pointer une série de marqueurs paralinguistiques générés par cette dernière :

- Recherche lexicale :

  On peut imaginer diverses raisons : manque de lexique dû à une éducation à parfaire, à une maitrise limitée de la langue, à une fatigue cognitive

- Respiration :

  Ce peut-être le marqueur d'une fatigue physique, mais aussi d'un état de stress accélérant le rythme respiratoire. Cela peut-être dû à une réaction de rire.

- Effet stylistique :

  L'information est relative au contenu oral, mais ces effets nous renseignent sur la personne les produisant.

ii. Les pauses sonores

La révélation de la mécanique fait partie de nombreuses expressions : tout enseignement permettant l'étude d'un geste et de ses erreurs (Artisanat, Sport, Musique...), tout expression artistique impliquant la mise en abime ou un concept questionnant l'existence même de l'œuvre (Cinéma, beaucoup de courants depuis les Arts Modernes...)... La révélation de cette mécanique est donc porteuse de sens. Dans notre exemple de la bibliothèque, on peut établir plusieurs postulats concernant le locuteur :

- Fatigue ou absence générale.
- Manque d'intérêt pour lieu.
- Gêne à l'évocation de la bibliothèque.

Ces pauses sont donc bien des indices sémiologiques très importants dans le discours oral car ces trois informations potentielles n'étaient absolument pas contenues dans le matériel linguistique. Il est possible de reporter les indices sémiologiques des pauses non sonores dans cette section.

iii. Les amorces complétées

Les informations sémiologiques transmises par ce genre de disfluences sont :

- Perturbation extérieure.
- Un accent mis sur le mot amorcé : il peut poser divers problèmes au locuteur, soit par son contenu (évocation inconsciente, gêne ou au contraire joie ou excitation), soit par sa prononciation. La tension créée par la disfluence peut donner la sensation à la fois d'une résolution, d'un apaisement mais aussi d'un aveu.
- Incertitude syntaxique/lexical dû à une défaillance physique : déficit d'attention ce qui retarde l'approbation du choix syntaxique/lexical et donc un piétinement.

- Défaillance des muscles articulatoires ou de l'appareil phonatoire : par exemple, des respirations mal placées empêchant de finir la production du mot, une tension dans le visage empêchant de continuer l'articulation.
- Incertitude syntaxique/lexical dû à un manque d'assurance : mauvaise maitrise de la langue, état émotionnel d'insécurité, contenu d'une idée dépassant le locuteur

iv. Les amorces modifiées

Ces amorces-là font office de corrections, de changements, de déviations. On perçoit la même sensation d'incertitude, de piétinement, seulement il y a ici légitimité puisque changement. La modification de l'amorce peut entraîner une grande modification syntaxique (dans le premier exemple, le verbe avoir est remplacé par le pronom je, la phrase est alors complétée avec le verbe être). Cette redirection a plusieurs causes et plusieurs effets, transmettant des informations sémiologiques :

- Perturbation extérieure.
- Erreur lexicale due à de la fatigue physique : déficit d'attention, déficit articulatoire. Le déficit articulatoire peut survenir : lorsque l'enchaînement de phonèmes se fait trop rapidement et les muscles articulatoires ne sont pas en parfaite synchronisation avec l'appareil phonatoire. On peut aussi avoir des problèmes de crispations ou de placement de la respiration.
- Erreur lexicale due à une mauvaise maitrise de la langue : choix d'un mauvais registre de langue, du mauvais mot, de la mauvaise prononciation, du mauvais genre, du mauvais accord.
- Ces scories peuvent aussi révéler un cheminement psychologique différent d'un cheminement oral public. Cela peut s'apparenter à un semi-lapsus, donc au presque-affleurement d'une information inconsciente. Sans entrer plus dans les détails, c'est en tout cas l'indice d'un cheminement parallèle à celui du résultat entendu. Cela crée une question sur l'information presque affleurée mais non révélée.

La principale différence entre l'amorce complétée et l'amorce modifiée, c'est bien l'affleurement d'un élément moins désirable (et non indésirable) qui implique une redirection. Que ce soit un élément du champ psychologique ou articulatoire, c'est bien le témoignage d'un circuit parallèle et donc paralinguistique dans la construction du discours oral.

v. Les amorces inachevées

Ici, la sensation de mystère, d'indécision peut-être plus grande. On peut l'interpréter comme différents phénomènes :

- Perturbation extérieure
- Un abandon d'une construction syntaxique/lexicale : mauvaise connaissance de la langue, construction inadaptée, propos trop simple ou complexe, erreur lexicale
- Fatigue physique : déficit de l'attention ou articulatoire (idem que pour les autres amorces) provoquant de l'incohérence.
- Révélation d'un mécanisme d'autocensure : produit la sensation d'un secret, d'un espace intime presque révélé. Le changement syntaxique appuie sur la nécessité d'une plus grande modification que dans l'amorce modifiée.
- Plus grande difficulté à construire un propos oral. Le passage de la pensée du propos à sa formulation est compliqué.

Les amorces sont soumises à un large spectre d'interprétations à la croisée des complexités du mécanisme phonatoire, de l'interprétation psychanalytique, de ces disfluence et de l'action même de la formulation orale comme d'un processus fait d'allers-retours, de modifications, de reprécisions et déconstructions.

vi. Les répétitions

La répétition se rapproche perceptivement de l'amorce, levant toutefois le mystère sur le mot de butée. On ressent donc de nouveau le piétinement et toute les suppositions sur l'état du locuteur. La répétition se rapproche perceptivement d'une amorce complétée.

vii. Les autocorrections

Cette correction provoque une perception similaire à l'amorce modifiée. La prononciation totale des mots à corriger atténue toutefois la notion de piétinement, créant aussi un rapport plus naïf à la langue employé. L'autocorrection peut véhiculer un aspect de découverte.

Voici un tableau récapitulatif des informations rencontrées.

Type de disfluence	**Sous-type de disfluence**	**Qualification**	**Perception potentiel**
Pause	Pause Silencieuse	200ms < x < 2000ms	- Besoin de temps pour une recherche lexicale - Fatigue physique ou psychologique / Absence - Effet stylistique, prosodique propre au locuteur
X	Allongement syllabique	x<1,8s Concerne en général une voyelle en fin de mot	- Besoin de temps pour une recherche lexicale - Fatigue physique ou psychologique / Absence - Désintérêt pour les mots suivant la zone d'allongement - Gêne pour les mots suivant la zone d'allongement - Intense réflexion - Refus de donner la parole - Manifestation d'une idée en cours
X	Morphème spécifique		- Besoin de temps pour une recherche lexicale - Fatigue physique ou psychologique / Absence - Désintérêt pour les mots suivant la zone d'allongement - Gêne pour les mots suivant la

			zone d'allongement - Intense réflexion - Refus de donner la parole - Manifestation d'une idée en cours
Les amorces	Complétées	1 amorce tout les 250 mots/ On ne peut amorcer que des mots d'au moins 1 phonème / Complexité et longueur	- Défaillance physique générale ou relative à la prononciation du mot amorcé, à la gestion de l'appareil phonatoire et des muscles articulatoires. - Incertitude syntaxique/lexical. Manque de confiance. - Accentuation sur le mot amorcé et questionnement de son rapport au locuteur : rapport émotionnel ou mémoriel fort (gêne, joie, excitation). - Créer une sensation d'obstacle - Le fait de surmonter l'obstacle peut produire une sensation d'aveu ou d'apaisement
	Modifiées	1 amorce tout les 250 mots / On ne peut amorcer que des mots d'au moins 1 phonème / Complexité et	- Défaillance physique : déficit de l'attention, rythme articulatoire trop élevé, crispation faciale ou respiratoire - Mauvaise situation de

		longueur / Topologie de la modification	langage : mauvais registre, mauvais mot, mauvaise prononciation, mauvais accord... - Semi-lapsus créant un questionnement sur le mot disparu
	Inachevées	1 amorce tout les 250 mots / On ne peut amorcer que des mots d'au moins 1 phonème / Complexité et longueur / Topologie de la modification Et appréciation subjective	- Abandon syntaxique/lexical : redirection du propos, remaniement de la syntaxe pour plus de précision, plus de simplicité. - Fatigue physique/articulatoire voire incohérence, - Construction du propos difficile, - Autocensure ou conflit intérieur
Répétitions		Redondance de mot entier sans sens grammatical	Idem que l'amorce complétée
Autocorrection		Répétition puis correction d'un court mot grammatical, pronom ou déterminant en vue de corriger une erreur de conjugaison, de grammaire de préciser un lieu ou un temps.	- Idem que l'amorce modifiée - La notion de piétinement est toutefois atténuée - Véhicule aussi fortement la pratique ou la découverte d'une langue en cours (en plus de l'élaboration d'un énoncé oral)

c. La question des profils

L'une des questions de ce mémoire et d'établir le lien qu'il existe entre l'identité vocale et la disfluence. Il est certain que la construction de l'oralité se fait à l'aide de la disfluence, et que chaque personne « disflue » selon des règles qui lui sont à la fois propres, tout en étant organisées par une construction d'influences divers.

On trouve peu de travaux concernant la synthèse des disfluences dans la langue française, toutefois leur suppression automatique pour créer des transcriptions textuelles « nettoyées » est un sujet très étudié. A travers deux études, on constate qu'il n'y a pas de modèles de disfluences apparents, genres, langues et situations confondus.

Dans l'étude de Adda-Decker faite au LIMSI-CNRS en 2004, à l'aide d'un corpus de 20 personnes fait à partir d'interviews, on constate les écarts de chiffres dont la variance ne semble pas répondre à une loi simple.

code	sens	#loc.	code	sens	#loc.
J	journaliste	9	**I**	interviewé	11
C	chairman	1	**w**	femme	1
e	natif anglais	1	**o**	personne âgée	1
r	*accent* région.	1	**c**	société civile	2
f	francophone	2			

Figure 11 Dénomination des sujets du corpus de Adda Decker

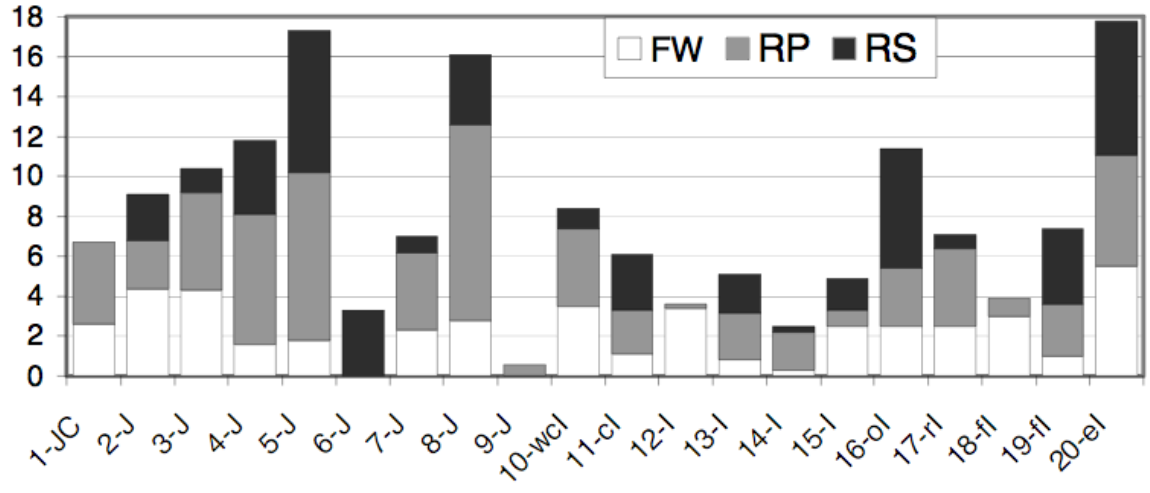

Figure 12 Résultat des quantités de disfluences constituant la totalité du discours en pourcentage / FW = Filler Word = Morphème spécifique + Allongement syllabique / RP = Répétition / RS= Restart = Correction + Amorce

On constate que pour les 9 journalistes, on obtient 9 résultats différents, il n'y a donc pas de comparaison à entreprendre avec les autres sujets (interviewé homme ou femme, ou

natif américain...). De plus l'absence d'informations relatives à l'état des sujets au moment de l'enregistrement nous éloigne de la problématique identitaire de ce mémoire.

Il aurait donc fallu avoir accès à un corpus d'enregistrements de voix françaises suffisamment conséquent et annoté pour pouvoir tirer des hypothèses satisfaisantes. Il s'avère que la tâche pour y accéder se révéla ardue.

Sur recommandation, je me suis orienté sur deux corpus :

- The Nijmegen corpus of casual french : Corpus annoté de 35 heures de français spontané. Il regroupe 46 sujets dont les seuls informations que l'on possède sont : sexe, âge, lieu de naissance, lieu dans lesquels le sujet a vécu, lieu de naissance des parents, accents des parents, accent du sujet, niveau d'étude, autres langues parlées... A partir de ce corpus ont été relevées un certain nombre de disfluences et les comportements que cela implique dans une discussion à deux. Il n'y a toutefois pas d'annotations sur un état émotionnel, sociologique ou affectif des sujets. Il n'était donc pas possible d'en extraire des informations relatives au sujet de l'identité du locuteur.
- Un corpus annoté en fonction d'états émotionnels fait à partir d'une plateforme nommé E-Wiz (en référence au magicien d'Oz, que l'on donne à voir au sujet avant qu'il ne restitue sa version du film à l'oral). Je n'ai tout simplement pas réussi à avoir accès à ce corpus, dont il aurait peut-être été possible de tirer des informations bien qu'il ait été établi pour la synthèse des affects et non des disfluences.

Malgré l'impossibilité de mettre en œuvre cette recherche, j'ai pu imaginer une démarche future en relation avec ma partie pratique pour établir un corpus à analyser. Ce protocole sera décrit en dernière partie du mémoire.

Puisque nous ne pouvons donc pas établir de profil concret à partir d'enregistrements, nous pouvons toutefois établir un axe sur lequel la disfluence passe d'une fatigue physique à des conditions psychologiques variables. Au centre de cet axe se trouvent les disfluences qui interviennent de manière sporadique illustrant le fonctionnement quotidien de l'oral, celui de l'accession au lexique alors qu'une syntaxe est déjà établie. À la vue de la variété des sens transmis par la disfluence et leurs rapports entretenus avec le lexique, la syntaxe, et l'identité vocale du locuteur, il serait bien périlleux de se lancer dans un profilage. Il s'agit donc d'établir de manière arbitraire des quantités de disfluences et d'écouter le résultat pour

potentiellement produire un corpus qui pourra être plus tard analysé perceptivement. A partir de cette analyse, on pourra alors nommer les quantités de disfluences en lien avec le lexique et la syntaxe et l'identité vocale, comme des profils.

D'une manière générale, toutes ces disfluences sont des éléments de mise en abyme qui font apparaître les mécanismes de construction du discours. Ils ont tous la faculté de passer toutefois inaperçus et d'être des mécanismes camouflés de l'oral qui ne se révèlent que si l'on y prête attention. Ils possèdent par contre une influence sur la perception de l'énoncé et de l'identité du locuteur, pouvant susciter autant de compassion, que d'agacement. L'augmentation du nombre de disfluences ou leur intensité (par leur complexification ou leur allongement) fait que l'on rentre dans des troubles du langage bien identifiés, dont le lapsus peut faire partie. Bien que nous ne développions par une étude de tous les troubles du langage, nous allons nous pencher sur un trouble particulier : le bégaiement.
Si une partie entière est dédiée au bégaiement au sein de ce mémoire sur les disfluences, c'est que ce trouble encapsule d'une part toutes les disfluences étudiées auparavant à des degrés d'intensité variable, mais aussi parce que les symptômes de ce trouble sont à la croisés des problématiques d'identité vocale, de voix performative et du rapport d'un individu à son corps, à son élocution.

3. Le bégaiement, le cas de l'individu

Le bégaiement, comme les disfluences, est difficile à définir en soi. Il est tellement compliqué d'y distinguer des patterns répétitifs, des motifs archétypaux, d'en tirer des analyses et des liens de causes à effets, qu'en dehors de la nomenclature médicale, le bégaiement ne semble être qu'une suite de cas particuliers (Patrick Poisson, 2002).

Pourtant le bégaiement est un trouble du langage référencé par une série de classification médicale : F89.5 selon la CIM[16]-10, 307.0 selon la CIM-9, 184450 selon la OMIM[17] et D013342 selon la MesH[18]. Cette terminologie décrit une série de symptômes que la plupart des auditeurs de la planète, quels qu'en soient la langue, le sexe, le milieu

[16] Classification Internationale des Maladies
[17] Online Mendelian Inheritance in Man qui est un catalogue de gènes pouvant mener à la compréhension de certaines maladies et à leur transmission
[18] Medical Subject Heading

socioculturel du locuteur, ont déjà rencontré. En effet, 1% des adultes sont touchés par le bégaiement (Bernadette Piérart, 2011). Il est possible d'établir une large définition en croisant tous les symptômes des nomenclatures ci-dessus: Le bégaiement est un trouble du langage altérant la fluence verbale d'un locuteur par l'utilisation de pauses intempestives ou d'allongements chaotiques, de répétitions et de bruits vocaux, empêchant la compréhension d'un locuteur.

Une série de questions se pose alors par rapport aux disfluences évoquées auparavant : Comment passe-t-on d'un discours oral spontané en train de se construire à l'incapacité de communiquer ? Qu'est-ce que l'incapacité de communiquer ? Comment se caractérise-t-elle ? Quelle est la sémiologie de ce trouble ?

Le corps du bègue semble être un corps en crise, d'après Maurice Blanchot dans « partenaire invisible ». Le diaphragme contrôlant la respiration, les muscles de la face permettant l'articulation et tout le système phonatoire semblent répondre à des règles difficilement prévisibles. Nous essaierons donc de décrire comment, dans son rapport étroit avec le corps, s'incarne l'identité vocale de l'individu atteint de bégaiement. Nous verrons enfin quel peut-être l'intérêt d'aller explorer cette limite au sein de la synthèse vocale.

i. Définition du bégaiement

La disfluence est un phénomène indissociable de la voix orale spontanée et se caractérise par des reliques de la construction d'un énoncé, donnant des sensations parfois de piétinements, de cahotements, de changements de directions. Nous avons aussi vu que c'était un phénomène auquel on prêtait peu d'attention. En effet, le nombre de disfluences est tellement grand que leur perception est automatisée. Bien que « non-conscientisées »[19], elles permettent la bonne compréhension d'un énoncé et de son rapport au locuteur. Le trouble du langage survient lorsque la fluence est tellement perturbée que le locuteur ne peut exprimer son énoncé, ne peut se faire comprendre de l'auditeur.

Quels sont les symptômes qui permettent de différencier un énoncé contenant un grand nombre de disfluences d'un énoncé de quelqu'un atteint de bégaiement ? Autrement dit, à quel moment la disfluence devient-elle suffisamment perceptible pour devenir un objet sémantique aussi fort que le contenu de l'énoncé.

[19] Dans ce mémoire, il n'est pas question du mécanisme de cognition de la disfluence. Le terme non-conscientisé permet donc de circonscrire un champ différent de celui de la prise de conscience auditive du premier degré.

Le bégaiement est un trouble connu au moins depuis 377 avant J-C (Hippocrate écrit ses premières analyses à son propos à cette époque) constituant un handicap physique et social avéré. Cette maladie accompagne l'Histoire des Sciences, menant à une symptomatologie multiforme dont les hypothèses pour la circonscrire traversent les champs médicaux (neurologie, génétique, audition, psychiatrie), psychologiques, religieux, sociologiques ect... Ce trouble intervient dans beaucoup de cas dès la petite enfance (dès le développement de la parole) en s'accentuant vers l'âge de quatre ans (dès la formation de phrases et la manipulation de la syntaxe).

La perturbation de la fluence n'est que la caractéristique la plus apparente de ce trouble, qui s'accompagne souvent d'autres symptômes linguistiques, émotionnels ou comportementaux (Bernadette Piérart, 2011).

Il existe une classification duale populaire opposant un bégaiement neurogénétique et un bégaiement psychogénétique. Le premier surviendrait lors d'une blessure à l'écorce cérébrale. Avant l'accident, le patient présente une fluence correcte. Après le choc, le patient montre une désorganisation de la parole et un trouble de la fluence constante et régulière. Ces cas de bégaiements neurogénétiques sont décelés selon des diagnostiques fortement discutés (Bernadette Piérart, 2011)

Le bégaiement psychogénétique, lui, est le bégaiement rencontré au quotidien. Les troubles ne sont pas réguliers ni constants et semblent dépendre d'une série de facteurs intérieurs et extérieurs qui varient en fonction du patient. Après 2000 ans d'étiologie[20] du bégaiement, des facteurs neurologiques semblent devenir une piste solide, sans impliquer de traumatismes préalables.

Toutefois, nous nous reposerons, pour le moment, sur le diagnostic du bégaiement par Wingate en 1964 :

- Spasmes et blocages des mouvements menant à l'émission de parole
- Répétitions involontaires audibles ou silencieuses
- Allongements de certaines unités brèves de la parole
- Mouvements parasites de l'appareil phonatoire, ou d'autres parties du corps non liées à la communication

[20] Etude des causes et des facteurs d'une maladie.

Cette « désintégration de la coordination de tous les mouvements qui participent à la parole »[21] mène aussi à des troubles plus discrets, internes, accessibles par « introspection du sujet »[22] (qui ne sont pas inclus dans le diagnostic de Wingate) et qui concernent moins l'appareil phonatoire :

- Etats visibles de tension, d'irritation ou d'embarras menant à une souffrance psychologique
- Troubles d'accès au lexique
- Perturbations morpho-syntaxique

On distingue donc deux sortes de troubles générés par le bégaiement : les troubles phonatoires (respiration, phonation, articulation) et les troubles psychologiques (psycholinguistique, psychologique, psychiatrique) dont les liens de causes à effets semblent bilatéraux.

Selon la définition de Wingate, les symptômes phonatoires du bégaiement sont des phénomènes similaires aux disfluences. Dans bégaiement et contexte syllabique de Patrick Poisson (2002), on identifie quatre types de bégaiement génériques :

- Blocage-tension / Trouble de la respiration :

  Ces blocages sont dus à des troubles de la respiration, des inspirations trop brèves, saccadées, mal contrôlées. Les patients font « des reprises respiratoires intempestives, en urgence, au beau milieu d'un mot » (Bernadette Piérart, 2011).

  Il y a désynchronisation entre l'inspiration et l'expiration (Perkins, 1976) menant à cet état d'urgence. Les blocages respiratoires ont tendance à survenir sur l'attaque des mots et correspondent à des poussées expiratoires. La désynchronisation du diaphragme et de l'émission phonatoire semble produire un objet similaire à l'amorce, sans pour autant qu'il remplisse cette fonction de construction de l'énoncé oral. La tension des muscles oraux-faciaux peut mener à ce que les phoniatres appellent « mâchoire de bois », ce qui correspond à la paralysie du bas du visage et à l'arrêt total de la phonation.

[21] Bernadette Piérart, 2011
[22] Bernadette Piérart, 2011

- Les allongements vocaliques

  Les allongements vocaliques sont comparables à des pauses remplies. C'est en général une matière produite de manière volontaire pour essayer de masquer le bégaiement.

- Les ajouts

  Les ajouts correspondent à deux catégories : l'utilisation de conjonctions d'appuis, qui s'apparentent à des tics de langage permettant de prendre de l'élan ou au contraire de structurer les arrêts dus au bégaiement. Ce sont aussi les bruits vocaux produis par les spasmes du larynx, la tension des muscles articulatoires menant à une émission explosive ou saccadée, absente de sens sémantique.

- Les reprises répétitions

  Les reprises et répétitions fonctionnent à l'image des amorces et des répétitions vues dans le chapitre sur les disfluences.

Quel est donc l'intervalle à franchir pour passer de la disfluence au bégaiement ? Nous avons vu que le bègue produit des sons buccaux plus variés que le locuteur lambda, de par la tension et le chaos qui règnent parfois dans l'appareil phonatoire.
Selon un référentiel psycholinguistique, un autre phénomène entre en jeu dans le bégaiement : « Il n'y a pas de préparation du moule mélodico-rythmique de la phrase » (Pichon, Borel-Maisonny, 1960)
D'autres études donnent des indices pour caractériser le fonctionnement du bégaiement par rapport à celui de la disfluence : chez les bègues, la prolongation de pauses vides s'accompagne de spasmes au niveau du larynx et de la face. Van Ripper (1971) considère qu'à part les mots typiques du bégaiement, les pauses du bègue ne sont qu'une exagération des mécanismes présidant aux pauses chez le locuteur normal. Il est donc possible de supposer que le bégaiement découle d'une impossibilité à concevoir une phrase avant son énonciation.
La disfluence fait partie du processus, seulement dans le cas du bègue, et serait le vestige de l'incapacité à pré-produire une prosodie. C'est un résumé schématique qui n'englobe qu'une partie de la symptomatologie du bégaiement, seulement il me semble important d'appuyer sur le fait que le bégaiement est finalement « une exagération de la disfluence ». Si l'on reprend d'ailleurs l'étude de Patrick Poisson, « il est démontré que le bégaiement n'est causé par

aucun type de syllabe en particulier ». Il ne semble pas y avoir de liens directs avec le geste phonatoire à produire. Bien que les bégaiements semblent survenir la plupart du temps sur des mots à contenu, les lexiques impliqués par ces mots à contenu sont relatifs à chaque individu. Pour finir, je cite la conclusion de cette étude :

« Bensalah(1997) mentionne que plusieurs recherches américaines concernant le bégaiement se sont avérées décevantes pour leurs auteurs. En effet nous remarquons que les bègues ne butent pas tous sur les mêmes phonèmes […] De plus, un regroupement de toutes les données ne permet pas de déterminer un groupe de sons difficiles sur lesquels les sujets buteraient systèmatiquement. […]

À en croire nos données, le bégaiement est propre à chaque individu. Malgré tout, il reste certainement des éléments de réponse à découvrir. »

Le bégaiement est donc une exagération individuelle des phénomènes de disfluences impliquant des difficultés d'énonciation à travers un trouble de la fluence verbales. Ces troubles sont d'ordres physique et psychologique.

On peut supposer que les limites de disfluences établies au chapitre précèdent, suggèrent qu'au-delà, l'énoncé se disperse, se dissout, s'efface, témoin d'un trouble du langage.
En reprenant le tableau (2)2)b)vii)), on peut alors établir la prise de conscience d'une disfluence qui malmène l'énonciation. Mais il serait simpliste de ne considérer que ce simple dépassement de limite pour entrer dans la zone du bégaiement. L'altération de la fluence implique aussi une altération de la vitesse d'énonciation. C'est donc aussi l'augmentation du nombre de disfluences, leur étirement et les variations de la vitesse d'énonciation, qui dilue l'attention de l'auditeur, empêchant la communication de l'énoncé premier.

Nous entrevoyons maintenant la complexité d'analyse que requiert le bégaiement. Le handicap de communication que cela engendre n'empêche pas pour autant la communication. De la même manière qu'être privé de jambes ne prive pas de manière absolue de mouvement, être bègue ne prive pas de communication. Cette dernière se fait toutefois dans une sphère hors de la norme, dans un référentiel différent, avec des signes et des gestes différents, qui appartiennent au registre de la crise.

ii. Sémiologie du bégaiement

Le bégaiement n'existe que dans la confrontation ; il faut être deux pour bégayer. Ce qui revient à dire que le bègue n'existe que dans un rapport à un contexte. Ce n'est pas pour rien qu'une grande quantité de groupes de soutien aux personnes atteintes de bégaiements se composent de discussions entre malades. Hors du contexte social établi comme fonctionnel, où l'on attend une information en tant de mots, avec un moule prosodique défini, il n'y a certainement pas la même tension insufflée dans l'appareil phonatoire du malade. Il y a donc un double rapport entre le bègue et l'extérieur : son rapport au locuteur, et son rapport à la fonction communicative. Ces deux rapports créent des réactions psycholinguistiques qui portent du sens : « un mouvement intérieur avec des effets extérieurs » dit Blanchot. Il s'agit ici de faire un effort identique à celui fait pour la disfluence, et de considérer le système de communication du bègue comme un système fait de signes et non d'erreurs.

« Le bégaiement est un élément actif qui amène à la reconsidération de la relation entre l'individu et la communauté » Blanchot

Le bégaiement serait un phénomène de porosité entre le collectif et l'individuel, survenant uniquement lors de performances collectives. La performance implique une série de gestes, et ces gestes créent une convention sociale : regarder son interlocuteur dans les yeux pour diriger le flux de parole selon un destinataire, parler rapidement pour marquer de l'excitation, des expressions faciales accompagnants des mots ... Ce mémoire n'a pas pour but d'établir la convention de la communication orale spontanée mais il serait possible d'utiliser les archétypes des disfluences établies au chapitre précèdent pour agrémenter une telle convention.

Cette convention – qui n'est pas un objet figé et simple à définir, mais tout aussi vivant qu'une langue orale - sous-tend, au sein de la communauté, les échanges oraux. Le bègue, s'inscrit donc dans la marge de cette convention, un peu comme le concept du « fou » qui fait exister la limite arbitraire du « normal » . Le bègue, dans sa performance vocale, est donc un élément primordial de la définition de la convention.

Cette convention est en relation directe avec la représentation orale ainsi que la transmission d'une idée mais aussi avec l'identité du porteur de l'idée, comme vu auparavant, par la syntaxe, le lexique, les disfluences, les marqueurs vocaux, la prosodie... Le bègue n'est pas le maître conscient de cette représentation.

« « Je » est un autre » Rimbaud

Cet autre qui s'incarne, qui pourrait être une des constructions faites sur le désert ascétique de Deleuze, qui pourrait aussi se traduire comme « l'impossibilité de se représenter ». Il y a une idée poétique derrière cette impossibilité, qui ne doit pas éclipser le mal être et la difficulté individuelle. Seulement, à une échelle collective, le bégaiement s'incarne comme l'éclatement de la norme. C'est dans cette relation à la représentation communicative du soi, que surgit autant l'impossibilité pour le bègue de correspondre à cette représentation, que le malaise de l'auditeur face à une crise qui touche à l'acte de parole. Les accidents du langage deviennent alors un parcours difficile, qui télescope la performance vocale au sein d'un paradoxe : l'envie et l'impossibilité de dire.

« Le bégaiement n'est pas un défaut individuel, mais plutôt la rétention du langage à un degré d'absence de parole. Et de cela émerge quelque chose qui étonne, effraie, dérange et repousse tous les locuteurs, tous les auditeurs, de leur état de confort » Blanchot

Le mouvement intérieur vers l'extérieur devient aussi une source de pression sur l'auditeur. Si l'on apparente le bégaiement à un trouble allant du TIC au TOC, il est certainement l'élément amenant à une compassion instantanée. L'impossibilité de dire, et donc de se transmettre, voire d'être, est quelque chose qui pousse à une relation compassionnelle avant d'être de l'agacement, de la gêne, du rejet... Nous verrons plus loin que cette compassion est certainement issue de la crise corporelle qui accompagne la phonation.

iii. Portrait potentiel du bègue

Le portrait psycholinguistique général de la personne atteinte de bégaiement a été établi à partir de l'ouvrage « Le bégaiement de l'adulte » et d'un corpus de voix spontanée oral de personnes atteintes de bégaiement. Il est élément archétypal qui ne fonctionne que d'un point de vue empirique : expériences de médecins spécialisés et de patients réunis.

Le bègue est une personne évitant certaines situations de communication complexe : en public, face à un supérieur hiérarchique, face à des inconnus, au téléphone. L'exemple du téléphone est un exemple intéressant puisque c'est le moment où le bègue est réduit à sa seule expression phonatoire. Dans de nombreuses confidences, les personnes bègues vivent mal la conversation téléphonique avec des inconnus. Le corps n'est pas là pour chorégraphier le dérèglement de la parole.

« Maintenant quand j'téléphone, je préviens de suite **mo- mo- +, mo- ++,** la **p- p- p- p-** personne que je suis bègue » raconte une femme dans une émission de téléréalité à propos de son bégaiement au téléphone.

Toutefois, le rapport émotionnel à la discussion engagée dépend de la situation modulant les efforts à fournir. Les mauvaises expériences familiales, les difficultés d'intégrations sociales, ainsi que la conscience du trouble participent de la construction du comportement du bègue. Souvent perçu comme anxieux, nerveux, incohérent voire atteint de désordre mental, on le perçoit aussi comme un battant s'il est capable de surmonter son handicap ou qu'il occupe une place à responsabilité. Son silence gêné peut parfois être perçu comme de l'entêtement, de la vanité ou du mépris. Ses symptômes s'aggravent lors de fatigues et de malaises physiques. Ils s'améliorent dans des conditions différentes du mode d'élocution habituel : chant, chuchotement, prosodie hachée...

Le profil psychologique du bègue ne s'établit que par le croisement de confidences faites aux psychologues, psychiatres, phoniatres, orthophonistes... Il s'en dégage deux axes : l'anxiété verbale et le continuum gêne-culpabilité (Bernadette Piérart, 2011). L'anxiété verbale est l'anxiété de la situation de communication, générative de malaise voire de colère contre sa propre-insuffisance. Le continuum gêne-culpabilité correspond à l'anticipation de l'échec potentiel de communication produisant de la gêne, augmentant la situation de tension. Si cet échec survient, un sentiment durable de honte s'installe, mettant un terme aux tentatives de communications. Toute la construction de l'identité vocale du bègue se fait autour de cette communication avec l'extérieur et de ces deux axes.

« Jean-Baptiste me dit qu'il a souvent rêvé d'un texte qui lui serait écrit et qu'il pourrait « réciter » aux jeunes filles dont il est amoureux. [...] Il l'associe à la situation de Christian dans Cyrano de Bergerac d'Edmond Rostand » Jean-Michel Vives, « Le bégaiement de l'adulte »

Cette citation illustre deux choses : la potentielle disparition, chez le bègue, de l'accès au texte avant son énonciation mais aussi la présence d'un second être, capable mais « non-soi ». Si la psychanalyse s'empare trop facilement du bégaiement, c'est parce qu'elle est le parfait chevalet pour la névrose[23] comme rupture intérieure. Si toutefois l'on revient à la

[23] Pour Jung, la névrose est une rupture intérieure entre plusieurs sources d'énergies. Cette rupture produit alors une ombre imprégnant l'identité du névrosé.

définition de l'identité de Deleuze, ce second être est un habitant de la faune et de la flore du désert qui nous peuple. Sans remettre en cause le handicap que produit le bégaiement sur la vie du malade, il est possible d'imaginer que ce qui créé le mystère autour de cette maladie vient du fait que ce n'est pas tant une maladie, qu'une rencontre entre une incompatibilité avec le monde et une configuration personnelle particulière. Une autre manière de l'exprimer consiste à dire que le bégaiement se développe dans un contexte où l'expression orale est une expression faite d'un langage organisé, voire fonctionnel. Le bègue rend la disfluence visible et devient alors un élément de marge. Tout le rapport qui se construit entre un bègue et sa manière de bégayer est selon moi un objet identitaire fort.

Comment parler de cette identité tout en sachant que, comme nous l'avons vu, chaque bégaiement est en pratique unique. Après avoir abordé la crise corporelle qui est un élément constitutif de l'identité du sujet bégayant, nous traverserons un corpus d'extraits sonores nous permettant d'établir une potentielle modélisation du bégaiement.

iv. Le corps en crise et l'identité du bègue

Le corps du bègue est un corps en tension, explosion puis relâchement. La crise corporelle ne s'explique pas selon un axe de cause à effet. C'est un système qui comprend au moins deux pôles certains : un dérèglement phonatoire dû à une situation de communication et une série de gestes d'accompagnements du trouble.
De plus ces gestes constituent la limite par rapport à notre sujet, car il n'est pas question de représenter le corps de notre voix synthétiques ; c'est un organe de communication dont Dennys sera privé.

« Les produits du bégaiement montrent la somatisation dans une totalité spasmodique » Blanchot

Les spasmes généraux qui peuvent aller jusqu'à secouer le corps du bègue portent le nom de syncinésies. Ce sont des mouvements involontaires d'un muscle ou groupe de muscles alors qu'un autre mouvement volontaire est en train d'être effectué : clignements d'yeux, froncements de sourcils, plissements du front, tremblements de la musculature faciale, claquements de langue, prostrusions linguales, tension laryngée, mouvements incohérents de la tête (n'accompagnant pas une idée syntaxique). D'autres spasmes sans lien neurologique avec l'appareil phonatoire surviennent : gestes de la main cachant le visage,

mouvements de la tête, du tronc, des jambes, élévations des bras ou des épaules, crispations des mains, claquements de doigts, appuis des pieds, rires nerveux... L'un des principaux handicaps de communication dans cette crise est la perte du contact visuel avec l'interlocuteur (Le Huche, 1998). Il reste enfin les réactions dites neurovégétatives : rougeurs du visage et du cou, hyper ou hypo salivation, hypertension, tachycardie, sudation excessive.

Cette crise corporelle est mise en place à la fois consciemment et inconsciemment par le bègue. Elle permet de dissimuler la gêne, d'attirer l'attention sur autre chose que l'émission de parole, mais aussi de réguler le rythme de la prosodie devenue chaotique à cause des spasmes. Une autre manière d'aborder cette crise, c'est de la prendre comme un acte performatif créatif :

« Ce qui doit arriver est un piratage du langage. Créer a toujours été une chose différente de la communication. Pour échapper au contrôle, peut-être que la chose la plus importante est de créer des vacuoles d'interruptions de la communication » G. Deleuze

La confrontation au collectif et la crise font que l'acte performatif devient inattendue, il est une création. Toutes ces imbrications font qu'il subsiste un mystère autour de cette maladie et de ses symptômes, poussant à qualifier chaque patient individuellement.
C'est pour cela que j'ai établi un corpus de bégaiements différents afin d'en noter les différences, les appréciations et les perceptions individuelles que j'en ai, pour essayer de modéliser plus fidèlement le bégaiement qu'à travers sa symptomatologie.
J'ai rencontré une grande difficulté d'accès à des enregistrements variés de voix orales spontanées de bègues français. Pour établir des directions de profils quelconques à partir de bégaiements, il aurait fallu organiser un corpus d'enregistrements annotés contenant les informations relatives aux locuteurs : des informations physiologiques (âge, sexe, origine, taille, état actuel) mais aussi une description personnelle de ce que le locuteur estime être son identité (milieu social, éducation, tempérament, goûts généraux, particularités...). Il aurait aussi fallu créer différentes situations de communication : avec une personne connue, une personne inconnue, en intérieur, en extérieur, au téléphone, avec deux personnes, trois personnes...

Ce corpus aurait alors pu être analysé pour déceler non pas un comportement logique au sein de la totalité du trouble, mais bien pour produire une cartographie du bégaiement de chaque personne s'étant soumis à l'analyse. Ces profils alors, constitueraient une information relativement fiable, qui par comparaison pourrait induire une meilleure connaissance du fonctionnement de mécanismes du bégaiement.

J'ai établi un corpus personnel en français et anglais, sans annotation, dont la situation est souvent identique : une interview face caméra, avec une ou plusieurs personnes. Les informations sur les sujets interviewés sont manquantes et la quantité est trop faible pour qu'il y ait un substrat à tirer des ces extraits. Ils constituent toutefois un panel de bégaiements variés qui traversent les différentes crises corporelles évoquées ci-dessus et peuvent illustrer certains éléments sémiologiques cités auparavant.

- NSA : Association de femmes américaines bègues parlant de la difficulté d'avoir ce handicap en tant que femme : https://www.youtube.com/watch?v=MR4RrTTciMw
- Reportage sur le bégaiement : Canada
- Documentaire Youtube : Dans l'ombre du discours, 2013, Réalisé par Adélaïde Palvadeau et Jérômine Lucas, comportant un grand nombre de portrait de personnes françaises atteintes de bégaiements témoignant face caméra : https://www.youtube.com/channel/UC2OJFJ3kMKaCVAMy-aQ8wsA
- C'est ma vie, M6, Décembre 2013 sur le bégaiement : https://www.youtube.com/watch?v=LaI_LfTVeYo

v. Modéliser le bégaiement

Un jeu de mots pourrait pousser à se dire : le bègue bug. Nous avons bien compris que ce n'est pas le cas. La problématique pour la synthèse de cette forme de disfluence exagérée, de ces nouvelles identités est la suivante : comment, dans la synthèse en se passant de la représentation du corps, allons nous donner à écouter le bégaiement sans transmettre la notion du bug ?

Dépasser la disfluence, c'est la rendre audible. Une machine sans intelligence artificielle indépendante et évoluée n'a pas de raisons de bégayer (pas de notion de troubles neurologiques, pas de sensation fatigue, pas de sensation de stress). S'il y a erreur, c'est que ce n'est pas prévu par le programme donc c'est un bug. Il ne peut donc pas y avoir de simulation du bégaiement, s'il n'y a pas d'une part une identité vocale cohérente, et la simulation d'une identité entière. Les programmes qui entrent dans ce groupe là seraient potentiellement des programmes dotés d'intelligence artificielle.

Il existe donc différents schémas de progression possible pour faire apparaître le bégaiement dans ce type de simulation. Considérons une erreur audible dans la fluence d'une voix. Si elle est unique, elle reste une erreur, un bug. Si cette erreur entraîne un dysfonctionnement et l'arrêt du programme, elle reste une erreur. Par contre, si cette erreur se répète, ou qu'elle entraîne une autre série d'erreurs, alors elle est un trouble dont la sévérité est variable.

Si on imagine qu'au contraire cette erreur se répète, elle peut le faire de deux manières : en créant une organisation rythmique, ou en créant de la désorganisation. L'organisation rythmique, par exemple répéter 5 fois une amorce à intervalle régulière, se retrouve de manière récurrente dans la plupart des bégaiements. Seulement il existe une différence entre ce que l'on appelle un intervalle perçu comme régulier, très légèrement variable par exemple entre 16 et 30 ms, et la répétition parfaitement cadencée que peut produire une machine. On constate à l'écoute du corpus que les intervalles entre chaque manifestation de bégaiement ne sont pas parfaitement réguliers. Je prends en exemple une phrase du jeune garçon du reportage C'est ma vie : « hé ben non, je vou- ou- ou- ou- ou- drais pas trop qu'il vienne. » Cette répétition rapide semble fait d'intervalles réguliers.

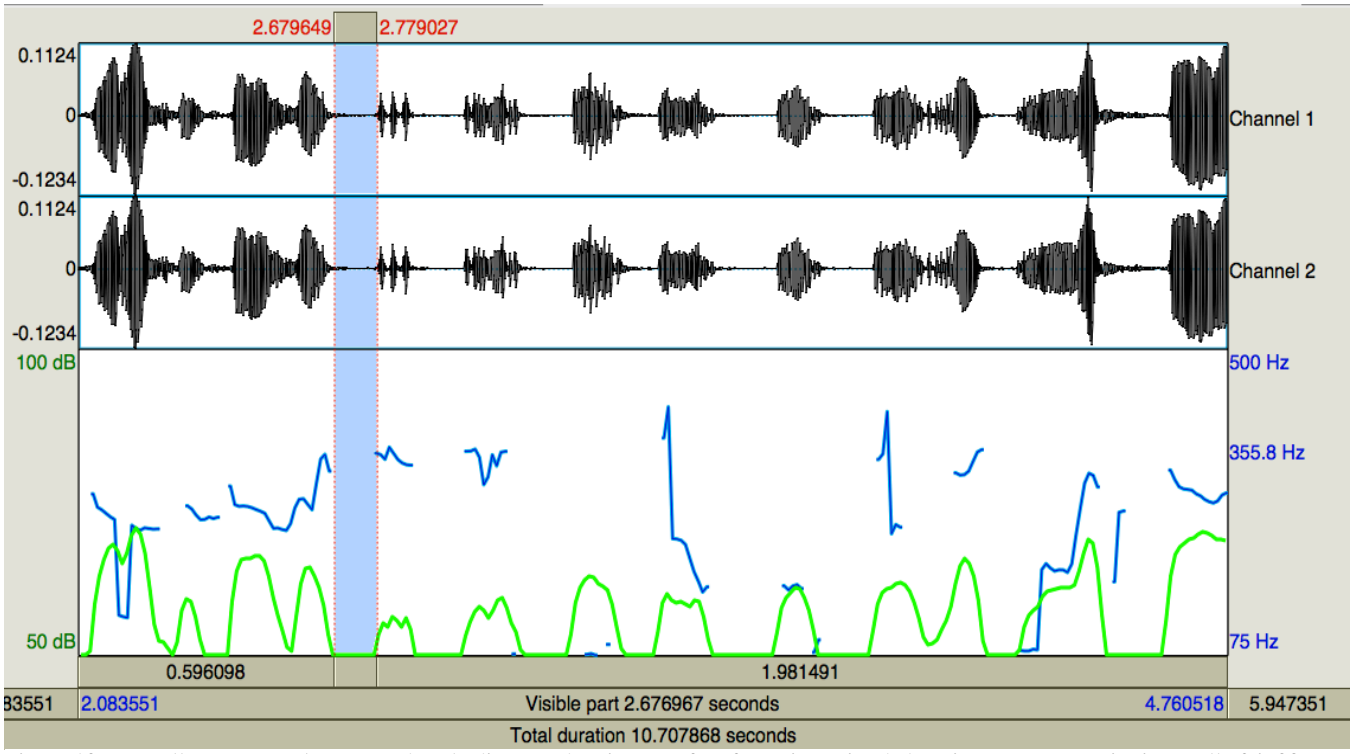

Figure 13 Intervalles séparant 5 amorces dans le discours d'un jeune enfant français atteint de bégaiement. Le premier intervalle fait 99ms

A l'aide de PRAAT, j'ai analysé la forme d'onde et les intervalles entre chaque amorce de « ou ». On observe alors les intervalles suivants : 99ms, 114ms, 110ms, 73ms, 142ms, 107ms. On voit donc que l'excursion maximum des intervalles est de 69 ms. La variation est donc suffisamment petite pour qu'au court d'un bégaiement de répétition d'amorce, il y ait une sensation de régularité, mais suffisamment importante pour que lors d'une simulation de bégaiement, une machine ne puisse pas simplement faire une boucle identique contenant un intervalle immuable de 99ms. On peut d'ailleurs rappeler le chiffre de 12ms, qui est la limite temporelle de perception d'une latence. En se servant de ce chiffre on pourrait alors se baser sur des variations minimums de 12ms entre chaque répétition.

La dernière question concerne le silence, lors d'un blocage tension par exemple, qui pour un programme induirait un arrêt, une impossibilité de calcul plus qu'une tension corporelle visible. Le silence est à l'image du blocage d'un bègue au téléphone, il n'y a pas de corps pour véhiculer l'information. Il est impossible de contourner cet état de fait. Le seul moyen de comprendre le silence d'un bègue sans son corps, c'est d'avoir au préalable saisi qu'il était bègue. Pour simuler le silence, il faudra donc avoir au préalable transmis une série d'autres troubles qui aiderons à la compréhension du dit silence.

On peut donc maintenant supposer une série de règles pour modéliser le bégaiement dans la synthèse vocale :

- Pas d'erreurs uniques
- Possibilité d'organisation arythmique des erreurs.
- Pas de répétitions exactement régulières. Toute répétition doit impliquer une légère variation, à la limite de la perception, au moins 12 ms.
- Il faut dépasser les seuils de disfluences pour entrer dans la zone du trouble du langage.
- Il ne faut simuler un blocage-tension impliquant du silence uniquement qu'après avoir simulé des troubles significatifs. Cette simulation significative se fera dans une période de test.

4. Conclusion sur la voix naturelle

En suivant la voix, de sa production mécanique à son impact à la fois pour l'émetteur et le récepteur, nous avons pu dresser une série de règles permettant de baliser et de comprendre sa forme spontanée, vecteur de l'identité vocale.
Grâce à l'établissement d'une série de descripteurs et de zones d'incarnations, nous avons maintenant des outils pour parler des voix synthétiques avec plus de précisions sans pour autant réduire l'identité vocale à une composante simpliste comme une manifestation esthétique ou une information biométrique.

La disfluence sous-tend une grande part de cette communication spontanée et bien qu'il ne fût pas possible de relier directement sa mécanique à des traits identitaires, nous avons pu établir une série de règles et d'impacts sémantiques et sémiologiques qui nous seront utiles dans leur modélisation.
À travers le bégaiement, nous avons étudié la limite de la disfluence avec le trouble et le témoignage corporelle comme trace d'un dérèglement de l'appareil phonatoire. Grâce à cette étude, il est maintenant possible d'avoir une cartographie des problématiques que nous rencontrerons dans la synthèse des disfluences pouvant mener à la synthèse d'un trouble et de ce que cela peut impliquer dans la perception d'une identité et d'un corps fantomatique.

Après avoir étudié l'histoire de la synthèse de la parole et ses techniques, nous nous concentrerons sur un synthétiseur en particulier, MaryTTS, dans lequel à l'aide du langage SSML, nous nous attèlerons plus tard à synthétiser la disfluence. A l'aide des informations récoltées, nous essaierons alors d'établir des profils identitaires de voix synthétiques en fonction des techniques et des emplois faites de ces voix.

3. La voix synthétique

5. La synthèse de la parole

A travers l'histoire de la synthèse vocale, on pourra distinguer deux catégories de production : la voix parlée et la voix chantée. Ce mémoire a pour objet uniquement la synthèse de la parole. Mais comment la définir ?
Un magnétophone n'est pas un synthétiseur. Pour qu'il y ait synthèse, il faut qu'il y ait une part de création de signaux, mais nous verrons que certains synthétiseurs fonctionnent aussi par l'association de portions d'enregistrements. Il faut définir la synthèse de la parole par son produit final et son message d'entrée.

« L'objectif de la synthèse de la parole est de produire des sons de parole à partir d'une représentation phonétique du message » (Calliope, 1989)

Comment à partir d'un message entrant segmenté (texte, commande) produit-on un signal acoustique continu ? Nous verrons dans un premier temps les étages de créations d'un signal acoustique continu – le synthétiseur – puis nous aborderons à travers le logiciel open-source MaryTTS, les étapes qui mènent d'un message à une commande pour le synthétiseur. Mais abordons d'abord la synthèse de la parole d'un point de vue historique.

a. Un peu d'histoire

La première machine à générer du langage articulé a été mise au point en 1791 par le baron Von Kempelen. C'était une machine imitant l'appareil phonatoire en utilisant un soufflet, un petit tuyau d'orgue et des matériaux résonants[24]. Il n'était possible de synthétiser uniquement que des suites de phonèmes très simples comme « papa » sous couvert d'un long entraînement pour manipuler machine.

[24] https://www.youtube.com/watch?v=zYRVqrfY3tQ

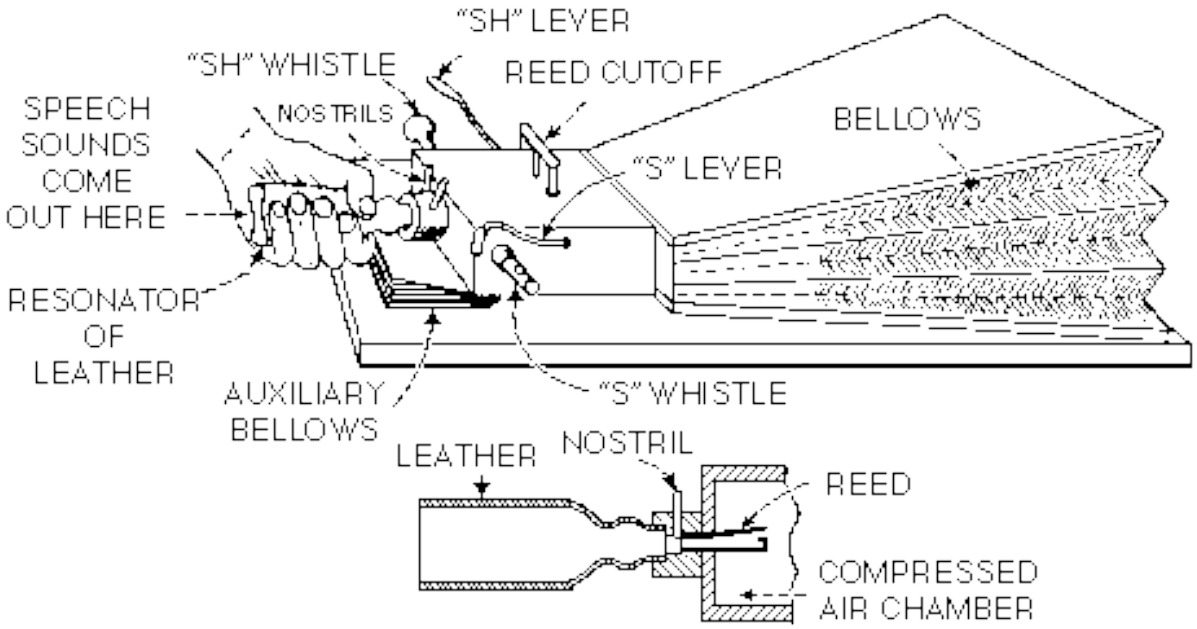

Figure 14 Schéma du fonctionnement de la première machine à synthèse de la parole

Les machines qui suivirent furent des améliorations de cette modélisation acoustique. Il faudra attendre l'électricité, puis l'électronique pour qu'en 1939 naisse dans les laboratoires Bell, le VODER[25]. C'est un appareil muni de deux claviers commandant deux oscillateurs électriques produisant du bruit blanc attribué à la respiration, et du signal périodique attribué à la vibration des cordes vocales. Muni de 10 touches correspondant à des valeurs de filtrages servant à générer des phonèmes et une pédale contrôlant la fondamentale, il était alors possible de créer des phrases, de gérer leur prosodie, de générer des phrases impératives, des questions, de créer du silence. C'est la première fois qu'une machine génère sa propre voix. Le son du VODER va alors influencer autant les Arts que le rapport de la société à la machine puisque maintenant cette dernière est capable d'imiter certaines émotions. La machine est toutefois dépendante d'un opérateur humain, dont l'entraînement pouvait prendre une année avant d'avoir une parfaite maitrise du VODER.

[25] Voice Operation Demonstrator : https://www.youtube.com/watch?v=0rAyrmm7vv0

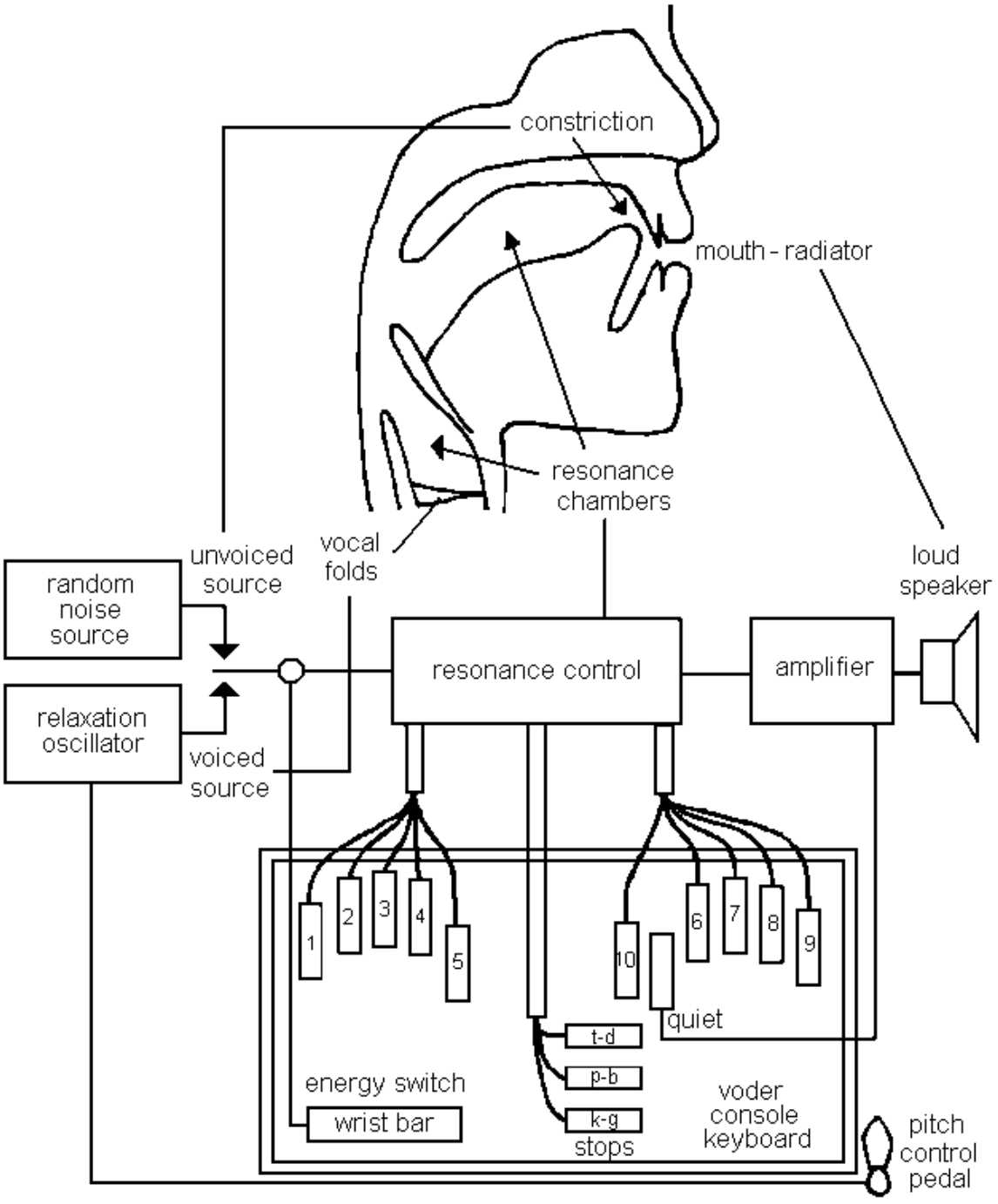

Figure 15 Diagramme du voder

Une grande série de recherches voient le jour à la suite du VODER et mettent en évidence l'importance des transitions dans l'enchaînement des phonèmes. A la suite on développera les premiers synthétiseurs de formants[26] (1953). Dans les années 70, les théories du traitement du signal numérique permettront de perfectionner deux types de modèles qui continuent leur chemin en parallèle :

[26] « On désigne par formant (acoustique) d'un son de parole l'un des maxima d'énergie du spectre sonore de ce son de parole. Il y a plusieurs définitions du mot "formant" (résonances du conduit vocal, pôles, etc.). » Wikipedia

- La synthèse de la parole à travers des modèles de filtre et de signaux (le VODER)
- La modélisation articulatoire d'une onde traversant le conduit vocal (la machine du baron Von Kempenen)

Chacun des deux modèles contient une série de technologies qui évoluent vers une imitation toujours plus réussie. Nous nous concentrerons dans ce mémoire sur le premier type de synthèse, technologie employée dans MaryTTS.

b. Les techniques

Il existe deux types de techniques qui sont relativement liées à l'espace mémoire des machines qui les embarquent : la synthèse « par modèle » et la synthèse « direct ».
La première va, à l'image du VODER imiter le signal de la voix par des procédés de filtrages. Bien qu'impliquant une faible consommation de mémoire, les timbres créés sont peu fidèles à la voix humaine.

i. Synthèse à travers un modèle

1. Synthétiseur à formant ou synthèse par règles

L'essentielle de l'information de la voix voisée est portée par les formants, il est donc logique, dans l'idée d'une synthèse efficace, d'avoir réduit la voix au codage de ces seuls formants. Une série de règles permettant de synthétiser tous les mouvements de phonèmes au cours du discours ont alors été établies. En définissant seulement trois formants, leur largeur de bande, leur amplitude et leur fréquence, il est alors possible en utilisant des filtres simulant l'effet de cavités supra-glottiques, de produire une partie du spectre de la voix parlée. En l'associant à des bandes de bruits générés aléatoirement, on obtient alors les sonorités fricatives manquantes. La complexité varie en fonction des synthétiseurs, dont certains embarquent des parties de synthèses granulaires ajoutant des paramètres.

La synthèse par formants se fait à l'origine à l'aide de plusieurs F.O.F (Fonction d'Onde Formantique) qui sont des sinus dits synchrones pour la partie voisée et asynchrone pour la partie fricative, modulés par une enveloppe générée par deux fonctions (cosinus et exponentiel). Plus le nombre de F.O.F créée augmente, plus la synthèse est complexe et le nombre de paramètres de modulation est grand.

2. *Synthétiseur à prédiction linéaire*

La synthèse par prédiction linéaire décrit un échantillon sonore comme une combinaison linéaire des échantillons précédents. C'est un modèle nécessitant un algorithme de 10 à 15 coefficients avant d'atteindre une qualité acceptable mais qui trouve ses limites dans la difficulté à gérer certains sons (les nasales, les constrictives voisées…) et la qualité générale du timbre obtenu.

Une autre solution était donc attendue. La synthèse directe fit son apparition avec l'augmentation de l'espace de stockage et de la mémoire vive. Cette technique implique une bibliothèque d'échantillons de voix préenregistrées donc une plus grande capacité mémorielle. Toutefois les timbres générés sont plus élaborés et fidèles à la voix humaine.

ii. Synthèse Directe

1. *Synthétiseur par concaténation d'unité*

La synthèse par concaténation d'unité consiste en la lecture d'une série d'échantillons sonores. Plus les échantillons sont nombreux et longs, plus la synthèse sera de bonne qualité. L'exemple le plus simple est la synthèse mot à mot de l'horloge parlante.

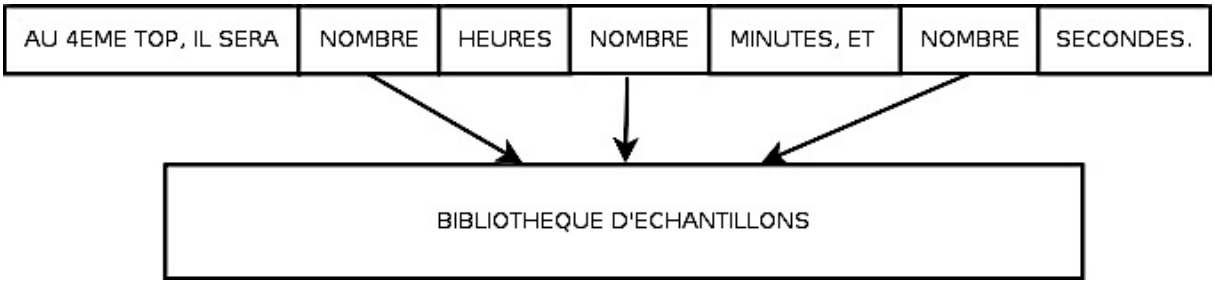

Figure 16 Schéma de la synthèse mot à mot de l'horloge parlante

C'est une synthèse sans aucune flexibilité ni naturelle dut à la grosseur des « grains » ou échantillons avec lesquels elle s'articule. La concaténation d'unité va alors se préciser pour descendre jusqu'au niveau du phonème avant de se confronter aux problèmes de coarticulation qui se posent par un enchaînement parfois impossible à réaliser. L'unité diphone fut alors définie : la transition entre deux phonèmes successifs, c'est-à-dire la portion du signal de parole comprise entre les noyaux stables de deux phonèmes consécutifs (Florent Xavier, 2003).

La synthèse par diphone fut la première synthèse concaténative sur le marché car elle ne nécessitait pas un stockage trop important (de deux à six minutes de paroles) pour un résultat fidèle. Plus l'enregistrement dure, plus la rencontre d'un même diphone dans

conditions différentes se multiplie. On augmente alors la possibilité du synthétiseur de fournir des modulations prosodiques.

Pour créer un synthétiseur à synthèse concaténative, il faut un texte bien défini et l'enregistrement d'une voix lisant ce texte. Il faut ensuite concaténer à différentes échelles – phrase, groupe modal, mot, syllabe, demi-syllabe, diphone, phonème – puis aligner chaque élément avec le texte correspondant ainsi que des informations acoustiques (pitch, durée …). Lors de la saisie d'un texte, il sera ensuite, de la même manière, décomposé à différentes échelles afin de créer un signal continu respectant le sens sémantique du message d'entré.

Il existe une possibilité entre la synthèse du signal pur (synthèse formantique) et la concaténation précise : la synthèse MMC. C'est le synthétiseur que nous utiliserons dans MaryTTS. Voyons plus en détail le fonctionnement de ce logiciel Open-Source en partant de la stratégie de synthèse jusqu'au message entrant.

c. MaryTTS

MaryTTS est un logiciel de synthèse de la parole à partir du texte ; TTS signifiant Text To Speech (Du texte à la parole). Cette solution open-source a été développée par le laboratoire de technologie du langage de DFKI et l'institut de phonétique de l'université de Saarland. Multi-plateforme (Linux, Windows, OsX, Solaris), multi-langue (Anglais, Français, Allemand, Tibétain, Italien, Russe …) et accessible grâce à son développement en java, sont autant d'éléments qui m'ont poussé à le sélectionner pour ce mémoire parmi toutes les autres solutions de synthèse.

Le logiciel se compose d'une série de modules permettant d'aller d'une saisie textuelle dans un grand nombre de langues à un signal acoustique dont les résultats qualitatifs varient en fonction des paramètres utilisés.

Voici un schéma général de l'architecture fonctionnelle du logiciel.

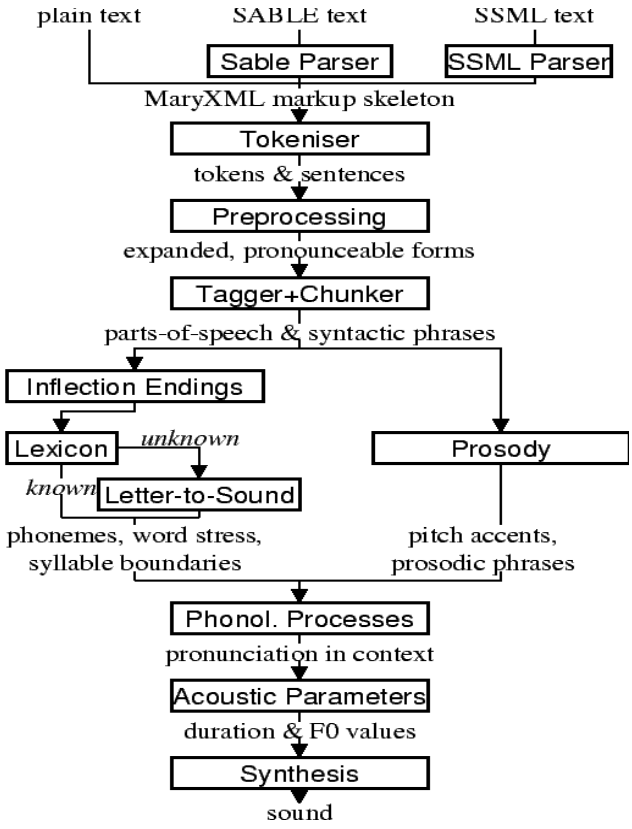

Figure 17 Architecture du logiciel MaryTTS issu du site de DFKI

On observe la possibilité de fournir plusieurs types de signaux d'entrées :

- « plain text » qui sera une entrée de texte normal.
- « SABLE text » qui est un langage XML adapté à la synthèse vocale. Les initiales de SABLE se réfèrent aux sociétés ayant participé à l'élaboration de ce langage : Sun Microsystem, AT&T, Bells Labs, Edinburgh University
- « SSML Text » ou Speech Synthesis Markup Language est le langage XML actuel pour la notation de texte destiné à la synthèse vocale.

Nous aborderons ces types d'entrées plus tard, une fois le fonctionnement général assimilé. Commençons par l'étage final : le synthétiseur.

i. La synthèse MMC avec HTS

HTS est un synthétiseur développé en 2002 par le Nagoya institute. Pour son bon fonctionnement, il faut passer par deux phases

1. La phase d'entraînement

C'est la phase qui comme dans la synthèse concaténative, va passer d'un enregistrement vocal à une bibliothèque d'échantillons. Il y a donc segmentation à différentes échelles puis extractions de données d'excitations et de spectres. La différence est qu'ici, pour pallier au manque de matière enregistrée, il va y avoir création d'arbres décisionnels pour les informations d'excitations et de spectres. Ces arbres sont établis à l'aide de Modèle de Markov Caché. Les modèles de Markov sont des algorithmes de probabilités permettant de

définir une séquence de transitions pour une quantité N d'états. La notion cachée implique qu'il est impossible de remonter aux états initiaux à partir de la séquence générée, car la séquence repose sur un fonctionnement aléatoire. Ces modèles permettent alors de réduire un élément figé (ici une voix) à un système de probabilité automatique qui fournit des séquences non redondantes.

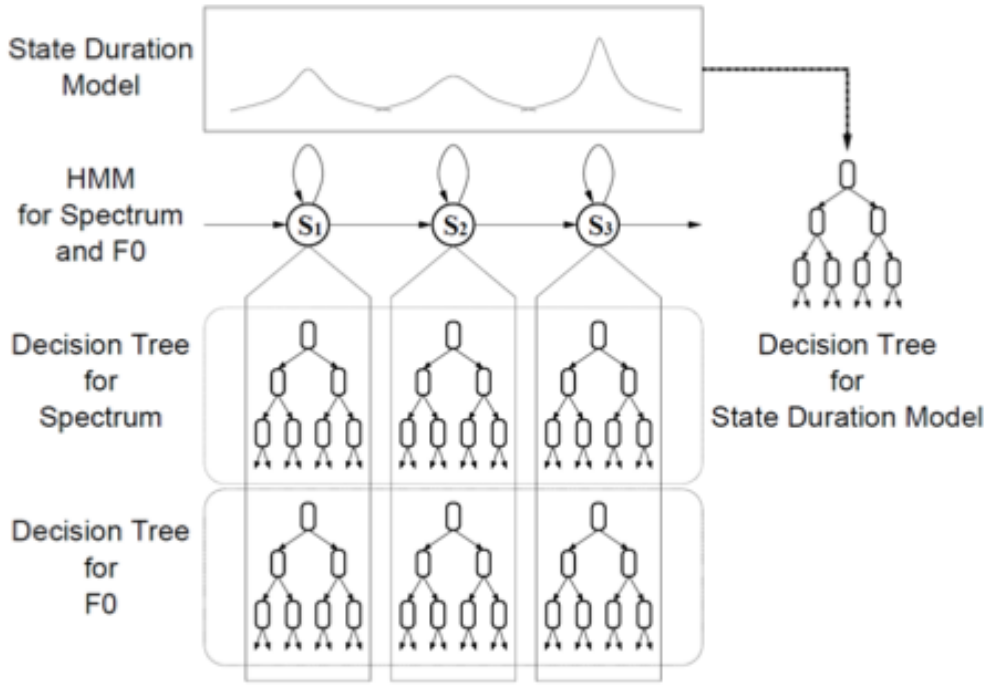

Fig I.10. *Arbre décisionnel pour le regroupement contextuel* (34)

Figure 18 Schéma de la phase d'entraînement de la synthèse MMC issu du mémoire de Xavier Florent, 2003

Ces arbres vont permettre une plus grande flexibilité lors de la synthèse en multipliant les possibilités et l'exactitude de la composition de l'excitation et du spectre d'un phonème attendu.

2. *La synthèse*

La synthèse va avoir pour but de sélectionner les paramètres les plus judicieux selon les arbres générés par le modèle avant de synthétiser la forme d'onde à partir de systèmes tels que celui-ci :

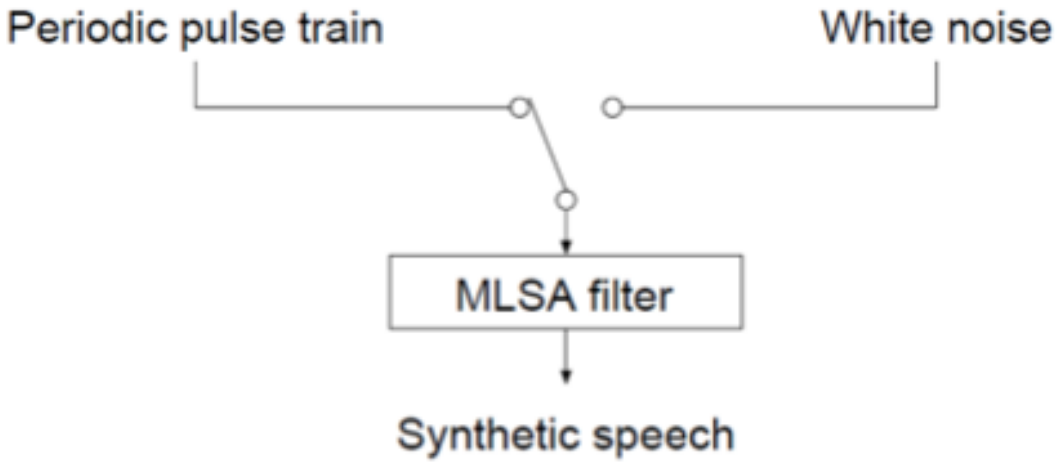

Figure 19 Module de synthèse de la forme d'onde

Ce dernier reprend le fonctionnement duel de la synthèse par modèle associant un son périodique contenant l'information dite voisée à un bruit blanc contenant les informations articulatoires et de constrictions. Les informations acoustiques, de filtrages, la prosodie et les phonèmes à coder vont être extraits par une série de modules qui font partie du TALN[27] ou NLP[28] en anglo-saxon. Nous allons décortiquer chaque étape de cette mission précise qui conduit un texte tapé par un utilisateur à des informations compréhensibles par un synthétiseur MMC.

ii. Les étapes du TALN

Chaque opération consiste en une segmentation de plus en plus petite et précise pour conduire à une traduction en phonèmes et à l'extraction de paramètres acoustiques. Nous prendrons comme exemple la phrase : Je suis 1 Robot, et je ne veux pas « mourir ».

1. Tokenizer

Le Tokenizer est un module de segmentation qui va séparer chaque mot, mots-composé, ponctuation et chiffre pour créer des unités indépendantes. Ce découpage va permettre de déterminer le fonctionnement grammatical du texte et de chacune de ces propositions à destination des autres modules. Il permet de gérer les ponctuations telles que

[27] Traitement automatique du langage naturel
[28] Natural Language Processing

les guillemets, les retours à la ligne etc. ... Après le module Tokenizer, le texte entré ressemblera à cela :

Je	suis	1	Robot	,	et	je	ne	veux	pas	«	mourir	»	.

2. Preprocessing

La phase de Preprocessing est une phase de mise en forme et de nettoyage permettant de limiter les ambiguïtés du texte. Les premières étapes sont la suppression de la casse du texte par la mise en minuscule de toutes les unités, la réaccentuation en fonction d'un dictionnaire interne, le remplacement des chiffres par leur équivalent littéraire ainsi que la gestion des abréviations.

Après l'étape de Preprocessing, notre phrase ressemblera à cela :

je	suis	un	robot	,	et	je	ne	veux	pas	«	mourir	»	.

3. Part-Of-Speech-Tagger

La traduction française du rôle de ce module porte le nom barbare d'étiquetage morpho-syntaxique. En d'autres termes, c'est le module qui va attribuer le rôle syntaxique supposé de chacune des unités. Il se joue à cet endroit trois choses :

- C'est ici que l'on va pouvoir déterminer des éléments de prosodie grâce à l'orientation syntaxique et à la ponctuation.
- C'est ici que l'on prend en compte la prononciation des liaisons ou non. La règle générale pour le français sera de prononcer la liaison si la première lettre du mot suivant celui concerné est une voyelle. Il faudra toutefois prendre en compte certaines liaisons interdites ou l'élision du schwa plus communément appelé « e » muet ou « e » caduque. Le dictionnaire de règles définit donc aussi une orientation identitaire. En effet, certaines règles ne s'appliquent pas de la même manière dans toutes les régions de France (par exemple la prononciation des « S » dans le sud-ouest de la France ou de certain « e » muet). Il existe donc différents modules de Part-Of-Speech-Tagger qui permettent d'influer sur l'orientation morpho-syntaxique et donc sur l'identité de la voix synthétisée.
- Et enfin, c'est ici que l'on va déterminer la prononciation de certains homographes hétérophones, ces mots dont l'écriture est identique mais la

prononciation différente. Par exemple : « Les poissonniers évident les sardines, évident non ? » ou « mon fils dénude les fils de cuivre » .

De par la position du mot, il est possible de savoir si c'est un verbe, un nom ou un adjectif... et donc en déduire sa prononciation.

je	suis	un	robot	,	et	je	ne	veux	pas	«	mourir	»	.
Suj	Ver	Det	Nom	Ponct	Int	Suj	Neg	Ver	Neg	Ponct	Verb Inf	Ponct	Ponct

4. Phonemizer

Nous arrivons à la phonétisation. Chaque mot étiqueté va être transformé en une série de phonèmes selon les règles contenues par un lexique et l'étiquetage morpho-syntaxique fait au préalable. Si ce mot n'est pas contenu dans le lexique, une traduction « lettre à son » est appliquée.

Il reste alors deux éléments constitutifs de la synthèse qui sont déterminés lors de l'étape de phonétisation : la prosodie et les paramètres acoustiques.

5. Le générateur de prosodie (Prosody Generic + Prononciation Model)

Nous avons vu que les informations prosodiques concernaient autant un rapport individualiste et un style personnel, qu'une inflexion informative permettant de déterminer si une phrase est une question ou une affirmation.

Il existe quatre fonctions prosodiques :

- La fonction distinctive qui permet de déterminer si une phrase est interrogative, impérative, déclarative, ou tout autre type d'énoncé.
- La fonction démarcative qui permet de marquer la séparation entre les mots pour bien les distinguer
- La fonction « focalisatrice » (Xavier Florent, 2010) qui permet de créer des accents et d'installer une emphase, d'appuyer sur une quantité d'informations. Cette fonction difficile à définir automatiquement relève d'habitude linguistique relative à une langue, un pays, une région, une ville et un individu. L'étiquette morpho-syntaxique permet d'obtenir des indices comme la position du verbe, la hiérarchie des propositions par rapport à la ponctuation.

- La dernière fonction relève de la prosodie individuelle, le style personnel, la chanson, l'accent. C'est la fonction « expressive ». Le rapport émotionnel n'est pas en corrélation directe avec la saisie d'un texte, il n'est pas pour le moment possible de synthétiser de telles informations avec un synthétiseur TTS. Nous pouvons toutefois les simuler de manière arbitraire mais nous verrons plus tard par quel moyen.

Les informations prosodiques sont donc générées selon les étiquettes morpho-syntaxiques, les phonèmes générés et les intentions de ponctuation. Elles sont ensuite agencées sous la forme de balise comme suivant :

```
<prosody pitch="+5%" range="+20%">
<phrase>
<t accent="L+H*" g2p_method="lexicon" ph="Z @" pos="content">
je
</t>
<t accent="L+H*" g2p_method="lexicon" ph="' s H i" pos="content">
suis
</t>
<t pos="content">
un
</t>
<t accent="L+H*" g2p_method="lexicon" ph="' R 0 - b o" pos="content">
robot
</t>
<t pos="$PUNCT">
,
</t>
<boundary breakindex="4" tone="H-L%"/>
</phrase>
</prosody>
<prosody pitch="-5%" range="-20%">
<phrase>
<t accent="L+H*" g2p_method="lexicon" ph="' e" pos="content">
et
</t>
<t accent="L+H*" g2p_method="lexicon" ph="Z @" pos="content">
je
</t>
<t accent="L+H*" g2p_method="lexicon" ph="n @" pos="content">
ne
</t>
<t accent="L+H*" g2p_method="lexicon" ph="v @" pos="content">
veux
</t>
<t accent="L+H*" g2p_method="lexicon" ph="' p a" pos="content">
pas
</t>
<t pos="content">
"
```

Figure 20 Extrait de la sortie du module Prosody Generic de Mary TTS

On peut observer dans une ligne la présence d'une l'étiquette très peu variée : « content » ou « punct ». On constate la présence de la traduction en phonème (ph="v @") mais aussi d'une information d'accentuation (<t accent="!H*") ainsi que des balises de « prosody pitch » (<prosody pitch="-5%" range="-20%">) Nous analyserons toutes ces subtilités de codage dans un deuxième temps. Pour le moment nous voyons qu'il y a, à cette étape, génération de marqueurs décisifs dans la synthèse de la prosodie.

6. AcoustiModeller

Le dernier module va traduire l'ensemble de ces étiquettes, marqueurs et balises en paramètres acoustiques, c'est à dire une ligne d'intonation de fondamentale (f0) et des pauses. Pour ce faire, on génère puis sélectionne des pairs de phonèmes légitimes à partir de la bibliothèque d'échantillons à l'aide de Chaîne de Markov ou d'un arbre décisionnel (CART).

En voici un exemple rapide pour les deux mots « je suis » :

```
        prosody pitch="+5%" range="+20%">
        <phrase>
        <t accent="L+H*" g2p_method="lexicon" ph="Z @" pos="content">
        Je
        <syllable ph="Z @">
        <ph d="80" end="0.08" f0="(12,140)(25,137)(37,134)(50,141)(62,143)(75,147)(87,150)(100,153)"
p="Z"/>
        <ph d="50" end="0.13"
f0="(10,160)(20,160)(30,163)(40,164)(50,164)(60,164)(70,163)(80,158)(90,158)(100,157)" p="@"/>
        </syllable>
        </t>
        <t accent="L+H*" g2p_method="lexicon" ph="' s H i" pos="content">
        suis
        <syllable accent="L+H*" ph="s H i" stress="1">
```

Figure 21 Exemple de la sortie du module Acoustic Modeller de MaryTTS

On observe des couples de nombres entre parenthèse qui constituent des modulations de fondamentale en fonction du temps.

Le synthétiseur a maintenant toutes les informations pour générer un signal continu et respectant le texte saisi en entrée. Voici un document récapitulatif issu du rapport de stage de Florent Xavier (2010) :

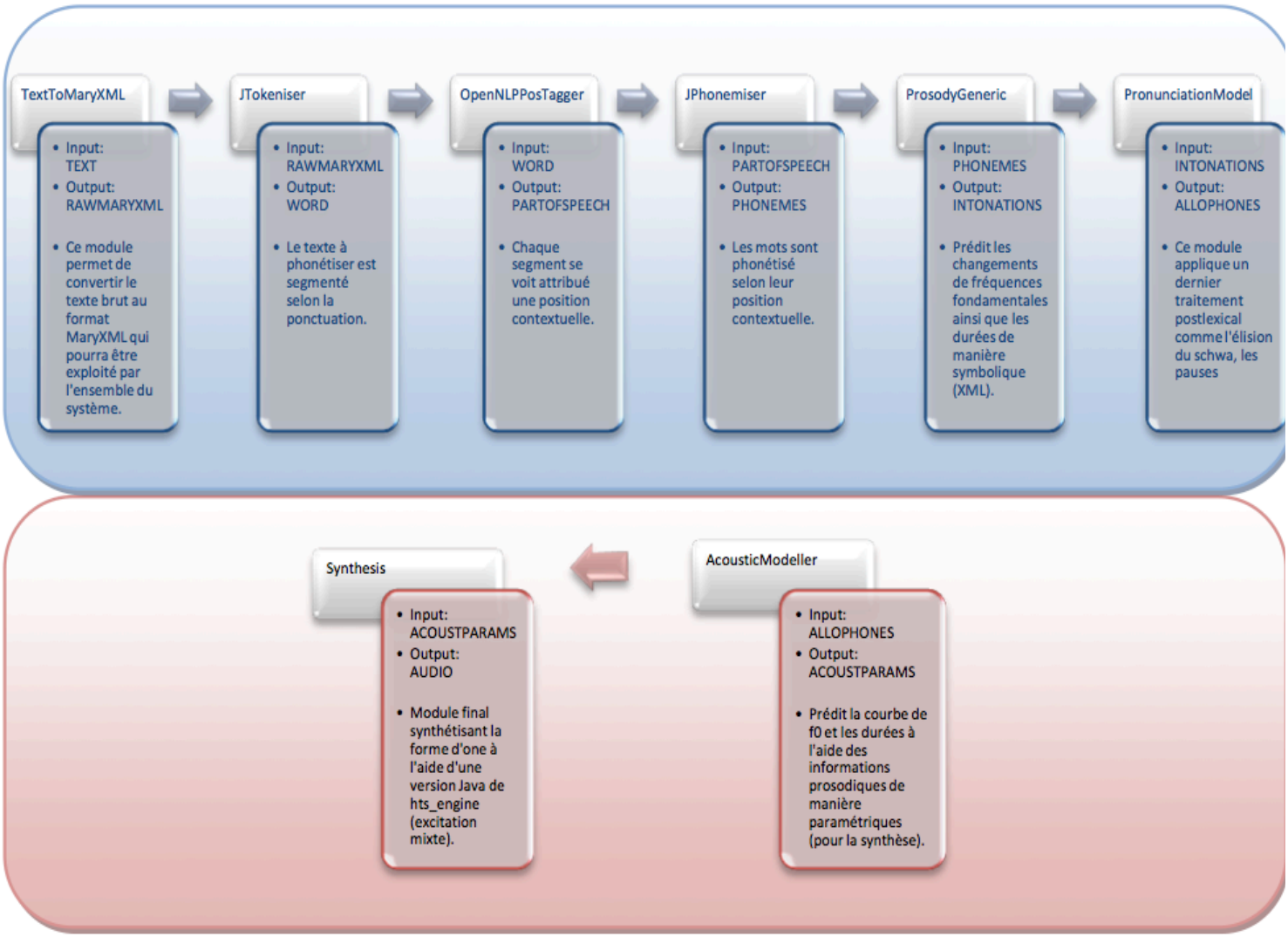

Figure 22 Modules de Mary TTS (Florent Xavier, 2010) La partie bleu correspond à la mise en forme de caractères. La partie rouge correspond à la synthèse sonore.

Maintenant que nous avons abordé l'ensemble du fonctionnement de MaryTTS, revenons au langage utilisé dans ces modules. Ces balises de cette forme <type = 'information'> sont des balises XML. Nous allons nous concentrer sur le SSML, ce fameux langage XML dédié à la synthèse vocale.

6. LE SSML

Le XML (eXtensible Markup Language) est un langage de balisage générique permettant de créer des zones de définition précises au sein de contenus destinés à plusieurs systèmes qui ne sont pas nécessairement compatibles. Par l'utilisation de chevrons « < >», il est alors possible de préciser un contenu, de lui donner une destination, des limites ou de préciser certaines de ses caractéristiques.

Le SSML (Speech Synthesis Markup Language) en est un dérivé formel adapté à la synthèse de la parole, c'est à dire qu'il permet de formater un texte à l'aide de balises renseignant sur des caractéristiques dudit texte, à destination d'un synthétiseur. Ce langage a été intensément développé de 1999 à 2004 à l'intention d'applications web utilisant la synthèse de la parole, à l'initiative du Voice Browser Working Group du W3C. Son développement devait permettre un contrôle des synthétiseurs sur la plupart des plateformes et la plupart des implémentations de software. Il devait être facilement déchiffrable par un humain et permettre un fonctionnement dans n'importe quelle langue. Nous verrons que tous ces points sont respectés. (http://www.w3.org/TR/speech-synthesis/).

Une fois son fonctionnement appréhendé, nous listerons les balises disponibles puis vérifierons leur fonctionnement au sein de MaryTTS. En dernier lieu, nous établirons les liens potentiels entre certaines balises et certaines disfluences.

a. Le XML dans MaryTTS

Voici un extrait de texte codé en SSML.

```
<?xml version="1.0"?> <speak xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:dc="http://purl.org/dc/elements/1.1/"version="1.0">
        <metadata>
        <dc:title xml:lang="en">Telephone Menu: Level 1</dc:title>
        </metadata>
<p>
        <s xml:lang="en-US">
        <voice name="David" gender="male" age="25"> For English, press
<emphasis>one</emphasis>.
        </voice>
        </s>
        <s xml:lang="es-MX">
        <voice name="Miguel" gender="male" age="25"> Para español, oprima el
<emphasis>dos</emphasis>.
</voice>
</s>
</p>
```

Figure 23 Extrait de langage SSML

Les chevrons délimitent donc deux types des balises : certaines encadrent du texte, fonctionnant comme des bornes. D'autres semblent contenir des informations, fonctionnant comme des marqueurs. D'une manière générale, les balises fonctionnent de cette manière : **<Sujet attribut=** "**caractéristique**">

Ici la balise <voice> possède plusieurs attributs : gender, age et name. A l'intérieur de cette balise, on accède à ces attributs, il devient alors possible de les définir : name = miguel, gender = male, age=25. Il est alors possible de transmettre des informations indiquant quelle type de synthèse il faut appliquer à cette zone de texte : pour synthétiser "Para español, oprima el dos" le synthétiseur devra s'orienter vers une voix de jeune adulte masculin. Il existe un grand nombre de balise dont la description est disponible en détail sur http://www.w3.org/TR/speech-synthesis/ . Nous en analyserons les principales, mais d'abord analysons la relation entre le SSML et le synthétiseur MaryTTS.

La première chose à constater est que la sortie du module de l'acousticModeller de MaryTTS (voir chapitre précédent) ressemble de très prês à du langage SSML. On peut alors supposer que le SSML est le codage transmis au module de synthèse de MaryTTS ?

Pas tout à fait à vrai dire. MaryTTS fournit à son module de synthèse du code XML qui possède les mêmes caractéristiques que le SSML, sans pour autant contenir les mêmes balises et possibilités. Toute l'arborescence XML de MaryTTS est disponible ici :

http://mary.dfki.de/lib/MaryXML.xsd/view

A chaque sortie de module, un code XML est généré, découpant lentement le texte en une arborescence significative pour le synthétiseur. Nous utiliserons comme exemple la phrase : « Je suis. Un robot ».

- Tokenizer :

A la sortie du Tokenizer, deux étapes ont eu lieu :

1/ la mise en place d'une balise de métadata renseignant la version XML employée, où la trouver, l'encodage du texte et quelle langue est employée.

2/ La mise en place de balise <s> </s> (s pour sentence) qui encadre une phrase, <t></t> pour chaque unité de cette phrase (t pour token) et la balise qui encadre les paragraphes.

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="0.5" xml:lang="fr">
<p>
        <s>
        <t>
        Je
        </t>
        <t>
        suis
        </t>
        <t>
        .
        </t>
        </s>

        <t>
        Un
        </t>
        <t>
        robot
        </t>
        <t>
        .
        </t>

</p>
</maryxml>
```

Figure 24 Exemple de la génération de MaryXML à la sortie du Tokenizer

- Part-Of-Speech Tagger :

Durant cette étape, chaque « token » va être étiqueté en fonction de sa position morphosyntaxique. On va utiliser l'attribut « pos » dans la balise <t>, puis on indiquera l'étiquette concernant le token encadré : <t pos= « content »>

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" version="0.5" xml:lang="fr">
<p>
        <s>
        <t pos="content">
        Je
        </t>
        <t pos="content">
        suis
        </t>
        <t pos="$PUNCT">
        .
        </t>
        </s>
```

```
        <t pos="content">
        Un
        </t>
        <t pos="content">
        robot
        </t>
        <t pos="$PUNCT">
        .
        </t>

</p>
</maryxml>
```

Figure 25 Exemple de la génération de MaryXML à la sortie du PartOfSpeechTagger

- Phonemizer :

Durant cette étape, chaque token va être phonémisé. On utilise ici deux attributs : « g2p_method » (g2p pour Grapheme To Phoneme) qui est la méthode de phonémisation. Il en existe plusieurs mais celle que nous utiliserons est « lexicon », c'est à dire une phonémisation à partir d'un lexique préétabli. On pourrait utiliser les attributs userdict, coumpound ou rules, qui permettent la création de méthodes de phonémisation personnalisées.

« ph » est l'attribut qui prend en caractéristique la traduction du graphème en phonème selon la méthode définit en g2p.

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" version="0.5" xml:lang="fr">
<p>
        <s>
        <t g2p_method="lexicon" ph="Z @" pos="content">
        Je
        </t>
        <t g2p_method="lexicon" ph="' s H i" pos="content">
        suis
        </t>
        <t pos="$PUNCT">
        .
        </t>
        </s>
        <s>
        <t g2p_method="lexicon" ph="' 9~" pos="content">
        Un
        </t>
        <t g2p_method="lexicon" ph="' R 0 - b o" pos="content">
        robot
        </t>
        <t pos="$PUNCT">
        .
        </t>
        </s>
```

Figure 26 Exemple de génération de MaryXML en sortie du Phonemizer

```
</p>
</maryxml>
```

- ProsodyGeneric :

Durant cette étape, trois nouvelles balises font leur apparition : <phrase>, <boundary> et <prosody>. De plus à l'intérieur de la balise <t> apparaît l'attribut « accent ». Tous ces éléments permettent de gérer la prosodie, et donc une corrélation entre un rythme d'énonciation de phonèmes et une variation de la fondamentale f0.

Abordons l'attribut « accent » en premier lieu qui comporte, selon l'arborescence XML, une grande quantité de caractéristiques possibles. C'est un attribut qui permet de gérer l'accentuation sur certain token.

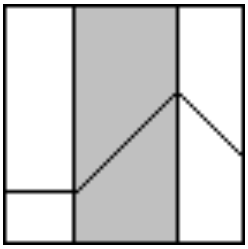

Figure 27
Représentation ToBI de L+H*

Sa notation se fait selon le ToBI (Tones and Break Indice) à l'aide d'un « L » pour Low, « H » pour High. L'utilisation de symbole tel que « * » (qui désigne le registre d'arrivé), « ! » (qui désigne ce que l'on appelle « la faille tonale », abaissement tonal distinctif sur une voyelle), « + » (qui permet d'associer les registres). Il est alors possible de créer des accentuations dans la prosodie, à ne pas confondre avec les accentuations de la phrase qui permettent de distinguer une phrase déclarative d'une phrase interrogative. Ici, l'accent L+H*, dit « scooped accent », donne une orientation concernant la fondamentale au moment de la synthèse.

Il n'est pas nécessaire de connaître les enchaînements[29] car la génération est automatique grâce au module ProsodyGeneric. Nous avons par exemple dans notre exemple généré pour « Je suis » :

```
<t accent="L+H*" g2p_method="lexicon" ph="Z @" pos="content">
Je
</t>
<t accent="!H*" g2p_method="lexicon" ph="' s H i" pos="content">
suis
</t>
```

Figure 28 Exemple de la génération d'accent dans le module ProsodyGeneric

Cela revient donc à partir d'un registre tonal bas pour aller vers le haut sur le « Je », de se maintenir dans le registre haut sur le « su » puis de revenir dans le bas sur le « is » (c'est l'affaissement tonal).

[29] Heureusement, car il existe 34 accentuations de phonèmes possible dans l'arborescence xml de MaryTTS. Cf : http://mary.dfki.de/lib/MaryXML.xsd/view

L'attribut « accent » fonctionne en association avec la balise <boundary>, qui permet de gérer les variations d'intonation sur une phrase entière. Le premier attribut est « breakindex », qui prend en paramètre un chiffre allant de 1 à 6. Le second attribut est « tone ». Il prend en paramètre des notations issu du ToBI, seulement ici c'est « - » qui définit la tonalité de la phrase. Le symbole « % » permet de créer des variations de registre. Par exemple, %L correspond à un départ de la phrase dans le registre bas, L% correspond à une arrivée dans le registre bas. Nous pouvons lister quatre exemples :

- La phrase déclarative : L-L%
- La phrase interrogative : H-H%
- La liste : L-H%
- Tonalité basse : L-

La dernière balise, extrêmement riche, est la balise <prosody>. Comme son nom l'indique, elle comporte une série d'attributs qui affectent la prosodie des éléments qu'elle encadre :

- Pitch : affecte le pitch, selon un argument en pourcentage. C'est donc un changement relatif.
- Pitch dynamics : permet de limiter les variations de pitch déterminées par les accents ou les intonations de <boundary tone>
- Range : Affecte l'ambitus de la voix à synthétisée
- Range dynamics

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml  xmlns="http://mary.dfki.de/2002/MaryXML"  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="0.5" xml:lang="fr">
<p>
<s>
<prosody pitch="+5%" range="+20%">
<phrase>
<t accent="L+H*" g2p_method="lexicon" ph="Z @" pos="content">
Je
</t>
<t accent=" H*+L" g2p_method="lexicon" ph="' s H i" pos="content">
suis
</t>
<t accent="L*" g2p_method="lexicon" ph="' 9~" pos="content">
un
</t>
<t accent="L+H*" g2p_method="rules" ph="' R 0" pos="content">
ro
</t>
<t pos="$PUNCT">
,
</t>
<boundary breakindex="4" tone="L-H%"/>
</phrase>
</prosody>
<prosody pitch="+0%" range="+0%">
<phrase>
<t accent="L+H*" g2p_method="lexicon" ph="' 9~" pos="content">
un
</t>
<t accent="L+H*" g2p_method="rules" ph="' R 0" pos="content">
ro
</t>
<t pos="$PUNCT">
,
</t>
<boundary breakindex="4" tone="L-H%"/>
</phrase>
</prosody>
<prosody pitch="-5%" range="-20%">
<phrase>
<t accent="L+H*" g2p_method="lexicon" ph="' 9~" pos="content">
un
</t>
<t accent="!H*" g2p_method="lexicon" ph="' R 0 - b o" pos="content">
robot
</t>
<boundary breakindex="5" tone="L-L%"/>
</phrase>
</prosody>
</s>
</p>
</maryxml>
```

Figure 29 Totalité du code MaryXML généré par le module ProsodyGeneric

Nous avons donc traversé tous les apports du générateur de prosodie. Il est possible grâce à cet outil, d'envisager la gestion au cours d'une disfluence, par exemple une amorce complétée, la gestion de son intonation. Imaginons : « Je suis un **ro-**, **un ro-,** un robot »
Après les deux amorces, il serait bon de finir dans le registre bas pour maintenir cette phrase comme inachevée, puis après « robot » de finir dans le registre bas pour conclure cette phrase, comme si l'obstacle avait été surmonté.
Il est aussi possible d'imaginer au contraire, finir « robot » dans le registre haut, créant une sensation de découverte du mot enfin prononcé. Il est en tout cas possible de modéliser précisément une grande série de disfluences.

Il subsiste toutefois une variable encore inconnue : le temps. Cette variable apparaît aux étapes suivantes, à savoir la génération de Allophone et de paramètres acoustiques. Ici sont décidées des durées et des pauses. Les token sont transformés en phonèmes et en syllabes, accompagnés d'un attribut « d » correspondant à la durée. Ce mécanisme automatique et complexe à manipuler ne rend pas pratique la manipulation des durées.

C'est à ce moment là qu'entre en jeu le SSML. Rappelons que jusque là nous n'avions qu'étudié le fonctionnement en XML de MaryTTS mais nous n'avons pas encore abordé les balises proposées par le SSML. Ces balises ont pour vocation de simplifier la gestion de la prosodie et de la synthèse en générale.

b. Le SSML dans MaryTTS

Il est possible comme nous l'avons vu, de spécifier l'entrée fournie à MaryTTS. A chacune des étapes il est possible d'intégrer du code SSML, supporté par chacun des modules. Il est important d'indiquer, comme pour le MaryXML, des balises de métadata permettant de trouver la structure du SSML, la langue employée, l'encodage des caractères écrits...

```
<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2001/10/synthesis                    http://www.w3.org/TR/speech-
synthesis/synthesis.xsd"
  xml:lang="fr">
```

Figure 30 Balise metadata du SSML

Il est ensuite possible d'intégrer des balises. Je les ai testées car toutes ne sont pas prises en compte par MaryTTS (à cause de la traduction automatique faite par le module SSMLParser, qui transforme mal certains caractères). La plupart des balises en MaryXML sont similaires en SSML : <p><s><ph><prosody>... Voici celles qui sont fonctionnelles et qui apportent des fonctionnalités supplémentaires utiles dans notre modélisation des disfluences

- **<emphasis>** qui permet de gérer quatre niveaux d'emphase différents qui sont renseignés à travers l'attribut « level » : « none », « reduced », « moderate » et « strong ». L'emphase se traduit par une gestion de l'intonation, de la vitesse de lecture et du pitch. « Reduced » mène à une synthèse dans le registre bas et une lecture accélérée. Cela donne une sensation de discrétion. « Moderate » amène à une lecture à la vitesse par défaut mais à une légère augmentation du pitch. « Strong » amène à une lecture à la vitesse un peu plus lente ainsi qu'à une grande augmentation du pitch.
- **<break>** permet de générer des pauses dans la prosodie. Ce n'est donc pas une balise qui encadre des éléments mais qui fonctionne à l'image d'une ponctuation. Elle prend deux attributs, « strength » et « time ». L'attribut « strength » ne semble par affecter la durée de la pause alors que l'attribut « time » semble, lui, fonctionnel. Il prend en argument une durée en millisecondes ou en secondes :

```
Test de <break time="1500ms"/> prosody
```

- **<prosody>**, que l'on trouve aussi dans MaryXML fonctionne ici avec des attributs légèrement différents : on retrouve **« pitch »** qui prend un nombre exprimé Hz en argument, affectant le pitch des mots à synthétiser encadré par la balise. **« range »** qui correspond à l'ambitus de la voix et qui prend aussi des Hz en argument. **« rate »** qui correspond à la gestion relative du débit de parole, exprimé en pourcentage.

  On découvre l'attribut **« contour »** qui permet de créer des intonations intuitivement en créant des couples (position temporelle exprimée en pourcentage, variation de pitch). On peut alors créer une courbe relative du pitch selon le temps, ou les coordonnées temporelles sont 0%, 10%, 50%.

```
je <prosody contour="(0%,-70Hz) (10%,-30Hz) (+50%,+30Hz) "> test l'attribut, contour</prosody>
```

Deux attributs très utiles ne semblent pas fonctionnels sous MaryTTS :
« duration » qui permet de gérer la durée de synthèse d'un bloc, et « volume » qui permet de gérer le volume de synthèse d'un bloc.
Pour « duration », il est possible d'imaginer des alternatives à l'aide de l'attribut « rate » et de la balise « break ». Pour la gestion du volume, il n'y a pour le moment pas de solution envisagée.
Une autre balise n'est pas fonctionnelle : <audio> qui permet la lecture de fichier audio au format .wav

Une piste pour corriger ces problèmes est ouverte ici : https://github.com/marytts/marytts/issues/167
Que ce soit pour la balise <audio> ou les attributs « volume » ou « duration », il s'agit d'un problème de lecture de l'argument qui leur est transmis. Le module « SSML Parser » n'interprète pas correctement ce qui est entre guillemets, empêchant la bonne transmission des informations au processeur de synthèse. Une partie de ma pratique consiste à corriger ces trois erreurs pour améliorer l'efficacité du logiciel mais surtout pour préciser la modélisation des disfluences (notamment à l'aide de la gestion du volume).

Nous pouvons maintenant dire que nous comprenons une petite partie de l'organe phonatoire des machines, son développement et son fonctionnement. A travers les deux voies entreprises (la simulation du conduit vocal et la modélisation de la voix), on peut voir deux manières d'envisager la voix des machines : une simulation ou une imitation de la voix humaine. La première a pour but de reproduire à la perfection jusqu'à l'impossibilité de discerner, la seconde vise plutôt à gérer un cahier des charges pour produire des sons les plus fidèles possibles tout en ayant conscience des limites du modèle proposé.

Mais la question de l'identité de ces synthétiseurs vocaux a-t-elle était déjà posée ? Nous allons voir que toutes ces technologies ont crée un paysage de voix synthétiques qu'il est intéressant de baliser.

7. Un paysage vocal varié : l'identité des voix synthétiques

Dans les jouets, les barrières, les ascenseurs, les téléphones, les ordinateurs, la musique, les appareils hospitaliers, les voix synthétiques se multiplient. Tout le monde les a entendues et possède une idée relativement précise des caractéristiques de ces voix. Il va s'agir d'explorer et de préciser le souvenir que la communauté a de ces voix. Peut-on définir une identité vocale de voix synthétiques permettant de mieux les cerner, les appréhender et les améliorer ?

Leur apparition dans la sphère publique, dans la communauté puis finalement dans la mémoire, se fait à travers le lien entre un emploi à grande échelle et une limitation technique. Il va donc s'agir de lister et d'étudier d'un côté les apparitions de la voix synthétique dans la société actuelle, puis d'un autre côté les relier aux techniques potentiellement employées pour mettre en œuvre ces voix. Il sera alors possible en raisonnant empiriquement, d'établir des identités vocales en liant une visée fonctionnelle à une technique. En réutilisant les catégories et descripteurs établis en (2.1.c), on pourra alors nommer et discriminer ces identités.

a. Les emplois des voix synthétiques

i. Les annonces publiques

L'une des premières voix citée lors des discussions entretenues à propos de mon mémoire avec des personnes non spécialisées sur le sujet, était Simone : la voix des annonces de la S.N.C.F. C'est une voix marquante car elle reste quasiment inchangée depuis 1981 où elle a commencé à se faire entendre dans toutes les gares de France.

C'est une voix duale, car elle n'est pas à proprement parler « synthétique » puisque ce sont de vrais échantillons de la taille de mots qui sont utilisés. Nous avons toutefois défini cette technique de concaténation comme une synthèse. C'est donc une voix synthétique « avec des gros bouts d'humain à l'intérieur ».

L'intérêt de Simone réside dans la marque mémorielle qu'elle a produit depuis 30 ans sur l'idée que l'on se fait d'une annonce publique. Ces annonces se retrouvent dans les lieux d'affluence où l'urbanisme ne suffit pas à organiser et à transmettre l'information. Elles sont des voix que nous nommerons « Voix de services », puisqu'elles sont en général au service de l'exploitation d'un lieu. Pour gérer des messages d'information répétitifs à l'intention de foules, ou des instructions d'utilisations pour de grands flux de personnes, la solution économique a été de créer des annonces publiques à partir de la synthèse vocale. Ces annonces se retrouvent dans les lieux dédiés aux transports en communs (gare, métro, train

de banlieue ...) et les lieux publics à haute fréquentation (Centre-Commerciaux ou Parc d'attractions ...). Il ne faut pas les confondre avec des annonces enregistrées, qui sont utilisés dans les situations de messages sans variables : par exemple les stations de métro parisien sont enregistrées, à la différence des stations du R.E.R D de la R.A.T.P.

Leur mission fonctionnelle est de transmettre le plus clairement et précisément un message au plus grand nombre de personnes possible : il faut au sein de ce message, identifier les destinataires potentiels, le sujet et les détails informatifs pour que l'auditeur puisse faire le tri. Il y a donc un impératif d'intelligibilité. Cet impératif se traduit à travers plusieurs points importants : une phonation parfaitement intelligible, un langage accessible, une identité sociale non excluante. Le choix lexical ne fait pas directement partie de la synthèse, mais va forger l'identité vocale de ces voix. Les deux autres points sont par contre déterminants dans la gestion de la synthèse et de la prosodie. La synthèse doit être d'une qualité suffisante pour ne pas produire de phonèmes ambigus, compréhensible pour des gens de différentes nationalités ou habitués à toutes sortes d'accents. Ce qui est entendu par « non excluant », c'est que cette voix s'adresse à toute classe sociale, toute origine ethnique, tout âge et tout sexe. Finalement cette voix s'adresse à tout les « clients » de la même manière. Cette notion de client implique toutefois une grande politesse. Un rapport d'intimité n'est pas attendu et serait peut-être même malvenu dans cette mission d'information.

Le rapport change-t-il lorsque les voix synthétiques s'adressent à un individu, de manière privé, dans le même cadre de service ? Nous allons voir qu'il existe plusieurs types d'utilisation privée des voix synthétiques à commencer par deux qui reviennent dans les discussions : le téléphone et le GPS.

ii. Les serveurs téléphoniques

La multiplication de services divers assistés par des automates a été l'un des catalyseurs de l'explosion des voix synthétiques : la distribution à grande échelle et la nécessité de personnaliser des contenus sont les deux facteurs de l'utilisation de la synthèse vocale. Le serveur téléphonique est un outil téléphonique qui permet de recevoir un appel d'un utilisateur, puis d'informer ce dernier ou de l'aiguiller vers un agent humain pouvant l'informer.

Nous sommes toujours dans une logique de service où un exploitant cherche à transmettre de l'information à propos de son produit. Nous ne parlons pas de répondeurs classiques voire même de certains répondeurs améliorés qui ne sont que des annonces interactives que l'on

gère avec le clavier du téléphone. Nous parlons ici de serveurs qui vont, toujours à l'aide du clavier, prendre en entrée des renseignements utilisateurs puis après interrogation de bases de données, synthétiser ces dernières pour les donner à entendre à l'utilisateur. Par exemple, le S.A.V du fournisseur d'accès Free, la consultation téléphonique de son compte en banque au Crédit Agricole...

Il y a plusieurs différences dans les modalités de synthèse par rapport à l'annonce publique : Ici l'auditeur est unique et supposé attentif. L'enjeu d'une transmission parfaite de l'information dès la première phonation est moins grand. Toutefois, ces serveurs entraînent un coût financier (surtout du côté utilisateur) qui suscite une plus grande intransigeance chez l'utilisateur. Il y a donc toujours la nécessité d'une synthèse rapide et efficace.

On peut de plus supposer que, à travers l'habitude téléphonique de l'utilisateur et le rapport unité (humain) à unité (machine) créant un cadre privé, il y ait l'attente d'une discussion ayant les mêmes modalités qu'une discussion orale d'humain à humain. Il ressort d'ailleurs de la plupart des expériences de confrontation avec les serveurs téléphoniques, une déception voire un agacement généralisé de par « un manque de souplesse », « une froideur », « un désintérêt »... Je citerai ici Gérard Pelé, enseignant chercheur de Louis-Lumière, qui, lors d'une discussion posait la question : « Qu'est ce qui fait que je sais aussi rapidement, lorsqu'à l'autre bout du fil j'entends une voix, si c'est une machine ou un humain ? »

Le serveur téléphonique a pour seule mission de remplacer l'agent humain dans un but de profit (gain de temps + réduction de postes = rentabilisation d'une installation technique). C'est donc dans son cahier des charges d'être une imitation plus proche de l'humain. Le serveur téléphonique est donc associé à un cliché populaire : « l'ennemi de l'humain », pâle copie qui dérobe le travaille. A la différence de l'annonce publique, il véhicule une sensation plus désagréable. Il est donc dans l'intérêt des sociétés les employant, d'améliorer cette image. Nous verrons que c'est à travers sa mise en place technique qu'il sera possible de préciser son identité.

iii. Les GPS

Le Global Positioning System (navigation par satellite) est un objet aidant à se diriger dans un véhicule ou à pied. La représentation visuelle ne suffisant pas, il a fallu faire appel à la voix. La synthèse vocale intervient donc pour énoncer l'itinéraire à emprunter. La synthèse permet alors d'indiquer toujours les variations de positions ou les longueurs avant des changements de directions.

La synthèse intervient ici au cours d'une opération qui peut potentiellement présenter un danger : la conduite. Il faut donc que cette voix ne soit pas un élément qui perturbe la conduite. Dans le cas du G.P.S, il y a de nombreuses similarités avec l'annonce publique : l'information doit être la plus intelligible et claire possible afin d'éviter l'ambiguïté, génératrice d'erreurs de conduites.

Le projet de départ du G.P.S n'est pas de transmettre autre chose qu'un itinéraire. Il synthétise des assertions extrêmement rigides et répétitives à travers des syntaxes et des temps invariables (Les changements de direction sont annoncés à intervalles réguliers normalisés eux même par la route et son code, selon la même formulation : « Dans 300 mètres, tourner à droite »). Seulement, en vue d'augmenter son utilisation, il était alors nécessaire de réduire l'irritabilité transmise par la mécanique inexorable de cette voix. Son aspect dispensable a fait qu'une fonctionnalité s'est ajoutée à la phonation de l'itinéraire : le G.P.S doit aussi être un co-pilote fiable. Il accompagne le conducteur sur la route et bien qu'il ne fasse pas la conversation, c'est lui qui rompt la torpeur des autoroutes pour annoncer un changement de direction. C'est en tout cas un argument qui sert à la vente de cet objet dispensable par nature.

Il y a donc un large choix de voix de synthèse dont la qualité suit le prix des offres de G.P.S. Pour la plupart il est en plus possible de choisir et de télécharger la voix désirée comme co-pilote. Certaines de ces voix rompent sans vergogne avec un cahier des charges sécuritaire (il est possible d'utiliser la voix de Bourville, Jacques Chirac ou de Homer Simpson). Le G.P.S est donc à la croisée entre l'assistant de route, et le produit social : celui que l'on montre, qui peut faire vitrine de nos goûts, de notre humour, ou qui nous divertit, qui fait passer la route. Le cas du G.P.S contient déjà la problématique que nous aborderons plus tard au sein de la domotique[30].

iv. Traducteurs/Dictionnaires/Jeux éducatifs

Ces voix synthétiques correspondent à des objets qui appartiennent aujourd'hui à une époque passée allant des années 80 aux années 2000, remplacées par les applications de téléphones portables et programmes d'ordinateurs. Pourtant ces voix-là ont marqué la mémoire collective.

[30] Domotique est un mot correspondant à un mélange entre Domestique et Robotique. C'est un ensemble de technique interopérable entre différents systèmes, son organisation et un lieu prenant en compte la physique du bâtiment, l'électricité, l'électronique, l'informatique, les télécommunications...

Ces objets avaient pour fonction de vocaliser un mot saisi afin d'en apprendre la prononciation. On peut penser à la dictée magique, appareil destiné aux enfants pour apprendre leur langue maternelle, commercialisé par Texas Instrument en 1978.

Figure 31 La dictée magique T.I

Ces objets dont le cahier des charges était d'être portatif et de pouvoir synthétiser tous les mots d'une ou plusieurs langues, n'avaient pas pour vocation de produire de la voix orale spontanée. Ici la transmission d'informations n'a pas besoin d'être idéale. En effet, l'utilisateur se trouve déjà dans une position d'écoute attentive et possède déjà une idée du mot à entendre puisqu'il l'a saisi.

Ces voix étaient donc souvent d'une qualité moyenne. Ils sont donc les ambassadeurs de la voix synthétique bas de gamme, extrêmement modulable en terme de phonème mais extrêmement rigide en terme de timbre ou de prosodie.

v. Les voix de hackers

Un autre emploi répandu des voix synthétiques vient du côté des hackers, des cyber-terroristes et de tout autre personne en marge d'un système (loi, société) ne désirant pas faire connaître son identité vocale mais souhaitant communiquer vocalement.

La voix synthétique devient alors un outil d'anonymat. Par le détournement de la synthèse populaire, l'utilisateur s'assure que son message sera entendu sans être mis en danger. A l'image du ravisseur utilisant des lettres découpées dans des magazines pour ne pas être découvert par son écriture, l'utilisateur va ici détourner un outil commun et populaire pour se fondre dans la masse.

Ici l'attente est donc triple : l'anonymat croisé avec une imagerie populaire et une intelligibilité suffisante pour faire passer un message. C'est en général un message qui balaye un spectre allant de l'angoisse (par exemple, un message à caractère menaçant) à un message d'espoir (par exemple, un message d'alternative politique ou sociale). Cette perception est en générale uniquement portée par le contenu du message puisque le synthétiseur est souvent un objet de consommation dont la facture ne permet pas avec simplicité une grande variation émotionnelle. Ce type de voix devient alors rattaché à des causes dites marginales.

Je prendrais en exemple le groupe modulaire de cyber-activiste, Anonymous, dont les communiqués[31], très diffusés (plusieurs millions de visionnage pour chaque communiqué), utilise le synthétiseur text-to-speech fourni avec le système d'exploitation windows XP.

vi. L'assistanat médical

L'emploi de la synthèse vocale dans la médecine survient lorsque la fonction phonatoire a été supprimée. Il existe un grand nombre de causes possibles menant à un mutisme définitif : accident neurologique, cancer du larynx, paralysie...

La fonction est donc de redonner la possibilité à un patient d'exprimer une idée à l'aide d'une voix. Ici la synthèse joue un rôle complexe puisqu'elle doit exprimer dans toute sa complexité, les pensées, volontés, idées, contradictions et finalement l'identité de son utilisateur. Il faut en plus fournir un système si possible portatif et dont la latence est viable pour le malade. La gestion de la prosodie, des émotions et dans notre cas, des disfluences, sont autant d'arguments qui vont améliorer l'imitation du synthétiseur, mais ralentir sa capacité à synthétiser. Ces technologies sont donc peu démocratisées, d'une part à cause de la difficulté de mise en place mais aussi car il existe un grand nombre de palliatifs à la disparition de la voix : par exemple l'écriture ou le langage des signes.

Pourtant il existe certains utilisateurs dont l'aura populaire a marqué la communauté. Je prendrais en exemple Stephen Hawking[32], scientifique et auteur de vulgarisation scientifique atteint de paralysie sévère qui utilise un synthétiseur vocal pour s'exprimer en public.

[31] https://www.youtube.com/watch?v=JCbKv9yiLiQ

[32] Vidéo d'une interview de Stephen Hawking utilisant son système de synthèse vocale https://www.youtube.com/watch?v=Jzkm1E5RqJo

Ici la voix synthétique est à la fois l'incarnation d'un gouffre entre l'expression humaine naturelle et la synthèse, et en même temps l'opportunité de la retrouver. Elle jouit d'une aura compassionnelle, quelle que soit sa qualité et son caractère, puisqu'elle permet de dépasser un handicap. Elle est à l'image de la prothèse et son objectif est l'intégration par le patient-utilisateur, bien que la technologie ne le permette pas encore.

vii. L'assistanat quotidien ou la domotique

La domotique est un terme associant « Domestique » et « Robotique », qui correspond à un ensemble de techniques liant la vie du quotidien à des techniques comprises entre l'électrique, l'électronique, les télécommunications et la robotique.

Du réglage automatique du thermostat d'une maison à la porte électrique du garage en passant par l'androïde assistant pour les tâches ménagères, le domaine de la domotique peut être scindé en deux : l'exécutant et l'assistant.

L'exécutant correspond à un domaine d'application précis dont la tâche est déterminée, le champ d'action connu et les moyens définis. La synthèse vocale dans ce cas sert à transmettre des informations, en général sous forme de réponses à des requêtes, ou de renseigner l'état de la tâche en cours. Nous prendrons comme exemple les maisons commandées vocalement par un système androïde[33] (androïde étant ici le système d'exploitation développé par google). La synthèse vocale est donc soit un « monitoring » de l'exécution d'une tâche, soit la phonation d'une information relative à une requête. Si par exemple, je réclame l'augmentation de la température de ma maison, une voix synthétique pourra me renseigner sur la tâche en cours et me prévenir une fois que la température souhaitée est atteinte. L'exécutant est purement fonctionnel, à l'image d'un lave vaisselle. Il n'est pas attendu de sa part de développer une identité, et encore moins une sensation d'humanité.

Seulement la mise en réseau de ces technologies et l'infiltration des contrôles à distance au sein de la plupart des appareils font qu'il est maintenant possible de centraliser un réseau de commande et son monitoring. La dissociation d'un thermostat, d'un ordinateur et d'une télé est en train de disparaître pour laisser place à un seul grand flux de commandes et d'actions. Ce grand flux devient alors une entité de contrôle et de renseignement. A l'image d'une décoration d'intérieur, de la forme d'un appareil ou de l'uniforme d'un employé, il semble naturel d'avoir une attente esthétique auprès de ce monitoring vocal.

[33] Vidéo de démonstration d'une maison gérée par contrôle vocale
https://www.youtube.com/watch?v=hYMpMt0lwUY

La vie quotidienne ne répond pas aux mêmes critères que les situations précédentes : le confort permet d'imaginer des systèmes plus amples en taille, poids, ressources. L'attente esthétique peut alors être comblée à l'aide de technologie sans limitation autre que la technologie elle même et les moyens financiers. Ces systèmes-là ne sont pas encore démocratisés à suffisamment grande échelle pour avoir créé une marque mémorielle à grande échelle.

Toutefois la multiplication de système de Voice-Over[34] entraîne une lente accommodation à ces voix synthétiques. Je prendrai pour exemple, l'utilitaire Voice-Over de OsX 10.6.8 qui fournit des fonctions d'aides diverses : aide aux personnes atteintes de cécités, utilisation en lien avec la commande vocale, lecture de contenu pour gagner en efficacité.

Il y a une quantité d'options permettant de choisir une voix, de gérer sa prosodie et finalement, comme à l'image du G.P.S de personnaliser la voix utilisée. Seulement aucune d'elles ne fournit une imitation humaine convaincante. L'enjeu n'est d'ailleurs pas d'imiter mais bien de phonétiser du texte. Dépasser cet enjeu c'est sortir du cadre de l'exécutant et passer au stade d'assistant.

L'un des assistants doté d'une voix synthétique et qui se répand à grande échelle est SIRI, système de reconnaissance et de synthèse vocale développé par Apple. Le cas de SIRI est différent puisqu'il est doté d'applications relatives à l'intelligence artificielle. Il est capable d'entretenir des conversations, de gérer du contenu et de donner l'illusion de prendre des décisions (par exemple, changer de chanson et en proposer une autre). Cette petite quantité d'actions supplémentaires le fait passer du statut d'exécutant à celui d'assistant. Les attentes vocales sont alors autres. On peut supposer que la diminution du rapport fonctionnel et les traits d'intelligence qu'on peut lui prêter, l'autorise à une plus grande largeur prosodique, à une palette d'émotions plus variées. Il est envisageable de la part de SIRI, d'employer un ton humoristique (si tant est qu'il en existe absolument un), alors que ce plus difficile pour un exécutant renseignant la température d'une maison.

On peut donc supposer que c'est au sein des assistants domotiques que l'imitation humaine est attendue comme un argument commercial valable. Cette fonction n'est toutefois pas encore réalisée, ni commercialisée.

[34] Système permettant la synthèse vocale TTS pour personne aveugle, voire libérer les yeux et gagner en efficacité

La question qui s'incarne ici, et qui était sous-jacente depuis le début de cette section est la suivante : Est-ce qu'une voix synthétiques a des raisons d'être disfluente, si elle n'est pas doté d'intelligence ? Est-ce qu'une voix peut développer une oralité qui se construit en temps réel, si elle ne donne pas l'illusion d'avoir du libre arbitre ? Autrement dit, la disfluence dans la synthèse vocale n'est-elle pas directement reliée à la notion d'intelligence artificielle ?

Pourtant, si il est si facile d'imaginer un tel futur, c'est parce que la problématique de la confrontation Humain-Machine, de l'Androïde, et finalement de réflexion transhumaniste[35] est au centre d'une série d'œuvres d'Art, dont l'influence, elle, est immense. C'est dans ce rapport à ces anticipations que sont forgées une part des évolutions technologiques. Il s'agira donc d'étudier la voix de synthèse et sa représentation dans les Arts afin de dresser un portrait fidèle de notre conception des voix de synthèses.

Emploi/But	**Nom**	**Moyen Attendu**
Voix de service informatif ou commercial ayant pour but l'exploitation de lieu à haute fréquentation	Annonce publique	Transmission non excluante, extrêmement précise, d'un message à la variabilité restreinte destiné au plus grand nombre.
Voix de service informatif ou commercial ayant pour but la transmission individuelle d'informations relatives à une société ou à un utilisateur	Serveur téléphonique	Transmission d'un message informatif individuel et d'une image de marque. Nécessité d'efficacité due au coût pour l'utilisateur. Nécessité d'imitation due au rapport quotidien de l'utilisateur avec le téléphone.
Voix de service personnel assistant à la conduite.	G.P.S	Au départ, transmission d'une petite variété de syntaxe mais un lexique infini, avec une clarté et une précision attendues pour ne pas déconcentrer la conduite. Il y a une exigence à propos de la taille du système et de sa rapidité

[35] Mouvement culturel, spirituel et intellectuel prônant le dépassement humain par l'usage des sciences

		pour s'intégrer à l'habitacle du véhicule et toujours ne pas gêner la conduite. S'ajoute à cela un atout commercial autour de la personnalisation de la voix.
Voix de service personnel permettant d'apprendre une langue	Traducteur, Dictionnaire, Jeu Educatif	Synthèse dans une ou plusieurs langues, de n'importe quel mot de lexique la composant. La taille attendue est réduite.
Voix d'information, de représentation	La voix des hackers	Synthèse anonyme utilisant des moyens largement répandus, populaires.
Voix palliative, de soin,	L'assistanat médical	Attentes indéterminées. De l'ordre de la science fiction.
Voix de service personnel divers : de l'exécutant loyal à l'assistant	La domotique	Attentes variées allant d'une synthèse rapide de mauvaise facture permettant de monitorer certaines informations à une imitation humaine de grande qualité.

b. Un lien entre la fonction et la technique de synthèse

i. La synthèse par concaténation d'unité et les annonces publiques :

Ces voix ont en commun le paradoxe de la synthèse par concaténation d'unité : un timbre fidèle mais une rigidité très facilement identifiable. L'illusion humaine que procure la synthèse concaténative est passagère, trahie par l'aspect mécanique de l'apparition des échantillon-phonèmes. Cette rigidité phonématique est due à la reproduction à grande échelle de ces messages et à l'infime variation qui les compose (souvent de la synthèse mot à mot) : Annoncer un quai, un numéro de train, l'action de quitter un lieu à tel horaire.

L'irritabilité face à ce genre de voix est due à l'aspect impératif de cette inexorable prosodie. Quel que soit l'individu, l'identité vocale restera inchangée, la formulation identique, à n'importe quelle heure de la journée, sans variation en fonction de la température ou de l'affluence. Il n'y a pas de dialogue, de discussion, d'échange, d'interactivité possible. L'action de l'humain est prise en compte par un algorithme qui n'a pour seule mission que de fournir une information par le biais de l'audition. Les mêmes segments sont générés à partir des mêmes diphones, triphones ect… A l'image du Jingle SNCF qui annonce la voix de Simone, ces voix doivent être étudiées pour répéter un grand nombre de fois des séquences identiques sans irriter. Malheureusement l'absence d'adaptabilité à une situation, ce qui se traduit en sentiment par « la compassion », est un fort vecteur d'irritabilité.

Nous avons ici une rencontre quotidienne du domaine publique de l'ordre de l'exploitation d'un lieu. En plus de véhiculer une information, on véhicule aussi l'esprit d'un lieu, de l'exploitant, de la marque vendue… Nous avons donc là une identité vocale protéiforme qui dépend de la voix enregistrée composant le corpus mais dont certains éléments communs peuvent mener à un archétype. Pour établir ce profil, j'ai établi une analyse perceptive personnelle à partir d'un corpus de voix composé de Simone, la voix de la SNCF, d'annonces de la RATP de R.E.R et de l'horloge parlante.

Voici le portrait :

Les informations physiques :

Quel que soit l'extrait envisagé, le lieu et la voix, la personne parlante a l'air en excellente santé, sans soucis articulatoires. D'un âge mûr et donc potentiellement responsable (entre 30 et 40 ans), sa voix semble issu d'un appareil phonatoire bien utilisé (la voix est placée correctement). Un aspect toutefois choquant est l'absence de respiration. Quelle que soit la longueur ou la rapidité de l'intervention, la respiration n'est pas synthétisée.

Les informations sociales

Elles sont difficiles à imaginer par le gommage volontaire d'appartenance communautaire ou de déterminismes sociaux. Ces annonces sont destinées à une foule hétérogène. On peut toutefois statuer sur l'accession à un niveau d'étude suffisant pour énoncer clairement des informations à un grand nombre. La politesse marquée nous permet de dire que c'est une voix qui cherche à aller dans le sens de l'auditeur. La répétitivité de ces intonations révèle toutefois une mécanique dont la compassion et l'écoute semblent absentes.

Les informations psychologiques

La variable commune à chacun des individus du corpus est une grande stabilité quelque soit l'annonce et un calme à toute épreuve. La fixité prosodique provoque toutefois la sensation d'un comportement autistique qui peut être irritant ou anxiogène.

Phonématique

Les informations phonématiques sont difficiles à définir de par le peu de variabilité des voix du corpus. Toutefois, cette extrême rigidité et régularité est en soit la définition de la phonématique de ces voix.

Prosodique

La prosodie est rigide et sans variation. Seules des ruptures hétérogènes sont parfois marquées autour des mots variables (quai, voix, horaires ...). La ponctuation est fortement marquée, accentuant parfois la sensation de lenteur de l'écoulement des propos.

Linguistique

Les phrases sont soit déclaratives, soit impératives. Composées d'une seule proposition, elles sont facilement compréhensibles grâce à un vocabulaire accessible et parfaitement clair.

Nous donnerons un nom générique à cette large identité : **la prévention bienveillante.**

ii. La synthèse formantique en générale :

Nous allons maintenant rencontrer une déclinaison d'emplois de la synthèse par règles (ou par formant) ou à prédiction linéaire. Comme nous l'avons vu auparavant, ce sont des techniques avantageuses en terme de mémoire, d'encombrement et de consommation. Il est donc logique de retrouver cette technologie dans les systèmes embarqués, les systèmes peu coûteux, les systèmes favorisant la rapidité à la qualité timbrale et donc finalement les objets de consommation de masse.

Ces synthèses fonctionnent selon des modèles réducteurs par rapport à la complexité de l'écoulement fluide qui permet la voix parlée. Les timbres, produis par des sinus filtrés (les fonctions d'onde formantique) ne permettent ni une grande richesse (autant en terme de spectre fréquentielle qu'en terme d'évolution temporelle), ni une grande variété mais autorisent une grande marge de modulations tonales et rythmiques. Il devient alors beaucoup plus simple de simuler précisément une prosodie.

Cette technologie se retrouvant dans trois emplois différents, nous allons d'abord décrire les possibilités de la synthèse formantique les plus répandues dans les outils populaires. Puis en les reliant à leur emploie, nous proposerons des identités empiriques.

Pour se faire je me baserais sur l'outil VoiceOver fournit avec le système d'exploitation OS 10.6.8 et les versions supérieures. Il est possible de choisir entre 5 voix féminines et 5 voix masculines et une série de voix d'effets. Seulement 3 voix masculines (Fred, Junior et Ralph) et 2 voix féminines (Princess et Kathy) fonctionnent à l'aide de la synthèse par formant. Je ne connais pas la technologie employée pour les autres voix mais il est possible d'identifier la présence du découpage des phonèmes propre à la synthèse MMC.

Pour chacune de nos 5 voix, il est aussi possible de moduler très simplement 4 paramètres de ces voix : débit, hauteur tonale, volume et intonation. Ces paramètres influent directement sur la gestion des fonctions d'ondes formantiques qui définissent les voix dans leur ensemble. Les règles prosodiques et la simulation des émotions est ici identique pour les 5 voix et effets, il s'agira à l'écoute d'essayer d'établir des caractéristiques liant ces 5 voix et effets.

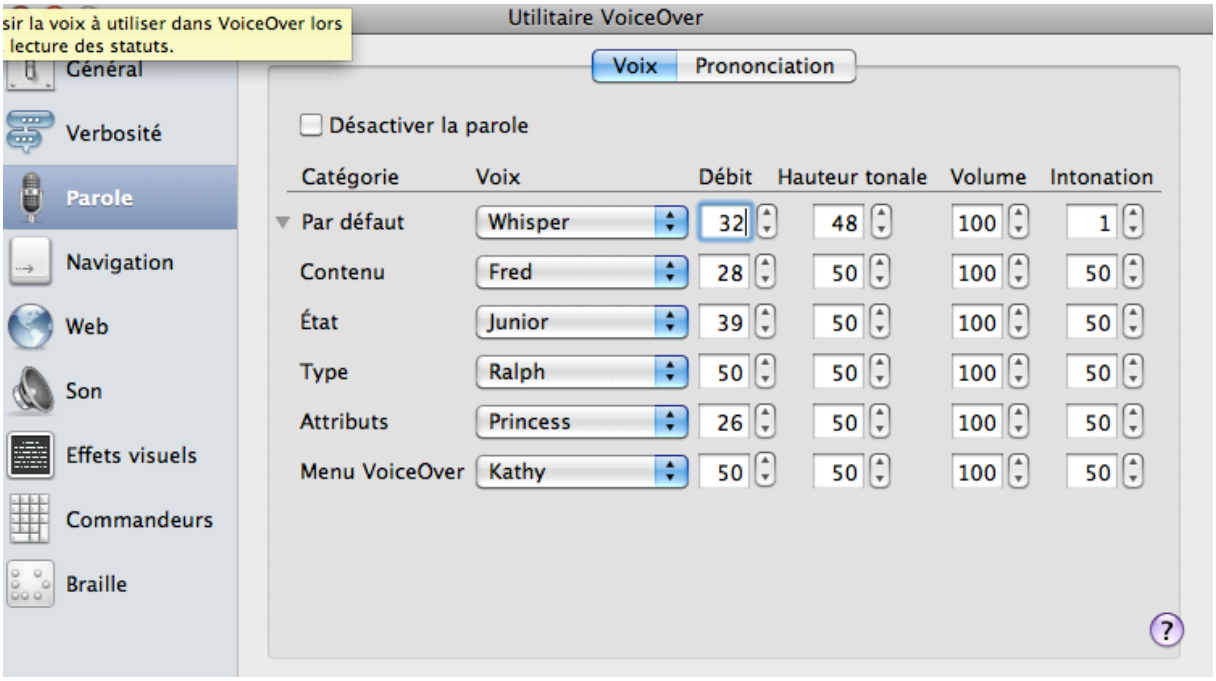

Figure 32 Voice-Over Apple

Après avoir enregistré un corpus de Fred, Junior, Ralph, Princess, Kathy et l'effet Whisper disant « Alright, so i am trying to hear what are the technologies of my voice », j'ai noté mon ressenti pour chacune d'elles. Il s'avère qu'elles provoquent une sensation relativement identique :

Une sensation de froideur et d'agressivité qui pourrait se traduire par un spectre ramassé dans le bas medium (vers 100 Hz) et la zone d'agressivité (en 2kHz et 4kHz). Cette sensation de froideur vient aussi certainement d'une gestion de la prosodie extrêmement mécanique répondant aux fameuses règles phonatoires évoquées à la section précédente.
Il n'y a pas de sensation de résonnance dans ces voix, ce qui efface la sensation de corps, de bouche, de cage thoracique ou de toute cavité supra-glottique. L'articulation n'a rien d'humain et semble bien être l'enchaînement de signaux segmentés.

L'aspect désincarné de ces voix créé donc à la fois une sensation d'empattement allant, pour certaines voix graves comme celle de Ralph, jusqu'à la rondeur, et en même temps une sensation de froideur voire d'agressivité. On peut supposer que ces sensations duales résultent de la sommation de sinus et de bruits, ainsi qu'à l'absence de critères acoustiques pour simuler les résonnances.
Grâce au logiciel PRAAT, j'ai pu faire une rapide analyse du corpus pour appuyer mes sensations avec des arguments plus tangibles (VOIR ANNEXE PRAAT). L'évolution tonale au cours du temps et la répartition fréquentielle de chaque voix nous renseigne d'une part des similitudes entre chaque voix, mais aussi de l'association effective de signaux sinusoïdaux pour la partie voisée et de bruit pour la partie fricative.
On constate que la variation tonale moyenne est d'environ 100Hz quelle que soit la voix considérée, que le spectre voisé se repartit de 20Hz à 1 kHz et la partie fricative de 1 kHz à 10 kHz. Le spectre ne s'étend pas au delà ce qui participe à la sensation synthétique (c'est une zone étrangement absente ou au contraire étrangement présente dans les sons de synthèses, qu'il est difficile à modéliser)
On observe pour chacune des voix deux pics à 4 et à 6,5 kHz, qui renforcent l'agressivité mais sont certainement les résultats d'une part du contenu de l'énoncé et aussi de la modélisation de certains phonèmes fricatifs.

Il existe donc des sensations duales au sein de ces voix, inhérentes au contenu phonétisé, autant qu'à la gestion de la prosodie, or nous avons vu qu'il était aisé de la manipuler. Il va donc être maintenant possible de statuer sur les identités possibles de voix formantiques en croisant ces informations avec les emplois listés précédemment.

iii. La synthèse formantique ou à prédiction et l'anonymat

L'utilisation de cette technique pour ses vertus d'anonymat, a créé au fil du temps un croisement esthétique aujourd'hui indissociable. L'aspect banal et bas de gamme s'est associé au signal électrique déshumanisé menant à une métaphore de l'humain, englobant tous les humains. Dans ce cas d'utilisation, la voix synthétique est à l'image de la forme ovoïde qui représenterait universellement l'être humain par association à la forme de son crâne. C'est à la fois un dénominateur commun identitaire, et en même temps une absence totale d'élément individualisant.

Cette voix est paradoxale, puisque son identité est en fait son absence d'identité. C'est une voix émissaire, une représentation, un hologramme, une occurrence. Elle n'est pas issue d'une entité intelligente (humaine ou machine).

Peut-on alors parler d'une identité vocale directe ? Elle est difficile à définir en dehors du message que cette voix porte, mais l'effacement d'un émissaire vis à vis de son message est en soit une modalité d'existence.

Pour établir ce profil, je me suis basé sur un corpus établi à partir 3 vidéos de propagande du groupe Anonymous.

Les informations physiques :

Absence corporelle totale. L'être parlant est un être numérique modulable. Genre et âge sont modifiables.

Les informations sociales

Absence d'appartenance sociale à un groupe humain. L'aspect social de cette voix est celui d'une machine exécutant les ordres d'une entité ordonnatrice.

Les informations psychologiques

L'aspect psychologique de cette voix est celui d'une machine exécutant les ordres d'une entité ordonnatrice.

Phonématique

Les phonèmes produits couvrent un spectre tronqué à partir de 10kHz. Ils s'organisent clairement selon une dualité entre voisement sinusoïdale et bruits blancs. L'équilibre timbral entre ces deux pôles permet de basculer d'une voix quasiment chantée à une voix chuchotée. Son aspect reste toutefois synthétique pour les raisons évoquées précédemment (Energie dans le bas medium et la zone d'agressivité accompagnées d'une enveloppe très peu variée)...

Prosodique

La prosodie est rigide. Les intonations peuvent être fortement marquées. Seulement la rencontre entre une prosodie invariable en fonction d'une syntaxe et les informations physiques nous renseignant sur l'aspect numérique de l'être en train de parler, transmettent la sensation d'une prosodie mécanique. Le style de ces synthétiseurs vient d'une prosodie faussement « idéale » issue des règles phonématiques. La totale absence de disfluences créé de plus cette sensation d'oralité factice.

Linguistique

Pas d'informations.

Nous donnerons un nom générique à cette large identité : **l'émissaire numérique**

La disfluence chez l'émissaire numérique n'est pas réellement désirée. Tout aspect individualisant pourrait nuire ou dégrader le contenu du message.

iv. La synthèse formantique ou à prédiction et la traduction

Le cas de la traduction est différent pour plusieurs raisons. La notion d'émissaire disparaît puisque la personne exerçant le contrôle sur la synthèse est l'utilisateur, de plus la construction syntaxique, l'intonation, la personnalisation et finalement l'imitation humaine ne sont pas directement nécessaires. La tâche de l'apprentissage ou de la recherche de connaissance ne crée pas le même rapport à cet être purement numérique. Cet outil est aujourd'hui obsolète, dépassé par les ordinateurs dont les contraintes techniques permettent d'explorer des technologies dont le rendu qualitatif est supérieur à celui à la synthèse formantique (typiquement les autres voix proposées par Voice-Over de Apple). L'esthétique « vintage » développe alors une idée de nostalgie. La voix est inoffensive à la différence de l'utilisation qu'en font les hackers. Elle est aussi produite par un objet portatif, il n'est alors pas question d'avoir une sensation d'humanité issue de ces simili-calculettes ou ces jouets.

Pour ce profil je me suis basé sur un corpus fait à partir d'enregistrements du jeu éducatif dictée magique.

Les informations physiques :

Absence corporelle totale. L'être parlant est un être numérique.

Les informations sociales

Absence d'appartenance sociale à un groupe humain. Absence totale, car transparence du système. L'aspect social de cette voix est celui d'une machine exécutant les ordres d'un utilisateur.

Les informations psychologiques

Absence totale, car transparence du système.

Phonématique

Les phonèmes produis couvrent un spectre tronqué à partir de 10kHz. Ils s'organisent clairement selon une dualité voisement sinusoïdale, bruits blancs. L'équilibre timbral entre ces deux pôles permet de basculer d'une voix quasiment chantée à une voix chuchotée. Son aspect reste toutefois synthétique pour les raisons évoquées précédemment (Energie dans le bas medium et la zone d'agressivité accompagné d'une enveloppe très peu variées)...

Prosodique

La prosodie est rigide. Seulement ici, les intonations peuvent être fortement marquées. Seulement la rencontre entre une prosodie invariable en fonction d'une syntaxe et les informations physiques nous renseignant sur l'aspect numérique de l'être en train de parler, transmettent la sensation d'une prosodie mécanique. Le style de ces synthétiseurs vient d'une prosodie faussement « idéale » entraîner par les règles phonématiques. La totale absence de disfluences créer de plus cette sensation d'oralité factice.

Linguistique

Les informations linguistiques dépendent du dictionnaire chargé dans le programme. Il y a donc un lien direct qualitatif entre le programme et les informations linguistique.

Nous donnerons un nom générique à cette large identité : **la bouche numérique**

La disfluence chez la bouche numérique est plutôt contre-productive. L'idée veut qu'une langue connue s'énonce selon une fluence parfaite, donc l'entité commerciale n'a pas d'intérêt à commercialiser un appareil produisant de la disfluence.

v. La synthèse formantique à prédiction et l'assistanat médical

Dans ce cas, l'utilité n'est plus commerciale. Le handicap menant au mutisme efface le jugement de valeur sur la qualité de la synthèse. L'utilisateur de cette technologie est dans un rapport de « besoin », ou en tout cas de nécessité plus grande que l'utilisateur lambda de la synthèse vocale. Chaque cas devient alors particulier puisqu'il est le croisement entre la sémiologie d'un handicape, le corps d'un utilisateur et la synthèse vocale. Une ablation du larynx ne transmet pas la même histoire, ni la même idée d'un corps qu'une ablation de la moelle épinière.

Ici, quelle que soit la configuration, la voix synthétique est rattachée à un corps réel. Elle devient l'émanation de ce corps. L'imitation de la voix disparue serait un idéal de l'ordre de la science-fiction, il semble alors implicitement inévitable qu'il y ait un gouffre entre la production synthétique et la voix disparue.

Lorsqu'on écoute Stephen Hawking parler, la voix formantique électronique permet la transmission du message sans créer l'illusion de la cure du malade. Sa voix ne reviendra pas et il n'est pas question ici d'imiter la défunte. Le décalage entre le résultat attendu et le résultat obtenu permet donc un travail de construction mentale : le contenu de l'énoncé est lié à la prosodie présentée par le système de synthèse. Puis la rencontre visuelle avec le corps du patient nous permet d'imaginer alors une identité hypothétique. Cela fonctionne comme un système de projection aboutissant à des suppositions.

Les informations physiques :

La voix ne contient pas d'information corporelle en soit. Genre et âge sont modifiables. La construction d'une identité physique se fait dans un rapport au corps visible de l'utilisateur.

Les informations sociales

Absence d'appartenance sociale à un groupe humain. Cependant son cadre d'utilisation confère à cette voix une composante « d'infirmier » ou « d'assistant de vie » fournissant une aura bienveillante, ou compassionnelle.

Les informations psychologiques

Absence totale, car transparence du système.

Phonématique

Les phonèmes produits couvrent un spectre tronqué à partir de 10kHz. Ils s'organisent clairement selon une dualité voisement sinusoïdale, bruits blancs. L'équilibre timbral entre ces deux pôles permet de basculer d'une voix quasiment chantée à une voix chuchotée. Son aspect reste toutefois synthétique pour les raisons évoquées précédemment (Energie dans le bas medium et la zone d'agressivité accompagnée d'une enveloppe très peu variée)...

Prosodique

La prosodie est rigide. Seulement ici, les intonations peuvent être fortement marquées. Seulement la rencontre entre une prosodie invariable en fonction d'une syntaxe et les informations physiques nous renseignant sur l'aspect numérique de l'être en train de parler, transmettent la sensation d'une prosodie mécanique. Le style de ces synthétiseurs vient d'une prosodie faussement « idéale » entraînée par les règles phonématiques. La totale absence de disfluences crée de plus cette sensation d'oralité factice.

Linguistique

X

Je nommerais cette identité : **La projection numérique**

vi. La synthèse avec le projet VocalID et l'assistanat médical

Il existe toutefois, au sein de cette problématique de l'identité vocale, des chercheurs qui tentent de produire des voix synthétiques uniques pour permettre à certains patients mutiques de retrouver une part de leur individuation. C'est le cas du professeur Rupal Patel (Northeastern University) et du docteur Tim Bunnel (Directeur du Center for Pediatric Auditory and Speech Science) à travers leur projet VocalID. Leur motivation vient de la constatation qu'il n'existe pas d'individuation des solutions de synthèse proposées aux malades atteints de mutismes sévères.

L'idée du projet s'apparente à celle de la convolution de deux fragments pour former un nouvel objet unique. En utilisant un signal dit source, on va exciter un signal dit de substitut qui mènera à la voix synthétique finale. On peut schématiquement décrire le fonctionnement de vocal ID comme le suivant : le signal source se comporte comme un générateur d'ondes sinusoïdales plus ou moins complexes et le signal de substitut va se comporter comme un filtre.

Les patients atteints de troubles du langage sévères ou d'un mutisme partiel peuvent encore produire des sons vocaux, permettant de fournir les données sources. Une réduction de ces données est faite pour mener à un générateur source.

Un donneur bénévole va alors enregistrer sa voix selon un protocole établi sur le site de VocalID, offrant la possibilité d'extraire des données de filtrage et de comportements timbraux uniques. Il fournit alors le signal de substitut. Par association de ces deux signaux, on aboutit à une voix synthétique assurée comme unique.

La technique qui n'est pas révélé sur le site ressemble à celle employée dans la synthèse MMC : corpus d'enregistrement et réductions à des marqueurs relatifs à cette voix, possibilité de large modulation prosodique.

Les malades peuvent alors choisir des signaux de substituts qui correspondent à ce qu'ils estiment être l'identité de leur voix disparue. Le choix de cette voix est fait en association avec les employés de VocalID.

Cette solution n'est pas encore popularisée, et implique une grande quantité de donneurs pour être effective. Elle n'est donc pas suffisamment médiatisée pour créer un archétype vocal en

soit qui sera identifiable comme étant une production issue de VocalID. Toutefois, à la différence de la projection numérique de Stephen Hawking, il y a une vocation d'imitation bien plus forte. L'hybride vocal créé cherche à s'associer au corps du patient.
Sans exemple précis à écouter, il n'est pas possible de statuer sur l'identité vocale qui se transmet. On peut toutefois imaginer la sensation extracorporelle de voir un corps mutique avec une voix lui correspondant s'exprimer sans impliquer de mouvements buccaux. Il y a toujours projection, mais ce n'est plus réellement une voix numérique mais bien une voix fantomatique, une réduction, une forme.

Cette notion de réduction d'une voix à des marqueurs nous invite à envisager notre exploration des identités vocales un peu différemment pour la suite. Puisque nous allons rencontrer des synthèses faites à partir de corpus d'êtres humains tous très différents, il va être dur de pouvoir statuer sur des informations physiques, sociologiques ou linguistiques. Nous pourrons par contre, grâce à un corpus varié, constater ce qui persiste pour chacune des voix et l'établir comme paramètre identitaire de ce type de voix.

vii. La synthèse MMC et la voix commerciale

A la différence des synthèses précédentes, nous entrons ici dans un domaine où la technique de synthèse permet une simulation très proche de la voix naturelle. Il faut donc changer de stratégie pour essayer de proposer une définition de l'identité des voix des proposées.

La voix commerciale est un terme englobant la voix représentant l'entreprise et qui propose un service qui n'est pas nécessairement tarifé mais qui véhicule une image commerciale, mais aussi la voix ayant pour but d'amener à une transaction commerciale d'ordre indéfini.
J'ai élargi le serveur téléphonique à la voix commerciale car la télécommunication se diversifie à travers internet. Toutefois son emploi reste identique : créer un contact efficace, rapide simulant la conversation humaine.

Pour avoir une idée des techniques de synthèse employées et des voix impliquées, il a fallu se rendre du côté du groupe Acapela[36], grande société qui propose des services de synthèses vocales pointues à un très grand nombre d'entreprises, parmi lesquelles on trouve des banques (Laposte, CIC...), des opérateurs téléphoniques (Bouygues, SFR), des assurances (AXA, services de renseignements (Les pages jaunes, Reverso, Babylon...) et une grande quantité d'autres services quotidiens.
C'est donc auprès d'Acapela que nous trouverons un corpus intéressant pour essayer d'établir une identité relative à la voix commerciale : https://soundcloud.com/acapelagroup
En analysant ce corpus, nous pourrons empiriquement se demander pourquoi ces voix sont identifiables en tant que matériel synthétique. Il est donc composé des voix de Alice, Antoine selon différentes humeurs, Julie, Margaux selon différentes humeurs, et Louise une voix canadienne.

Il est alors possible d'identifier des caractéristiques communes à toutes ces voix :

- Technique :

Certains enchaînements de phonèmes se font avec difficulté et on identifie un artefact de synthèse dans la fluence continue de la voix. Cette erreur n'a rien d'un trouble de l'appareil phonatoire, mais s'apparente bien à un problème de synthèse où le début d'un phonème se juxtapose mal au suivant, mettant en difficulté l'intelligibilité, à l'image d'un défaut d'articulation synthétique. Ces accidents surviennent de manière sporadique et semblent être issus d'une mauvaise concaténation du corpus d'enregistrement menant à certains phonèmes tronqués. Au vu des extraits du corpus, il semble que ces erreurs augmentent lorsqu'il y a aussi manipulation de la fondamentale pour gérer la prosodie.
Chacune des voix semble aussi fortement compressée, évoluant dans une dynamique restreinte. Après analyse avec PRAAT, on observe une dynamique de 6dB même pour l'extrait de Arthur contenant une version expressive criée.
De plus le spectre de ces voix se concentre essentiellement dans les 1kHz et se retrouve tronqué 10,5kHz certainement dut à des contraintes d'échantillonnages.
On observe toutefois des variations de la fondamentale différentes pour chacune des voix.

[36] Liste des compagnies ayant utilisé les services d'Acapela : http://www.acapela-group.com/gallery/?lang=fr

- Esthétique :

On observe toujours une absence totale de respiration. Cela renforce une sensation de malaise difficile à distinguer mais cela appuie l'aspect synthétique de ces voix lorsqu'à la fin d'une phrase, s'enchaîne alors une nouvelle sans repos. L'impossibilité d'un tel débit est certainement le facteur principal d'identification des voix synthétiques.

De plus, l'organisation prosodique correspond à un idéal de lecture : des intonations illustrant parfaitement une ponctuation, l'absence de bruits buccaux parasites, l'absence d'inflexions qui pourraient sembler involontaires. On retrouve l'un des aspects de la prévention bienveillante incarnée par exemple par Simone, qui est une sensation de politesse exagérée et invariable.

Il y a une absence totale de disfluence ou de tout autre mécanisme de la communication orale spontanée. Cette voix semble en représentation, à l'image d'un acteur lisant un texte. L'interactivité est un élément qui est souvent manquant au système de communication commerciale impliquant la synthèse vocale et quand bien même, la synthèse n'est pas encore envisagée comme un processus de dialogue où l'utilisateur modifie par son utilisation la prosodie de la synthèse.

Cette représentation donne la sensation d'une image d'Épinal de la voix humaine, comme un visage ayant abusé de chirurgie esthétique et dont la symétrie serait trop parfaite. De plus, dans ce qui est sensé être une simulation de discussion, l'impossibilité ne serait-ce que de produire une réaction chez son interlocuteur est quelque chose qui devient rapidement agaçant.

Cette sensation de perfection factice et absurde peut aussi s'incarner dans des simulations d'émotions telle que la tristesse, la colère. Même dans ces exemples, la sensation d'un rail prosodique est identifiable.

J'appellerais donc ces voix en raison de leur invariabilité et pourtant la réussite de la simulation : **le miroir trop parfait**.

viii. La synthèse vocale et le G.P.S

Plusieurs techniques remplissent le cahier des charges fonctionnelles établi auparavant mais la technique choisie a été dans la plupart des cas la synthèse concaténative. La petite variété de syntaxe, le lexique fini, et l'attente d'une voix fidèle par la relation qu'entretient un conducteur avec son co-pilote ont certainement orienté vers ce choix.

La petitesse des systèmes (que ce soit des boitiers externes faisant environ 10*10cm ou des téléphones portables) fait que la qualité timbrale est doublement diminuée : réduction de la fréquence d'échantillonnage des échantillons sonores et diffusion sur des haut-parleurs de 1 à 2 cm de diamètre.

Bien qu'il soit envisageable d'employer une technique formantique, de la synthèse MMC, ou une technique par prédiction linéaire, la plupart des voix que j'ai rencontrées au cours de ma recherche d'exemples s'apparentaient à de la synthèse concaténative de par leur grande rigidité et leur timbre proche de la voix humaine (toujours à l'image de Simone de la S.N.C.F). Les exemples choisis sont extrêmement hétérogènes : voix téléchargeable pour les systèmes TOMTOM, voix du GPS Voice Navigation, application pour windows phone.[37]

A l'écoute de ces exemples, il est difficile d'établir des profils pour chacune des voix. Il en ressort toutefois plusieurs informations relatives à la technique de la concaténation et à son utilisation. Il est possible de catégoriser les voix de G.P.S en deux groupes : les voix de guidages et les voix de divertissements.

- Les voix de guidages sont les voix par défaut fournies avec les périphériques. En général, il existe un version masculine et féminine qui pourrait se définir de la même manière que la voix de la prévention bienveillante décrite auparavant : Prosodie inexorable, absence de respiration, politesse excessive**.**
- Les voix de divertissements sont des voix de qualités variables, enregistrées par les sociétés qui commercialisent les G.P.S voire des indépendants. Elles sont en général des imitations de personnages médiatiques réels ou non. L'aspect comique apparaît souvent comme gadget puisque l'extrême répétitivité des interventions du G.P.S réduit la surprise première de la blague à une mécanique sans interaction. La tentative de vitrine humoristique ou d'accompagnement du conducteur par un divertissement est souvent un échec, remplacée par l'agacement de ces blagues incessantes.

[37] Lien vers les extraits de voix de G.P.S :
https://www.youtube.com/watch?v=wio1wkhyrLg
http://tomtomvoix.free.fr/index.php?cat=voix&page=2
http://www8.garmin.com/vehicles/voices/
http://www.today.com/video/today/52668916#52668916

Je prendrais en exemple les références vaseuses très peu variées que peut faire l'imitation de D.S.K ou l'imitation du rire niais qui ponctue chaque intervention de la voix de Bourville...

La rigueur de la synthèse concaténative alliée à la rigidité syntaxique de l'indication routière (A tant de mètres, vous prendrez, telle direction, tenez telle ligne, vous êtes arrivé) font de la voix de divertissement de G.P.S, un automate guignol**.**

Ici toutefois, à la différence des voix commerciales, le produit est déjà vendu et il n'y a pas de troubles quant à la relation avec la machine : l'utilisateur sait que la voix émane d'un boitîer, d'un outil. L'humain donne un ordre, ici une direction, et l'outil fournit un service, lui indiquer vocalement la direction. Il y a donc nécessairement un rapport d'automate car il n'y a pas d'illusion dans le G.P.S. Le développement des voix de G.P.S se concentre sur des imitations grossières plus que sur l'illusion de la discussion avec son outil. Le changement viendra peut-être si un utilisateur converse avec l'entièreté de sa voiture à laquelle on prête plus volontiers une conscience (cf Asimov dans le Grand livre des robots, Christine de John Carpenter, K2000...).

Pour le moment, de par l'accessibilité des outils et leur aspect rigide et leur limite linguistique de les qualifier **d'automates numériques**.

ix. La synthèse MMC et la domotique

En domotique, il existe deux identités possibles : l'exécutant et l'assistant.

L'exécutant est la voix qui permet le monitoring d'une tâche, renvoyant une information relative à un protocole. Il ne simule pas d'intelligence artificielle, et ne prends n'a pas de pouvoir décisionnel, de libre arbitre.

Actuellement, l'exécutant s'incarne le plus souvent dans des sortes de « Maisons Assistées par Ordinateur » dans lesquels il est possible d'effectuer un certain nombre de tâches par contrôle vocal. Une voix nous renseigne alors sur les états des actions. A travers des exemples de maison contrôlée par des applications Windows et Androïde, on constate que la technique employée semble être une synthèse MMC.

Les voix répondent selon des syntaxes préétablies, ne simulent aucune émotion en relation avec la situation évoquée. L'exécutant est à l'image de la voix de G.P.S, un **automate numérique**.

Il reste donc l'assistant, dont la seule incarnation que j'ai pu rencontrer est SIRI, le programme de reconnaissance et synthèse vocale développé par Apple. C'est le seul programme simulant un échange intelligent entre l'agent humain et la voix de synthèses. SIRI possède une voix d'homme adulte dont l'accent plat ne nous permet de statuer sur une identité administrative avec précision. Au vu de la variabilité des textes qu'il synthétise, des possibilités de modulations prosodiques et de la phonématique générale de la voix, SIRI semble être créé à partir d'une synthèse MMC.

A l'écoute des exemples de SIRI disponible sur internet, on constate que malgré l'intelligence artificielle, la rigidité prosodique empêche toujours de considérer cette voix autrement que comme celle d'un programme bien délimité.

Ici, plus que dans tous les autres cas, SIRI par l'entretien d'une discussion témoigne d'une forme d'intelligence et de libre arbitre qui simule une identité. Son émanation vocale semble alors constituer une limite empêchant d'exprimer toute cette identité. SIRI simule des conversations mais synthétise une oralité qui ressemble à de la voix lue. De plus, cette simulation omet toujours la simulation de l'appareil phonatoire, de sa respiration, de son épuisement autant que de sa stabilité.

SIRI n'est donc pas différent de **l'automate numérique**.

Ces exemples montrent que la problématique de la disfluence dans la synthèse vocale de la voix spontanée est primordiale. L'intelligence artificielle derrière le synthétiseur ne suffit pas à témoigner de l'incarnation d'une identité, il faut aussi mettre en place cette mécanique de recherches, d'erreurs et d'aller-retours décrit à travers les disfluences.

La domotique est donc un champ d'exploration très important à venir pour le secteur de la voix synthétique. Les Maisons Assistées par Ordinateur ou SIRI ne sont que des essais face à ce qui est envisagé par certaines œuvres de science-fiction (cf Her de Spike Jonze, dans lequel il est possible de vivre des histoires d'amour avec des OS). Il semble alors impératif de traverser les Arts pour finir notre cartographie des identités des voix synthétiques qui nous entourent.

c. La représentation des voix synthétiques dans les Arts

Les Arts fonctionnent sur des modes de représentation de phénomènes ou d'idées. La représentation (qu'elle soit figurée, abstraite, conceptuelle...) côtoie souvent les notions d'imitation et de simulation. Que ce soit à travers un tableau pompier, une peinture rupestre, une performance dansée ou une installation concentrée autour de la donnée numérique, la notion d'imitation ou de simulation s'incarne.

Il s'avère que comme une grande quantité d'éléments technologiques, les Arts se sont emparés des techniques de synthèse vocale. La reproduction de masse d'une même voix, qui produisait l'anonymat des « cyber-activistes », peut aussi devenir une idée artistique aux multiples facettes : médiation de masse, culture technologique, aspect science fictionnelle, déshumanisation, langage martial, pur plaisir esthétique... L'intérêt de la production artistique réside souvent dans la difficulté à identifier la frontière entre la fonctionnalité et l'aspect commercial de l'objet produit. Les œuvres poussent à parler différemment des voix synthétiques.

L'enjeu étant d'arrivé à des portraits dont l'aura rayonne plus largement qu'à une échelle personnelle, nous nous intéresserons plus particulièrement à des œuvres ayant connu un certains succès publiques, impliquant donc une trace mémorielle importante.

Au sein des Arts, l'utilisation de la synthèse vocale a répondu à des attentes différentes. Pour chacune de ces raisons, nous prendrons un ou plusieurs exemples et étudierons quelle représentation de la synthèse vocale a été transmise.

i. Esthétique

La synthèse de la voix parlée, quelle que soit la technique, a été maintes fois utilisée pour ses qualités esthétiques intrinsèques. Les outils de synthèses sonores ont été essentiellement pris en main à partir des années 80 par deux courants musicaux : la musique électro-acoustique et la musique électronique. Par détournement, elle devint un outil de création vocale au service d'une partition ou même de systèmes musicaux improvisés.

La musique électronique est une branche musicale très populaire qui englobe aujourd'hui des styles qui occupent l'essentiel du paysage actuel : l'électro, la techno et d'autres déclinaisons et évolutions... Kraftwerk, pionnier du genre, pratique une musique que l'on a qualifiée de Synth-Pop, Krautrock, Techno... Il est sûr qu'ils sont responsables de la popularisation d'une musique technologique à base de synthèse, créant un lien étroit entre un outil-machine et un agent humain, jusqu'à aller à être représentés sur scène par des androïdes.

Ils emploient la synthèse de la voix parlée à de nombreuses reprises dans leur morceau, rythmant alors les pistes de danse. Il n'est pas ici question d'oralité spontanée, mais bien d'une organisation vocale musicale autour d'une rythmique.

Je prendrais en exemple Musique Non-Stop (ils étaient aussi l'étendard d'une musique Européenne, mais c'est une autre histoire) qui fut classé 13eme des ventes en Allemagne, dont la rythmique et les paroles sont faites par des synthétiseurs à formants issus d'un traducteur Texas Instrument et d'un synthétiseur NED. La chanson commence sur des voix synthétiques égrenant une rythmique à la manière beatbox. Puis d'autre voix vont scander selon des modes différents : « musique, non-stop »

Les voix synthétiques dans ce cas sont utilisées pour produire un décalage. L'imitation de la voix parlée n'est que grossière, car l'élément recherché est au contraire son aspect synthétique. Kraftwerk passera l'essentielle de sa carrière à magnifier la production de la machine, à la rendre mystérieuse, spirituelle ou pleine d'humour, en n'utilisant que des périphériques électroniques pour composer leur musique.

Ici le décalage produit par le remplacement d'une voix habituellement humaine, par une synthèse dont l'imitation est approximative, produit un effet se déroulant en deux temps : une sensation de déshumanisation, ou en tout cas la disparition d'une trace directe de la production humaine (puisque ce sont quand même des instruments programmés par les membres de Kraftwerk), puis la sensation musicale. L'organisation musicale aide à entrevoir l'aspect synthétique comme un choix artistique et à redécouvrir les timbres de ces synthétiseurs à formants comme des potentialités en soi, plus que des limitations technologiques. L'enjeu artistique chez Kraftwerk, a selon moi, essentiellement porté sur une esthétisation de la machine.

Dans « musique non-stop », la prosodie hachée, les timbres riches en haut médium, l'articulation approximative et le côté grinçant de certaines voix créée une chorale de machine, une chorale de « freaks » dont la danse frénétique est répétitive. C'est toujours ce décalage qui est magnifié. Ces voix là sont de qualité « médiocre », dans un rapport à la simulation humaine, et c'est ce handicape qui les rends machine, et donc intéressante. Ce type de profil de voix, dont seul l'esthétisme compte, sera nommé « **pure machine** ».

Chez Kraftwerk, les machines chantent, les machines font danser, et peut-être même que les hypothétiques corps de ces voix, dansent. Il n'est pas d'autres idées, que la seule beauté de cette production synthétique.

ii. Les machines imitant humains

Tous les projets artistiques n'ont pas nécessairement pour vocation de magnifier la production synthétique. Au contraire, la dialectique de nombreuses œuvres repose sur le questionnement d'un grand nombre de notions relatives à l'humanité et sa production technique, à un maître et son créateur, à la notion de vie et d'existence ou à son absence, à la notion de hasard, d'erreur ou encore de reproductibilité, tout cela à travers une dialectique synthétique/organique.

Le questionnement méta-physique de l'identité, la honte prométhéenne[38] s'entrecroise dès qu'il est question de simuler une partie de l'humain et de l'inscrire dans un processus de monstration.

Dans notre cas, l'incarnation de voix synthétiques dans l'art est souvent rattachée à des questionnements sur le flux d'informations, sa reproductibilité et son rapport avec l'action humaine (**A piece of Work** d'Anne Dorsen) ainsi que son aspect mécanique, contrôlé ou au contraire imprévisible de la communication au sein de la société. Le réseau et ce qu'il produit tout comme les identités qui le traversent sont aussi des sujets reliés à la synthèse vocale (**Listening** Post de Mark Hansen et Ben Rubin ou **Ip Poetry** de Gustavo Romano). En dernier lieu, la synthèse vocale est un outil de questionnement sur le flou identitaire qui résulte de certains modes de communication et la possibilité de simulation (**Marylin** de Philippe Parenno).

Sans perdre de vue notre définition de l'identité vocale comme la somme d'un contexte donné à un temps t, de marqueurs biométriques et d'une série de comportements individuels et collectifs se traduisant par des variations dans la fluence, nous allons aborder les œuvres citées ci-dessus. Nous verrons que la production artistique cristallise avec beaucoup de justesse les questionnements d'identité vocale et met en exergue l'exercice corporel performatif (parfois hors du langage) qu'est la voix parlée et son rapport à la spontanéité et donc à la disfluence.

[38] Concept relatif à la honte de l'homme face aux outils qu'il a crée et à son infériorité créé dans le rapport de dépendance. Günther Anders parle d'obsolescence de l'homme. Ce concept est orienté comme une critique de la technique, nous rattachons inévitablement à un processus de consommation

Le projet Ip Poetry[39] de Gustavo Romano questionne la subjectivité de la technologie et la place des poètes à travers une installation : un flux de poésie est généré à partir d'internet (à l'image des suggestions dans la barre de recherche de google.fr) et d'un début de phrase prédéfinie.

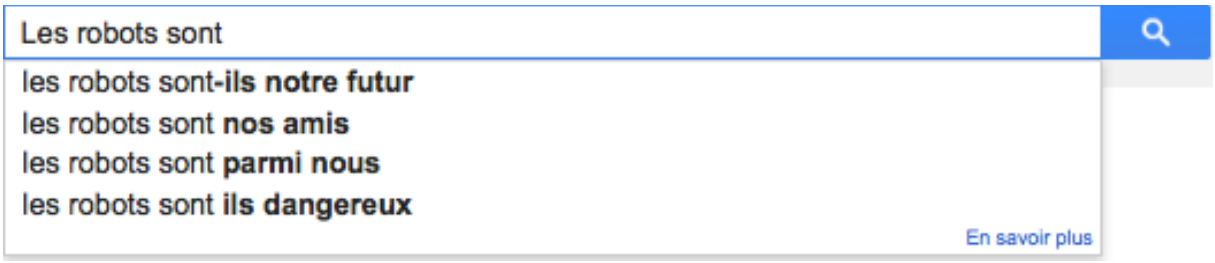

Figure 33 Suggestion Google pour la phrase "les robots sont"

Le texte poétique est alors lu par une voix synthétique espagnole ou anglaise selon une syntaxe préétablie. La voix est représentée par une bouche en gros plan animé de manière saccadée.

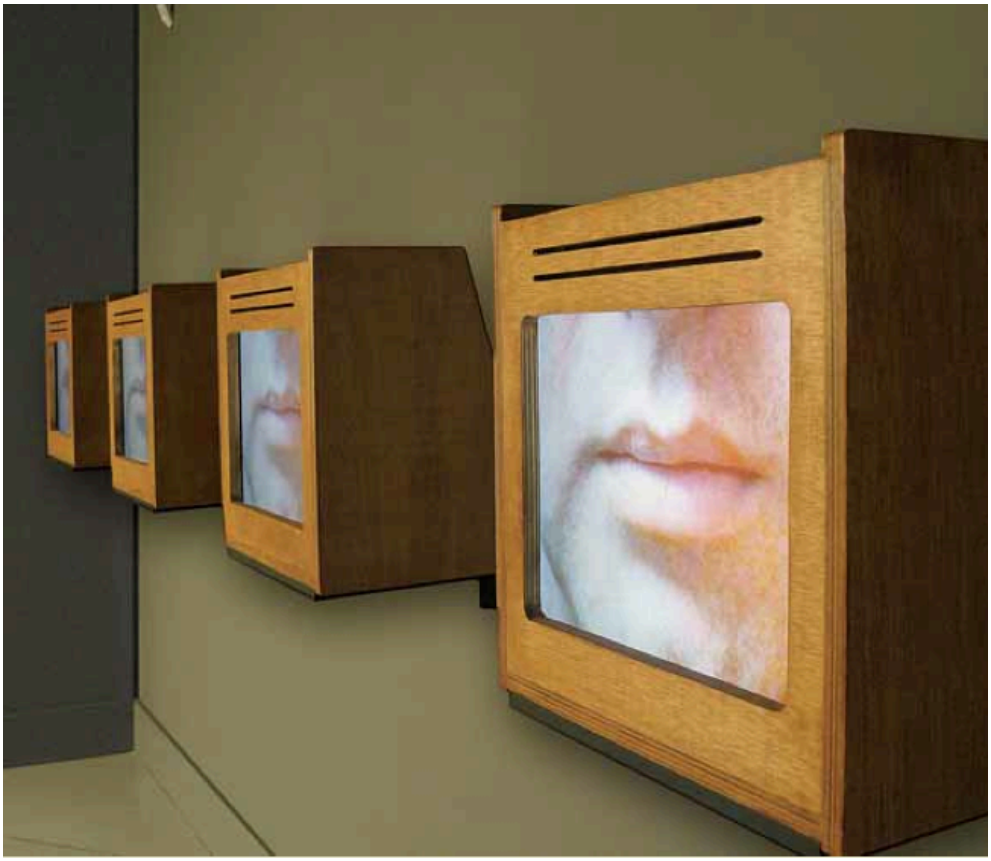

Figure 34 Une des formes de l'installation IP Poetry de Gustavo Romano

[39] Extrait vidéo d'un des poèmes générés dans IP Poetry : http://www.youtube.com/watch?v=GryZYv3Li6w
http://www.youtube.com/watch?v=U5MwEeO7LFI
Une des scénographies de l'installation :
http://www.youtube.com/watch?v=F6ayKoYyfE4&index=3&list=PL9633E6EB2E036C22

La technique employée est une synthèse formantique dont le timbre est relativement restreint et agressif mais l'intelligibilité excellente. La voix interpelle le spectateur dans le contexte d'une lecture poétique. La poéticité est d'abord suggérée par le nom de l'œuvre, mais aussi pas l'absence de fonction informative de la production vocale. Bien que la syntaxe induise une mécanique inexorable, il y a aussi une création rythmique amenant à la découverte surprenante de chaque fin de phrase :

« Sometimes, I dream that I am director in chief
I dream, that I am an american teenager
I dream, that I am compatible with humans
I dream, that I am only a dream
I dream, that I am Marianna
I dream, that I am one of them
I dream, that I am a prisoner on deathrow
My Life is a horrible nightmare,
Sometimes » *Extrait d'un poème généré dans IP Poetry*

Cette technique se rapproche d'une forme de cut-up, popularisé par le grand poète américain William Burroughs. Le contexte de la lecture poétique devient d'autant plus clair, que cette œuvre est généralement exposée dans des lieux dédiés à l'art, orientant donc la lecture de phénomènes comme des manifestations potentiellement poétique. La vocalisation est donc organisée - on pourrait dire « intelligente » - par une énergie poétique.
Quelle est l'identité de cette voix ?
Selon les profils établis ci-dessus, nous sommes dans le cas de l'émissaire numérique : la projection d'une entité numérique qui représente un groupe, une idée, une intention et dont la conscience est attribuée à l'entité qui manipule la voix. Nous ne sommes en effet pas dans la projection de la voix d'un corps (Stephen Hawking), ni dans la vocalisation d'une tâche mécanique (G.P.S).
La lecture poétique est une lecture différente de la voix orale spontanée. Le texte déjà présent n'implique pas un processus de disfluence illustrant l'établissement du discours oral. La simulation de disfluences serait alors un propos anti-réaliste, qui transformerait l'œuvre. Toutefois l'absence de respiration, la répétitivité prosodique, la fluence inexorable, finalement l'émissaire numérique, noyau de l'œuvre, amènent la lecture d'une production technologique

par un outil technologique. La représentation seule de la bouche comme appareil phonatoire total appuie la disparition d'un corps. La bouche est alors un potentiel, une chose qui produit un hors champ aux contours incertains, se déployant autour de l'écran. Ce hors champ est le flux de poésie issu du « réseau », le corps tentaculaire et technologique qui semble couper des flux pour les vocaliser dans une forme poétique.
Cette bouche dans l'écran renseigne toutefois, sans certitude, sur le physique de la voix : un homme blanc entre 30-40 ans. La synthèse formantique travaille donc dans un registre tonal grave (de l'ordre de 120 Hz) avec un débit modéré.
Ici, comme dans l'emploi fait par les hackers d'anonymous, le contexte crée un hors-champ à la fois inquiétant et fascinant.

Listening Post de Mark Hansen et Ben Rubin fonctionne sur la même idée de l'extraction de parties du flux qui constituent le réseau. Depuis 2001, cette installation composée de panneau d'affichage et de huit haut-parleurs regroupe des flux d'informations venus de chat, forum, BBS (bulletin boards system), les découpe, les agence par similitude puis les affiche et les synthétise à l'aide d'un synthétiseur formantique.
Les séquences de sonification sont variables, mais chaque phrase est séparée par un « bip » identifiable. De plus la syntaxe des phrases est extrêmement rigide suite à l'action de regroupement par similitude. Le tout est accompagné par une musique étrangement mélancolique.
Ici, la forme poétique subsiste par l'ampleur de l'installation, l'enveloppement musical, et la cadence des voix synthétiques. L'extrait présente une séquence au sein de laquelle expriment uniquement des phrases commençant par « I AM ».

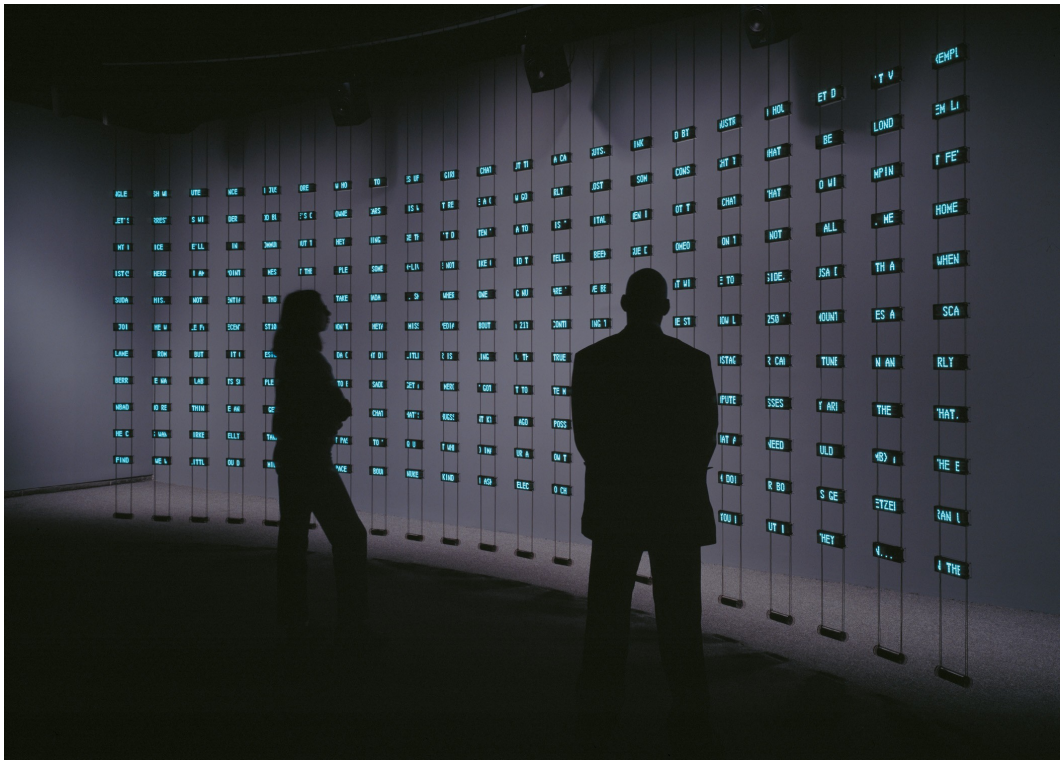

Figure 35 Listening Post de Mark Hansen et Ben Rubin

« I AM

I'M BI

I AM OFF

I AM 18M

I AM

I AM NICE

I AM 26

I AM HOT

I AM OPERATING APES

I AM FREEZING

I AM TODD

I AM 13

I AM GOING

I AM FROM LATVIA

I AM

I AM HERE

I AM BORED

I AM NOT REPEATING »

Extrait d'une séquence de Listening Post

Ici, nous sommes en face d'une quantité infinie d'émissaires numériques. Fenêtre sur un flux continu, la phonétisation permet de donner à entendre, de passer de l'écrit à l'oral, et surtout d'incarner à la fois l'incroyable anonymat de ces flux, la juxtaposition permanente d'informations de degrés totalement différents, et en même temps la potentialité d'identité qui se dégage de ces phrases.
La valeur sémantique de ces phrases n'est pas annoncée comme poétique, seulement la génération humaine est annoncée, ce qui leur confère une puissance communicatrice. Chaque phrase vient potentiellement d'un agent humain différent et chaque phrase est synthétisée avec l'unique même voix. Le paradoxe est donc posé : des milliers de personnes soumis à la même forme, des milliers de pixels individuels. Ici l'exemple prit avec « I AM », pose la question de l'être, autant pour le réseau qui fournit la matière que pour les oreilles impliquées dans le processus d'écoute.

A la différence de IP Poetry, ici le texte pourrait être imaginé comme une forme non définie à l'avance, donc spontanée. Une partie des phrases collectée est issue de réseaux de discussion instantanée. La disfluence aurait été alors plus légitime dans une imitation humaine. Mais l'enjeu ici est d'incarner le réseau à travers cette voix et le montrer comme la route inexorable qu'il est.

Ces deux œuvres expriment la rencontre entre l'être de la machine et l'être de l'humain. L'identité vocale inexistante reflète donc en premier lieu les paradoxes de ce que peut être une identité sur un réseau, une identité numérique, et comment se transmet-elle, ou pas, dans l'information. En second lieu, la disparition du corps engendre un hors champ tentaculaire, qu'il est difficile de définir comme étant composé d'une seule ou de plusieurs entités. Le corps du réseau, des machines, est là, puisqu'il y a phonation. Mais il n'est pas similaire à l'humain, il n'est pas une simulation, ces contours ne sont pas faits d'un tronc, de deux jambes, deux bras et d'une tête ovoïde. Il n'y a pas d'image pour cette voix.

Ici l'être de la machine est absolument opaque. Certes, les syntaxes prédéfinies et les algorithmes sont accessibles, mais l'organisation de telles installations en une entité, digérant des informations générées par l'humain reste difficile à entrevoir comme un simple outil, un simple poste d'écoute. Ici l'émissaire numérique croise donc une notion très rapidement abordée : celle d'une intelligence artificielle. Nous appellerons donc cette sensation étrange d'identité, **l'être numérique.**

Figure 36 L'intérieur décrit par la voix synthétique de Marylin Monroe dans "Marylin" de Parenno

Certaines œuvres, au contraire de Krafwerk ou Gustavo Romano, questionnent l'identité grâce à la synthèse vocale dans un souci de réalisme maximum.

Dans Marylin de Philippe Parenno, installation complexe impliquant une sculpture de polystyrène et une vidéo mettant en scène la voix de Marylin Monroe, la technique de synthèse vocale est poussée dans ses retranchements pour simuler à la perfection la voix de l'actrice décédée. Le film est projeté dans une pièce dont la température est suffisamment froide pour condenser la respiration des spectateurs. Il montre l'intérieur d'une pièce à l'aide de travellings arrière, pendant que la voix décrit ce qui est donné à voir, mais aussi un hypothétique hors-champ. Sera alors révélé une machine imitant l'autographe de Marylin, ainsi que l'équipe technique, pour aboutir à la révélation : il y a un paysage glacé derrière l'écran.

La justification sensorielle de la température par la glace fait alors sens, mais lorsqu'on s'approche du paysage de glace, il n'est en fait que de plastique.

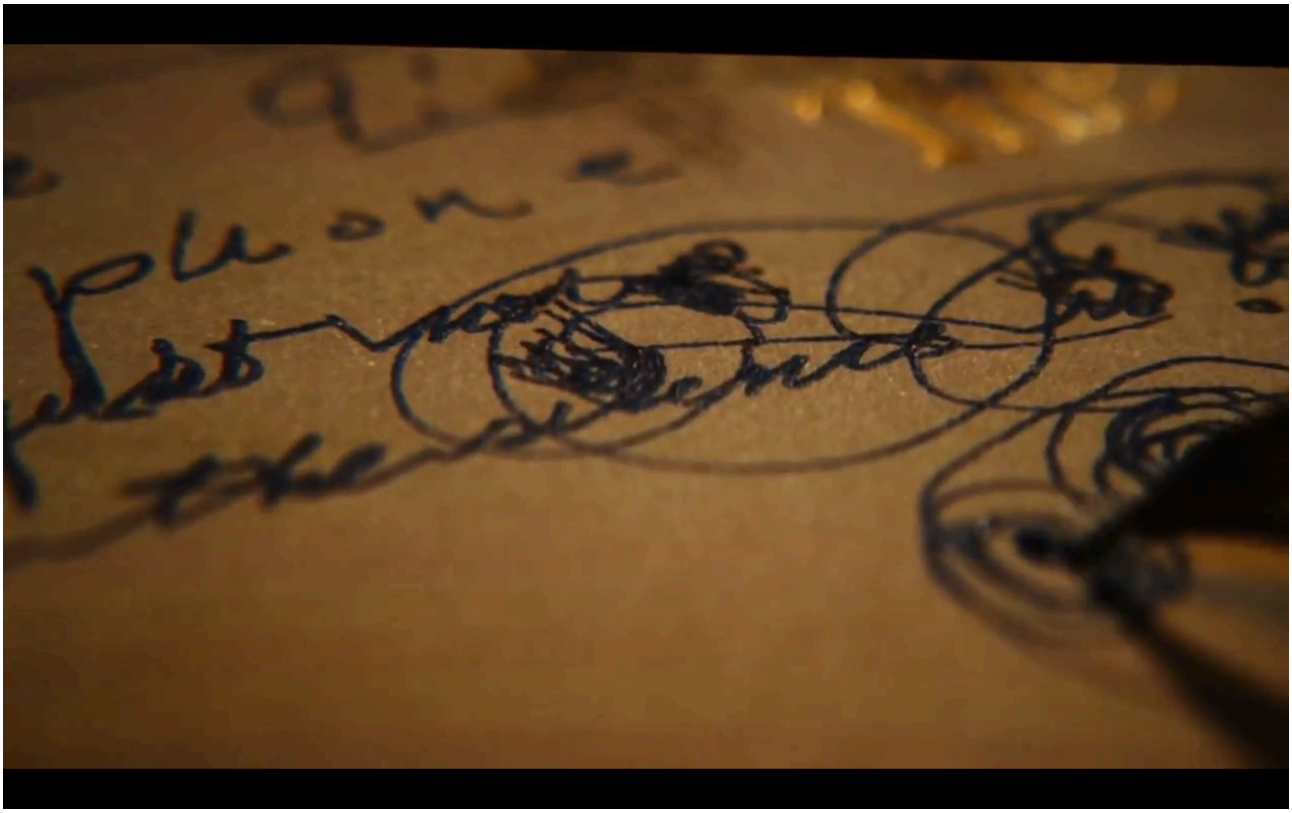

Figure 37 Le robot imitant l'écriture de Marylin dans "Marylin" de Parenno

Nicolas Obin, chercheur de l'IRCAM ayant participé à l'élaboration du TRAX de Flux-Ircam, a crée un corpus à partir d'enregistrements de la voix de Marylin puis travaillé sur des transformations prosodiques pour créer de l'expressivité.

La qualité de cette voix réside dans la richesse timbral, l'absence d'artefacts numériques, l'impossibilité de distinguer les échantillons du corpus. La simulation est réussie puisque l'on retrouve les attributs de la voix de Marylin : un ton légèrement lascif, avec un débit lent et une grande quantité d'air dans la voix, dont la tonalité générale assez basse pour la moyenne des sujets féminins, induit peu de variations de fondamentales.

On observe toutefois de nouveau, l'absence de respirations, de bruits buccaux et toute autre production non langagière. Toutefois, si l'on se réfère à un archétype de voix-off, comme étant une voix hors de l'action filmique, souvent omnisciente, et souvent sans corps (sauf révélation au cours du film), on constate que cette même voix est souvent nettoyée de tout son autre que langagier : diminution des respirations, disparition des disfluences, effacement des bruits vocaux... La voix-off n'est généralement pas un acte spontané de communication orale, il n'y a donc pas de disfluences attendues sans discussion. La voix synthétique de Marylin est donc une simulation parfaite de voix-off.

La réussite de la simulation créée un autre contexte : celui de la comédie. L'expressivité, le dispositif filmique et le rapport à l'actrice, projettent la voix à travers le prisme du jeu. La simulation est donc double ici : celle de l'acteur et celle de la voix synthétique. Sans la dialectique de l'œuvre de Parenno, la disparition de Marylin, le robot signature et le mensonge de la glace en plastique, il ne serait pas possible d'imaginer que la voix off de ce film est générée par un programme. La tromperie des sens, illustrée par la surprise de la glace, rappelle le cabaret Silencio dans Mullholand Drive.

Comment s'incarne alors l'identité de Marylin ?

C'est d'abord travers sa voix synthétique et l'imitation de son écriture, comme des fragments suffisants pour une reconstruction complète. Il serait intéressant de comparer l'analyse des marqueurs biométriques de cette voix synthétique avec celle d'un enregistrement naturel. En théorie, il devrait y avoir concordance. L'identité physiologique devrait être donc respectée. Seulement cette identité physique croise l'image de star créée tout au long de la carrière de Marylin Monroe. La problématique de la simulation est relative à la problématique de l'acteur. Elle traverse autant le cinéma de Bresson, s'incarnant dans une direction faite de modèles et de voix blanches quasiment inexpressives, que le cinéma de Ferrara où ce sont des corps en danger dans des environnements urbains. Ici donc, une composante identitaire est respectée pour mieux rendre palpable le flou entourant la définition de Marylin Monroe, et finalement la complexité de cerner un être, aussi médiatisé soit-il.

La question de l'identité dans cette œuvre est traitée à travers des fragments et des traces, un peu à l'image d'un corpus de samples pour la mise en place d'un synthétiseur MMC. La question de la tromperie concerne autant le monde synthétique que le monde dit « naturel » que l'on pose par facilité comme antinomique. Ici le hors-champ qui entoure la voix, n'est pas le réseau, mais bien l'imaginaire commun qui entoure Marylin.

Ici, la voix de Marylin est une trace. Une trace aux contours parfaits, aussi réussie que l'enregistrement direct de sa voix sur une bande. Seulement ici, le contenu et l'expressivité sont modifiables, comme une trace vivante. Une trace vivante mais unilatérale, sans discussion ni interactivité à la fluence figée. Une trace vivante, dont le corps -appareil phonatoire- n'est pas représentée. Nous l'appellerons **la voix fantôme.**

La voix fantôme ne compose qu'une portion de l'identité vocale : peut-être la faune et la flore qui se construit sur le désert ascétique de Deleuze. La sensation que véhicule la synthèse MMC par l'extraction de marqueurs de la voix, porte aussi cette fragmentation de l'identité, sa potentielle mutation. Le projet VocalID, rencontré en section 7)b)vi), représente d'une certaine manière cette fragmentation. L'extraction de données de filtrage d'un donneur, mélangé aux données sources d'un malade, permet de créer une voix synthétique unique, amenant à une nouvelle identité, fait de deux fragments. Il y a création d'une nouvelle identité qui correspond à un corps hypothétique.

Figure 38 Marilyn

Anne Dorsen s'est aussi beaucoup intéressée aux voix synthétiques acteurs avec **A piece of work.** Durant cette représentation de Hamlet, chaque personnage est incarné par une voix synthétique de facture différente. A l'aide d'une scénographie minimaliste (projection de texte, fumé et lumière), le passage du texte de shakespear à l'oral se fait sans corps, à l'aide synthèse formantique, concaténative et MMC.

Figure 39 A piece of work de Anne Dorsen. Le spectre du père de Hamlet s'exprime

Nous avons croisé le contexte de la lecture poétique avec IP Poetry, de la sonification poétique avec Listening Post. Ici nous sommes dans la performance d'acteur mais à la différence de Marilyn, pour A piece of work l'action se déroule en direct. Cette instantanéité et l'idée de comédie mènent à la recherche d'expressivité. La mécanique toujours inexorable de ces voix par l'invariance des intonations, l'absence de respiration, l'imitation très policée des comportements tonaux mène à une réflexion duale : qu'est-ce qu'incarner un texte dont la rigidité n'a d'égal que son style écrit et son poids culturel ?

Les identités des voix synthétiques qui jouent dans cette pièce sont de l'ordre de représentations figées, des statuts reproductibles dont le visage imparfait est toutefois un visage. De plus, cette pièce (Hamlet) dont l'une des questions centrale est l'existence de Dieu et l'existence de l'Homme prend de l'envergue quand elle est jouée par des entités créées par l'homme.

Je prendrais pour exemple le trio Hamlet (synthèse MMC), Spectre (synthèse formantique), et la voix annonçant les scènes que nous nommerons voix-didascalie (synthèse formantique). La réaction du public face à l'entente de la voix-didascalie, ronde, sans attaque, avec un timbre pauvre lui donnant un air maladroit est celle du rire. La capacité humoristique de ces interventions réside autant dans le choix esthétique que dans l'absurdité d'oraliser une portion de texte sensée rester une balise écrite. De plus la confrontation avec la voix de Hamlet dont l'aspect humain, bien que perfectible, simuler grâce à une synthèse MMC, rend le manque de définition (essentiellement du au comportement dans le haut du spectre) de la voix-didascalie d'autant plus comique. La voix clown, à l'image d'une grimace. La voix que l'on fait en bloquant son larynx, en entrant sa tête dans ses épaules, bloquant certains résonateurs ce qui produit cette sensation de fond de gorge.

C'est alors que la voix du spectre, synthèse formantique ne contenant que la composante « bruit blanc » se met à chuchoter. On se retrouve avec l'émissaire numérique, rattaché à un personnage fantomatique, le père défunt, la présence inquiétante autant que la révélation.

Ces voix jouent donc la comédie, à l'oral, personnifiées par des types d'éclairages ou de la fumée. L'expressivité invariable devient un objet sémantique croisant les problématiques de la pièce et la question du jeu d'acteur.

Si nous revenons à cette idée de performance, de jeu direct, ce que nous attendons c'est une inflexion qui n'est pas prévue dans le texte, qui individualiserait la représentation ou pousserait à trouver tel acteur mieux qu'un autre. Ce que l'on attend du direct, ce n'est pas l'erreur ou l'accident, mais la surprise qui résulte de certains mécanismes qui nous échappent parfois. Ce lâcher prise n'est pas envisagé dans les algorithmes mis en jeu pour synthétiser les voix de la pièce. Pourtant, Anne Dorsen va lentement amener sa scénographie à ce dérèglement, à l'explosion du texte.

Ici l'oralité spontanée n'est pas questionnée, mais l'écartèlement progressif du texte, lui, va s'incarner à travers la synthèse de phrases sans sens apparent entre elles, de plus en plus rapidement. Le débit des voix augmente, leur hauteur aussi, la longueur des phrases rétrécies. La rythmique extrêmement cadencée du déclenchement des phrases affichées à l'écran n'illustre pas l'emballement d'une machine, mais bien au contraire sa maitrise dans des conditions où l'humain est limité.

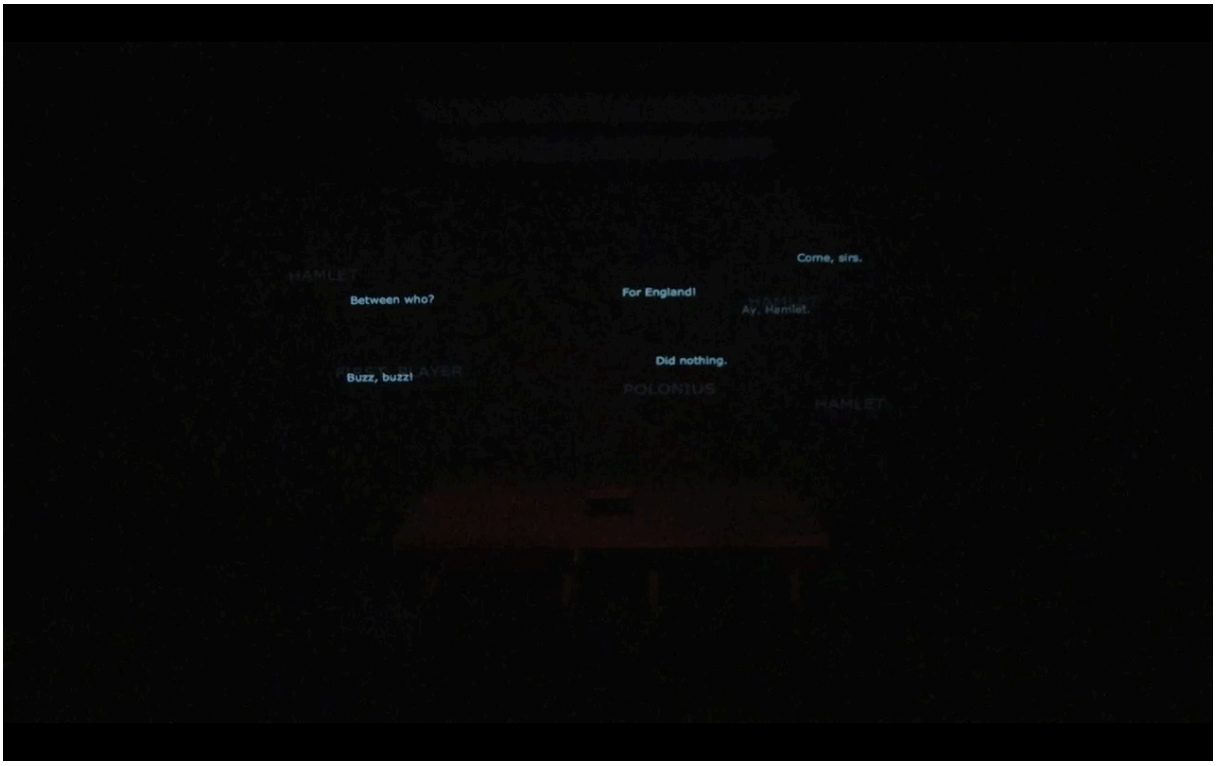

Figure 40 Eclatement du texte synthétisé au court de A piece of work de Anne Dorsen

A la manière de Kraftwerk, cette séquence intervenant 11 minutes après le début de la pièce, magnifie les potentialités de la synthèse en créant des rythmes vocaux sur une durée difficilement gérable par un humain. La logique est au départ visible : ce sont des phrases employant le Ô de l'invocation, synthétisées de la plus longue à la plus courte. Seulement cette logique disparaît, et le sens syntaxique avec. A la différence de Listening Post ou IP Poetry, il est difficile de tirer un concept clair de l'enchevêtrement cadencé d'assertion issu de l'éclatement du texte de shakespear. Peut-être une génération nouvelle à partir du texte à chaque représentation ? Ce qui est sûr c'est que l'absence d'un cadre conceptuel met en difficulté la construction logique. On se rapproche alors d'un dérèglement de l'ordre du trouble du langage. Comment troubler le langage d'un programme qui n'a pas d'appareils phonatoire ? Et selon quel modalité ?

Ici se développe une sensation que je n'avais pas envisagée. Le hachage de plus en plus petit des syntaxes, l'augmentation du débit, se perçoit autant comme une surprise que comme une règle : les intervalles et les ruptures de tons s'identifient comme faisant partie d'un algorithme, comme étant l'agent d'un accident organisé. Ce dérèglement n'est donc pas le lâché prise ou la tension d'un bègue. Il n'y a pas de ruptures.

Sans répondre à cette question Anne Dorsen projette ces problématiques encore plus loin dans une troisième phase où munie d'une oreillette, elle va imiter la prosodie inexorable de ces synthétiseurs. Ici, les phrases ont disparu ne laissant plus que des mots, dont les intonations sont finalement difficiles à simuler pour un humain.

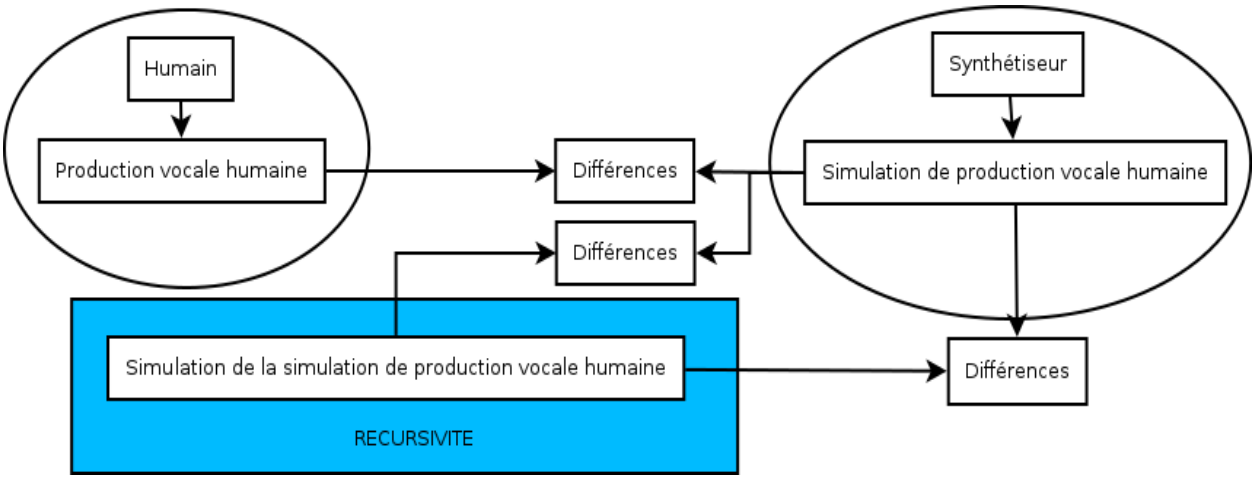

Figure 41 La récursivité dans A piece of work de Anne Dorsen

Dans sa scénographie son corps est au début absent. Seul un isoloir est éclairé. L'illusion d'une voix synthétique s'étiole très rapidement lorsque des respirations font leur apparition. Anne interprète alors un texte qui lui est dicté par les synthétiseurs qui jouaient auparavant la comédie. Elle renverse donc le processus de simulation qui a mené à la voix synthétique. Elle prélève les informations d'intonations, de débits, de rythme et les appliques à son propre appareil phonatoire pour phonétiser le texte qu'elle entend. Le résultat est une imitation imparfaite de par les limitations technologiques du corps humain. Ce retournement produit alors les erreurs que l'on attend d'un jeu en direct, d'un corps qui travaille, qui se fatigue et s'épuise. Il y a alors des arrêts, des reprises. Leur cause est toutefois inconnue : peut-être est-ce le texte synthétisé qui à cette forme. La syntaxe s'effrite de la même manière que dans la séquence synthétique, menant à des mots seuls, puis des syllabes.

Figure 42 L'isoloir personnifiant le corps de la voix indéfinie de Anne. Et Anne Dorsen au premier rang dans A piece of work

La récursivité que met en place Anne Dorsen amène à se questionner sur la notion de reproductibilité et des différences que cela génère de manière inévitable. De plus, la personnification de la voix par un objet ou une matière fonctionne autant que l'un corps. La voix ne porte pas nécessairement l'information d'un corps, parce qu'elle est aussi une construction linguistique, sémantique et sémiologique. C'est ce qui se ressent par le jeu de l'isoloir et la révélation du corps de Anne, et de la différence d'interprétation entre un texte sans expression corporelle ou expression du visage, puis avec.

Le poète sonore Anne James Chaton s'est lui aussi illustré dans la répétition de syntaxe et la lecture de liste, de trace, d'informations factuelles... Dans décade, il raconte l'itinéraire d'un homme à travers les lieux, les heures, les rencontres et les achats qu'il fait. Sans être concerné par la synthèse vocale, il s'illustre dans la mise en place d'un mécanisme inexorable à l'image d'un quotidien de faits, mis bout à bout.

Dans décade, la respiration a disparu. L'absence de variations et la répétition tonale invariante durant les 10'34 de la pièce se confronte aux timbres extrêmement particulier de Anne James Chaton ; particuliers par son assise dans les graves. A l'image de la récursivité de Anne Dorsen, Anne-James cherche à questionner ce qu'est le « dire » et quelle est l'identité de celui qui « dit ». Il raconte ici l'itinéraire de cet homme à la troisième personne dont on imagine l'histoire, donc un comportement potentiel et peut-être un physique, relatif aux traces qu'il laisse et aux archétypes que nous contenons.

A l'aide de PRAAT j'ai voulu vérifier la rigidité de la prosodie de la voix de Annes-James dans la pièce décade. J'ai donc fait un relevé de la variation de fondamentale à 1 minutes puis à 6 minutes, et on constate une similitude presque totale :

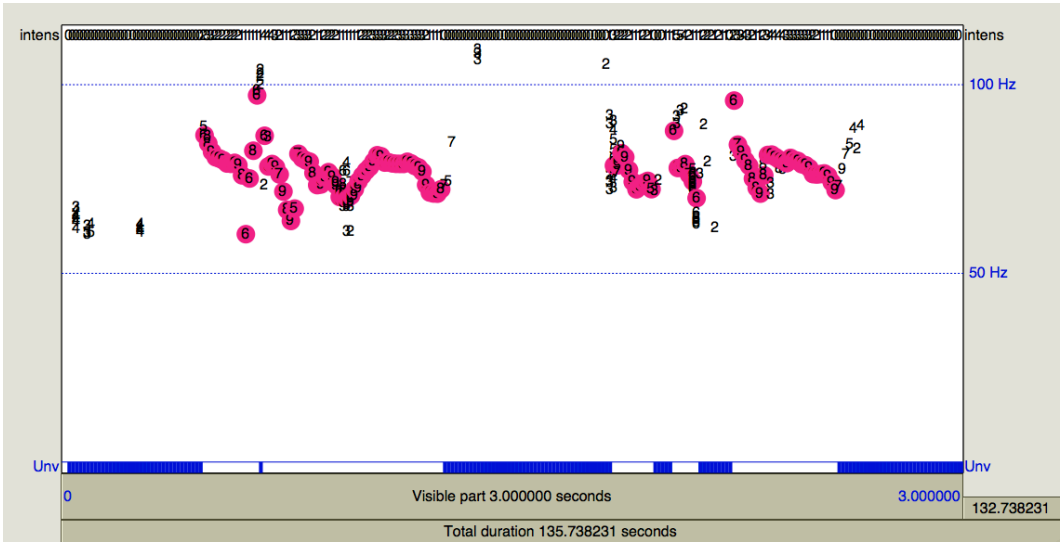

Figure 43 Analyse de la fréquence fondamentale de deux phrases consécutives de Décade de Anne James Chaton à 1'00

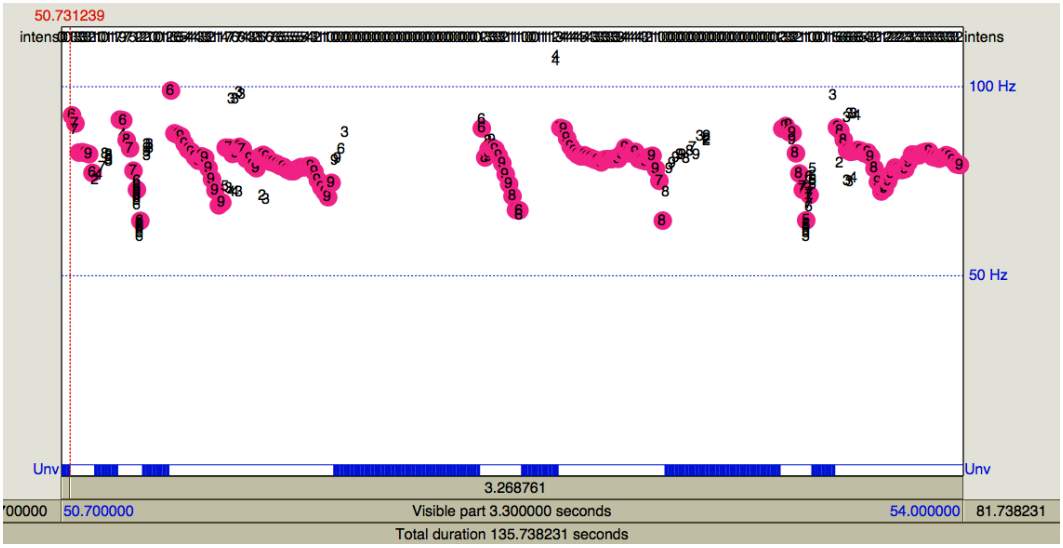

Figure 44 Analyse de la fréquence fondamentale de trois phrases consécutives de Décade de Anne James Chaton à 6'00

C'est d'ailleurs certainement ce « presque total » qui fait que la voix de Annes-James est identifiée comme humaine. Ce sont ces imperceptibles micro-variations qui accompagnent les deux mouvements tonaux descendant répétés, observable sur les graphiques.

La durée des pauses est-elle aussi respectée durant toute la pièce :

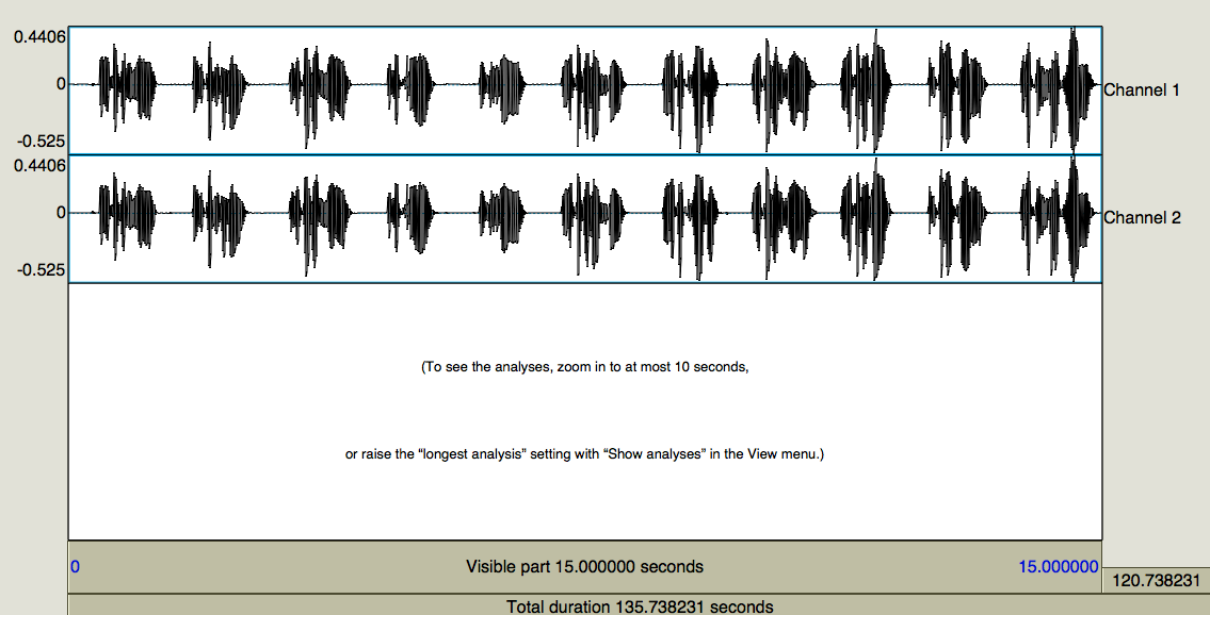

Figure 45 Observation de pauses silencieuses. Représentation graphique des 15 premières secondes de Décade de Anne James Chaton

A travers les œuvres rencontrées, plusieurs points importants se sont révélés dans la manière d'envisager l'identité dans la synthèse vocale et son rapport à la disfluence.

- Toute voix synthétique implique un corps. Ce corps peut-être conceptuel ou non, mais il est le lieu de production de cette voix à l'image de l'appareil phonatoire humain. Cette relation implique elle même une identité vocale. Cette identité vocale, quelle que soit sa qualité est d'apparence reproductible. Toutefois, la complexité et la variation des contextes d'utilisation produisent des individuations possibles.

- La simulation est un mécanisme bilatéral qui objective les modèles qu'elle produit. Ce phénomène récursif augmente les individuations possibles.

- Il n'y a pas de disfluence si il n'y a pas d'oralité spontanée. Il n'y a pas d'oralité spontanée si il n'y a pas une intelligence. Dans notre cas, l'intelligence devra être artificielle.

- Il est possible de personnifier une voix par l'action simultanée de l'entendre et du voir. Une simple lumière synchronisée permet de personnifier l'hypothétique corps de la voix synthétique.

- Le dérèglement de la voix synthétique selon les modalités d'un algorithme se perçoit rapidement comme une règle de l'algorithme. Pour donner la sensation d'une désorganisation du langage, il faut créer une rupture.

iii. La représentation des machines parlantes dans l'audiovisuel :

De nombreuses entités non-humaines, souvent électroniques, souvent machines, se sont vues représentés : K2000, R2D2, Wall-E, H.A.L, tout type d'androïdes (A.I, Alien, Blade Runner...) ... Pourtant dans aucun cas, un dispositif n'a impliqué l'utilisation d'une voix synthétique. De plus, on constate que malgré la démocratisation de l'image de synthèse, les voix des personnages animés restent, elles, humaines. Il est par contre imaginable que le secteur du jeu-vidéo soit en attente de techniques de synthèse plus précises, notamment en terme d'interactivité.

L'expressivité est au centre du jeu d'acteur qui compose ce que l'on pourrait appeler « le cinéma majoritaire ». C'est un cinéma qui tend à créer l'illusion que ce qui se déroule à l'écran et au son est une potentielle réalité. Cette illusion traverse autant la fiction que le documentaire.

Cette attente entre donc en conflit direct avec les difficultés technologiques de la gestion de l'expressivité dans la synthèse vocale (difficulté à cerner les stratégies de gestion de la fondamentale et de la prosodie en générale tout en conservant une qualité timbral). Nous avons toutefois vu qu'il serait possible de créer une illusion parfaite (cf Marilyn de Parreno). Mais en ce cas quel est l'intérêt économique et pratique d'impliquer un si grand travail pour simuler à la perfection ce qu'un acteur peut faire à moindre effort (en comparaison à la contrainte technologique).

On peut donc statuer sur deux éléments : actuellement le coût d'une synthèse dont l'illusion est totale est moins rentable que le coût d'un acteur. Lorsqu'un tel rapport sera renversé, il faudra réenvisager les typologies d'identité vocales car elles se multiplieront. Le second point réside dans la méconnaissance des possibilités esthétiques de la synthèse vocale et les

questions d'identité qui sont abordées dans ce mémoire. Si lors d'une réalisation hypothétique, un outil de création sonore de qualité était mis à disposition d'un réalisateur, je ne doute pas que l'opportunité créative de créer entièrement la voix d'un corps tout aussi synthétique ne soit pas attractive.

On peut toutefois s'attarder sur un point intéressant au sein du cinéma d'animation. Les images de synthèses n'ont pas de voix synthétiques sans que cela interpelle. L'illusion cinématographique est conservée par la synchronicité de la voix et des mouvements du personnage. Sans entrée dans des questionnements déjà extrêmement poussé par les théoriciens tel que Michel Chion, on constate que dans l'audiovisuel, il n'y a pas de rapports directs entre la matière d'un corps sonore et le son entendu (cf doublage).

Dans les films en images de synthèse, le corps est porteur d'une série d'indices identitaire que l'on espère retrouver dans les traits physiques, linguistique voire sociologique. Si l'on voit un perroquet rouge et bleu fait de pixels, de petite taille, on s'attend à une voix aigüe, un vocabulaire limité dut à l'archétype du perroquet qui répète, et un tempérament peut-être assez extraverti de par les couleurs qu'il porte. C'est dans ce jeu d'attente que l'on peut créer toutefois de la surprise. Il n'est par contre pas attendu d'entendre des indices portant à croire que la voix du perroquet est synthétique. Il n'est pas attendu de révéler les mécanismes de l'illusion.

Il n'y a donc de place légitime pour aucune des identités citées auparavant excepté pour la « voix fantomatique » au sein du paysage audiovisuel conventionnel.

Sortons toutefois d'un cadre cinématographique figé narrativement par le déroulement de la bobine ou du fichier, pour s'orienter vers le Jeu-Vidéo, ou la spontanéité est au centre de ce que l'on appelle le game-play.

Le game-play c'est la façon d'aborder un jeu, d'y progresser, d'y rester, de le terminer. Le game-play représente le champ des possibles d'un joueur et les modalités d'interactions avec l'univers qui l'entoure. Tout ceci est tendu vers un but ou une absence de but ce qui créer alors un jeu-vidéo. Ce dernier implique quasiment exclusivement des matières visuelles synthétiques. Pourtant comme le cinéma d'animation, les voix des personnages habitants les jeux sont pratiquement tous joués par des humains. Or si au sein du Cinéma, l'utilisation de la synthèse est un atout esthétique de création, dans le jeu-vidéo, la synthèse vocale devrait être un pilier pouvant développer l'expérience de jeu en augmentant l'interactivité.

L'interactivité vient de l'impossibilité de prédire le geste du joueur. Cette interactivité peut-être variable au sein des jeux, mais elle est la condition sinequanone de la notion de jeu. Les personnages habitant les jeux-vidéos sont souvent amenés à s'exprimer vocalement pour aider au déroulement du jeu ou simplement créer la sensation d'un univers réaliste, discuter, blaguer, divertir... L'imprévisible joueur amène donc des situations de spontanéités, qui se retrouvent pourtant limitées par des enregistrements de voix figés.
Nous nous trouvons donc dans un cas similaire à l'installation proposé en partie pratique : une intelligence artificielle qui va tenter de simuler les modalités de la discussion orale spontanée. Je vois donc le jeu vidéo comme l'un des secteurs d'activités potentiellement intéressés par l'utilisation des disfluences. De plus, d'un point de vue narratif, les tentatives de profilages précédents des disfluences peuvent aider à créer des personnages d'autant plus cohérents.

d. Les secteurs d'activités potentiellement intéressés par la disfluence

Nous avons donc observé que les cas nécessitant la simulation de disfluences étaient les cas de discussion orale spontanée ce qui implique dans notre cas l'utilisation d'une intelligence artificielle ou d'un humain pour piloter le synthétiseur vocal.

On peut donc imaginer que l'industrie du jeu vidéo serait intéressée dans plusieurs cas :

- Génération de personnages secondaires variés au comportement vocaux plus originaux, à l'expressivité plus fine ce qui éviterait d'engager des acteurs.
- Expérience de jeu plus immersive. En synthétisant le texte au cours du jeu, il est possible d'imaginer générer les dialogues d'un scénario en train de s'écrire, et de l'oraliser avec qualité grâce à la gestion de disfluences. On pourrait imaginer ce dispositif dans des jeux qui travaillent sur la sensation de liberté (GTA-Like ou SIMS-Like).

Le secteur des télécommunications automatiques et autres serveur téléphonique pourraient gagner en efficacité de vente si, grâce à un système interactif, les voix de vendeur ou d'informateurs s'adapter au locuteur à l'autre bout du fil. Par une analyse de l'agent humain de l'un des côtés du combiné, il serait possible d'adapté le débit, de gérer des pauses pour créer des tensions ou au contraire des créer des situations impliquant de la compassion par l'utilisation de tous les champs de disfluence, ce qui diminuerait l'antipathie envers ces miroirs trop parfaits et augmenterait la potentialité de réussite de la mission commerciale ou publicitaire.

Mais finalement, l'enjeu réel de ce mémoire concerne la création au sens large. L'enjeu serait de proposer une manière d'intégrer la disfluence au processus de synthèse vocale, d'abord à travers un programme Open-Source puis de le montrer à l'aide d'une installation qui impliquerait la simulation d'une situation de discussion orale spontanée. Cela permettrait alors de démocratiser ces problématiques et de potentiellement les rendre accessibles au sein de toutes sortes de projets : Installations, Musique, Cinéma, Recherche...

3. Partie pratique

8. Partie pratique

Grâce à MaryTTS, à sa communauté et à la documentation accessible, nous allons pouvoir mettre en place, à l'aide du SSML, des mécanismes de disfluences selon les règles établies auparavant.

Il faut donc trouver le moyen de modifier un texte pour y insérer des balises altérant la fluence de la voix synthétique. Pour ce faire, je me suis orienté sur les conseils de Greg Beller vers le python[40]. Le python est un langage de programmation interprété (c'est-à-dire qu'il n'y a pas de compilateur mais un interpréteur, comme par exemple le PHP) intuitif et rapide qui permet de manipuler simplement des librairies systèmes et d'automatiser un certain nombre de tâches. On peut donc imaginer que grâce à Python, nous allons modifier automatiquement des textes pour y insérer des balises SSML simulant pauses ou amorces...

Nous avons toutefois vu que pour amener une situation logique de disfluence, il fallait une situation d'oralité spontanée. Nous allons donc utiliser un agent conversationnel AIML[41]. Les AIML sont des robots qui permettent de gérer plus ou moins bien une conversation avec un agent humain. Ils prennent du texte en entrée et génèrent un texte en sortie.

Grâce à un langage, lui aussi sous forme XML, il est alors possible d'orienter des discussions, de gérer toutes sortes de réponses et surtout de façonner l'identité d'un robot. Ils prennent en compte des processus d'apprentissage encadrés ou non. Nous nous servirons d'un cerveau AIML français au départ établi par Jean Louis Campion et Siewlan Tan dérivé de ALICE[42] (Artificial Linguistic Internet Computer Entity), robot AIML développé par Richard Wallace, ayant remporté le prix Loebner en 2000, 2001 et 2004. Ce prix est attribué en fonction des résultats d'un test de Turing : établie par Alan Turing, ce test sensé tester la qualité d'une intelligence artificielle consiste à mettre un humain N°1 en relation écrite avec une intelligence artificielle et un autre humainN°2. Il est alors établie combien de temps met l'humain N°1 à distinguer qui est l'intelligence artificielle et qui est l'humain N°2.

Cet AIML contient la double problématique du mémoire : il sera le générateur d'un texte qu'il faudra oraliser, mais aussi le générateur de la spontanéité que la disfluence requiert.

[40] https://www.python.org/
[41] Artificial Intelligence Markup Language
[42] Il est possible d'essayer ALICE ici : http://alice.pandorabots.com/

La dernière question est d'imaginer quel sera le mode de communication entre un utilisateur et cet AIML, et comment sera donnée à entendre la voix synthétique ? Faut-il la personnifier ? Comment gérer l'intégration des disfluences ? Comment seront-elles définies ? Faut-il rendre visible ce processus ?

a. Présentation de l'installation

L'installation prendra la forme d'une conversation de l'ordre de l'intime (uniquement deux entités engagées) entre un humain et un être de synthèse.
L'enjeu de Dennys est d'offrir, à travers cette conversation, un questionnement sur certaines frontières théoriques de l'identité. Elle y est questionnée selon plusieurs axes :
La seule personne avec qui l'utilisateur converse, c'est lui-même. L'algorithme de discussion s'oriente selon les choix de l'utilisateur induits par les saisis claviers. On se retrouve très rapidement dans une discussion introspective. Les premiers agents conversationnels avaient d'ailleurs pour but de remplacer la séance de thérapie que dispense un psychanalyste.

Dans ce rapport introspectif, l'aspect synthétique de Dennys sera affiché par ses caractéristiques vocales et la scénographie de l'installation elle-même. Il y a donc une zone de friction populaire ici : ce qui est synthétique est souvent catégorisé comme irréel. Or que ce soit le monde numérique comme le monde synthétique ou virtuel, il ne peuvent être mis en opposition à la réalité, car ils en font tout simplement partie. La discussion avec Dennys, si elle a lieu, va nécessairement produire quelque chose (réflexion, ennui, joie, dérangement, rigolade...) qui est par nature réel.

De plus Dennys est un être qui semblera intelligent pour deux raisons : il est capable de mener une discussion et il semble posséder une identité vocale (voix nasillarde, peu d'articulation, accent canadien, disfluences ...). La synthèse de la voix de Dennys fera appel aux disfluences pour simuler la sensation de spontanéité, renforçant l'idée que Dennys construit ses réponses sur le même mode de fonctionnement que le cerveau humain, simulant les mécanismes de l'oralité.

C'est donc dans cette relation ambivalente entre un programme qui se comporte comme un humain sans en avoir l'apparence et un humain qui tente de converser avec un programme, entre un effort intellectuel pour produire une phrase et un effort virtuel pour produire une réponse, entre un texte et une voix orale que va se poser la question cible : « Qui est Dennys ». Je l'appelle question cible car le but de Dennys est de dépasser la question « Qu'est-ce que Dennys » pour atteindre cette limite entre la chose et la vie qui apparaît floue

dans le cadre de simulation complexe. J'applique le questionnement identitaire auquel j'ai soumis toutes les voix de synthèses rencontrées depuis le début de ce mémoire à Dennys, en espérant que l'expérience de discussion produira un questionnement identitaire pour l'utilisateur sur ce qui constitue son désert ascétique (cf définition de l'identité).

Le projet qui était initialement celui de pouvoir converser oralement avec l'être de synthèse s'est vu réorienté pour plusieurs raisons. Il aurait tout d'abord fallu imaginer un logiciel de reconnaissance vocale suffisamment performant pour transcrire la voix de l'utilisateur. J'ai donc entamé une série de tests sur des solutions freeware peu convaincantes (comme NaturalReader) avant de tester Dragon Naturally Speaking développé par la société Nuance. Le résultat, bien que plus convaincant, ne résolvait toutefois pas deux problèmes fondamentaux : très rapidement, des erreurs apparaissaient lors de l'utilisation de mot d'argot ou lors d'une prononciation un peu lésée dont le niveau d'intelligibilité devenait critique pour le logiciel. Mais aussi et surtout, il n'était pas possible d'envisager pour cette installation un logiciel transcription ne transcrivant pas les disfluences de l'utilisateur.

Cette difficulté fit donc apparaître un paradoxe fécond pour le dispositif de l'installation : l'utilisateur humain communiquera avec la machine à l'aide du clavier. Le clavier servira naturellement d'une restriction de langage, celui sans disfluences : celui de l'écrit.

L'écoute se fera à l'aide d'un casque fermé. Au départ l'idée était d'utiliser une enceinte, mais l'importance de créer un rapport intime avec l'agent conversationnel nécessité d'empêcher l'expérience de groupe et de focaliser l'écoute d'un seul utilisateur d'où le dispositif casque. Le multiplieur permettra toutefois d'écouter les résultats de la conversation sans pour autant nécessairement perturber la discussion en train de se dérouler, à l'image du combiné d'écoute supplémentaire qu'il existait sur les vieux téléphones.

Cette installation portera le nom de Dennys, qui est le nom de la voix synthétisée à l'intérieur de MaryTTS.

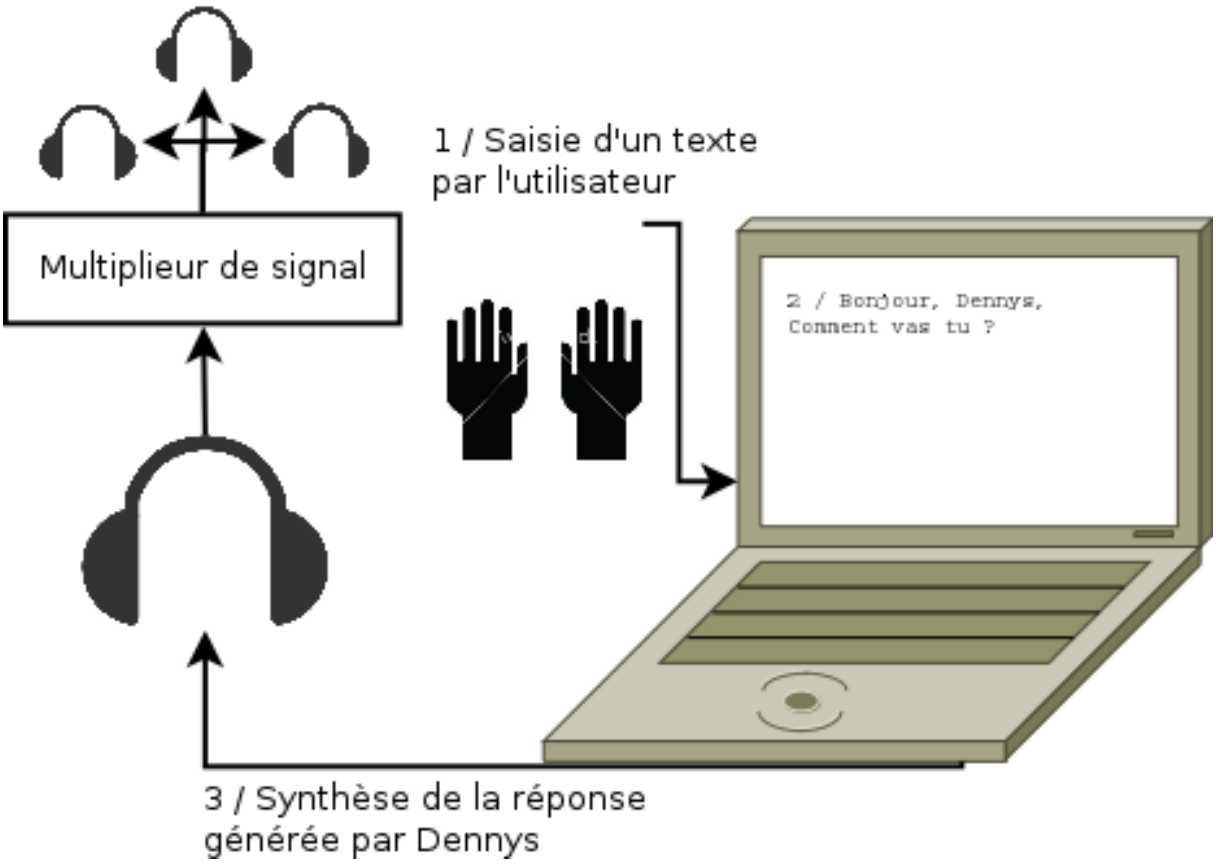

Figure 46 Descriptif de l'installation

b. Fonctionnement technique de Dennys

Le fonctionnement technique de Dennys réside sur la communication entre l'agent conversationnel, le synthétiseur MaryTTS, et un script en python organisant l'ajout de disfluences.

L'agent conversationnel et le synthétiseur fonctionnent à l'aide de serveurs avec lesquels le script en python communique. Ce dernier gère aussi les saisis claviers de l'utilisateur. Toutes les informations textuelles seront sauvegardées et modifiées à partir de fichiers textes. L'ajout de disfluences se fera à l'aide des balises SSML décrites auparavant, qui seront ensuite transformées en fichier son, dont la lecture sera automatisée par le script en python.

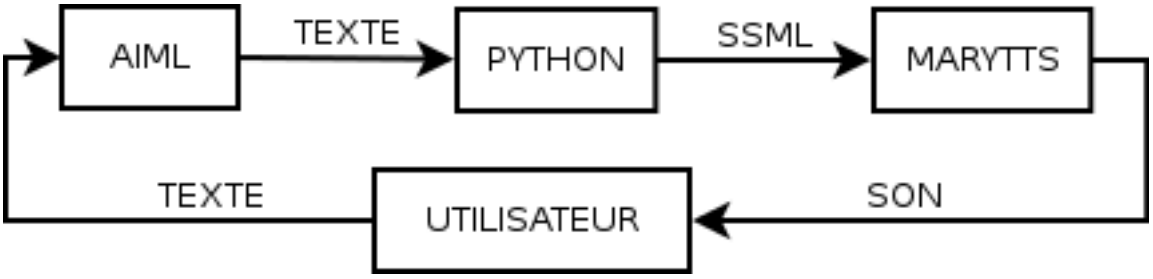

Figure 47 Synoptique 1 général de Dennys

i. Le programme en Python

Le script en Python est la colonne vertébrale du programme. Il va organiser l'aiguillage des textes, le déclenchement de la synthèse et les durées au sein du programme. L'idée est de fonctionner à l'aide de fichier texte qui stockerons les saisis de l'utilisateur et les réponses générées par l'AIML, en vue d'une transformation qui mettra alors à jour les fichiers texte préalablement créés. Les fichiers se nommeront u+N° et m+N° où « u » signifie utilisateur, « m » machine, et N° est un chiffre incrémenté à chaque nouvel échange.

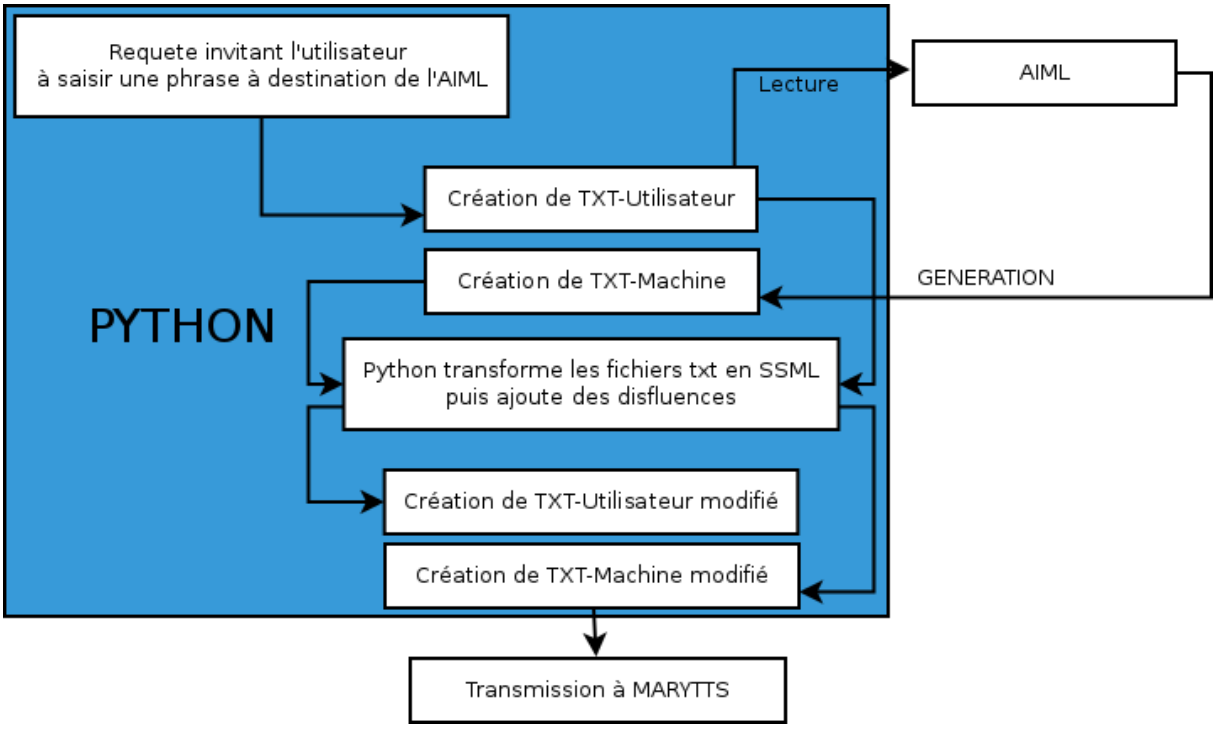

Figure 48 Fonctionnement général du script en python

ii. Le serveur MaryTTS

Le logiciel MaryTTS fonctionne à l'aide d'un serveur avec lequel l'utilisateur peut converser de différentes manières. Dans notre cas, nous utiliserons des requêtes http avec lesquels nous nous connectons sur le port 59125 du réseau local, port dédié à MaryTTS. Nous pourrons alors transmettre les informations suivantes :

- Le texte à synthétiser contenu dans la variable « phraseTraduire »
- Le format d'entrée : SSML
- Le format de sortie : Audio
- La langue : français
- La voix choisie : enst-dennys-hsmm

Voici la requête transmise à l'aide de la librairie requests :

```
UserResponse =
requests.get('http://localhost:59125/process?INPUT_TEXT='+phraseTraduire+'&INPUT_TYPE=SSML&OU
TPUT_TYPE=AUDIO&AUDIO=WAVE_FILE&LOCALE=fr&VOICE=enst-dennys-hsmm')
```

Figure 49 Requête python à destination du serveur de MaryTTS

iii. Le choix de la voix de Dennys

En vue d'obtenir des possibilités de modulations prosodiques permettant la bonne synthèse des paramètres de disfluences tout en conservant un timbre vocal de qualité, je me suis orienté vers des voix utilisant les modèles de Markov caché.

On trouve toutefois peu de voix synthèses françaises accessibles en Open-Source. J'ai donc contacté Florent Xavier, auteur d'un mémoire d'intégration au français de MaryTTS ayant développé deux voix en français basé sur la technologie MMC : Florent et Catherine. Leur facture était étrange: le volume de ces voix semble varier, de plus le filtre appliqué au bruit blanc à un comportement étrange ce qui produit des artefacts numériques.

Il m'a toutefois orienté vers la voix de Dennys, une voix Canadienne de technologie HSMM, développée à l'Ecole Nationale Supérieure des Télécommunications. HSMM signifie Hidden Semi-Markov Model. Le terme « semi » implique que la probabilité de transition d'un état vers un autre dépend aussi du temps de durée du phénomène, alors qu'il est indépendant dans un modèle de Markov caché classique.

Dennys remplit des conditions d'intelligibilité, tout en conservant des possibilités de modulations prosodiques. Bien que la qualité du timbre ne soit pas optimale, il jouit d'une

cohérence invariable au cours de la synthèse d'un court texte comme d'un long. L'accent Canadien de Dennys servira de plus d'élément identitaire fort de cette voix. Il s'agira de créer chez l'agent conversationnel un lien logique entre ces indices sociaux vocaux et une potentielle histoire personnelle.

iv. Le fonctionnement du serveur AIML

A l'image du serveur MaryTTS, nous converserons avec l'agent AIML à l'aide de requêtes http sur le port 2001 du réseau local, qui transmettrons le texte de l'utilisateur contenu dans la variable « phrase »
Voici la requête transmise à l'aide de la librairie requests :

```
questionAlice= requests.get('http://localhost:2001/?botid=Dennys&template=alice&text='+phrase)
```

Figure 50 Requête python à destination du serveur AIML

Nous avons maintenant tous les moyens pour mettre en œuvre l'échange entre un agent AIML français et un utilisateur humain. Les questions qui subsistent sont celles de la génération et l'intégration de disfluences, le profil de l'agent AIML et l'interface graphique qui sera montrée à l'utilisateur.

c. Modélisation des disfluences et des bégaiements et enregistrement des séquences.

Comme nous l'avons au chapitre (la question des profils), il n'est pas possible de partir sur des profils préexistant pour générer les disfluences ou les bégaiements. Nous allons créer des séquences en interprétant les règles établies dans notre étude de ces phénomènes. Ces séquences seront enregistrées dans un fichier texte. Il sera ensuite possible de relier la séquence au fichier son de voix de synthèse, ce qui permettra alors d'avoir une annotation précise des paramètres de la disfluence en question, ce qui pourra peut-être conduire à un corpus de voix synthétiques disfluentes intéressant.
Nous allons définir les règles pour chacune des disfluences puis en remontant progressivement, nous définirons le fonctionnement de la génération de séquence et son intégration au programme en python.

i. Les règles individuelles pour chaque disfluence

Pour chaque disfluence, nous allons définir sa notation et les paramètres avec lesquels il sera possible de la modifier. Toutes ces informations nous permettront d'une part de créer une fonction python équivalente et de normaliser la notation dans le fichier texte de sauvegarde.

- Pause silencieuse :

  Annotation : PS

  Durée : 200ms<x<2000ms

  Position syntaxique d'intervention : toutes sauf avant et après un « . »

- Allongement syllabique :

  Annotation : AS

  Position syntaxique d'intervention : toutes sauf avant un « . »

  Position syllabique : Début ou fin de mot

  Durée : x<1800ms

- Morphème Spécifique :

  Annotation : MS

  Dictionnaire

  Durée : même règle que l'allongement syllabique, 600<x< 1800ms

  Position syntaxique d'intervention : toutes sauf avant un « . »

- Conjonction d'appui :

  Annotation : CA

  Position : Interviens uniquement après une disfluence (toute sauf elle même)

  Dictionnaire

- Amorce complétée :

  Annotation : AmC

  Position : mot d'au moins deux phonèmes. Partout sauf avant une ponctuation. Il n'y a pas de problème théorique d'accessibilité lexicale avant une ponctuation car elle offre justement un temps de répits. Une amorce complétée avant ponctuation entre dans le domaine du trouble du langage.

  Fenêtre : nombre de mots constituant l'amorce

  Nombre de répétition : Combien de fois l'amorce est répétée avant d'être complétée 1<x<3.

  Pause inter-répétition : A l'image d'une pause silencieuse, 200ms<y<2000ms

- Amorce modifiée :

  Annotation : AmM

  Cible : Dictionnaire de cibles similaire aux autocorrections pouvant simuler le remplacement à la même place syntaxique

  Fenêtre : nombre de mot constituant l'amorce

  Nombre de répétition : Combien de fois l'amorce est répétée avant d'être modifiée 1<x<3.

  Pause inter-répétition : A l'image d'une pause silencieuse, 200ms<y<2000ms Chaque pause doit varier au minimum de 12ms avec la précédente.

- Amorce inachevée :

  Annotation : AmI

  Cible : Dictionnaire contenant des amorces inachevées prédéfinies : Elles seront ensuite choisies soit par analogie si il y en a une dans la phrase générée, soit par tirage au sort.

  Nombre de répétition : Combien de fois l'amorce est répétée avant d'être modifiée 1<x<3.

  Pause inter-répétition : A l'image d'une pause silencieuse, 200ms<y<2000ms Chaque pause doit varier au minimum de 12ms avec la précédente.

- Répétition :

  Annotation : R

  Position : Toutes sauf avant une ponctuation

  Fenêtre : nombre de mots constituant la répétition

  Nombre de répétition : 1<x<3.

  Pause inter-répétition : A l'image d'une pause silencieuse, 200ms<y<2000ms. Chaque pause doit varier au minimum de 12ms avec la précédente.

- Autocorrections :

  Annotation : Aut

  Cible : Dictionnaire de cible similaire aux amorces modifiées pouvant simuler la correction à la même place syntaxique.

ii. Les dictionnaires :

- Morphème spécifique
  - ⇒ Heu
  - ⇒ Hm
  - ⇒ Hé
  - ⇒ ...
- Conjonction d'appui
  - ⇒ Bon
  - ⇒ Enfin
  - ⇒ Alors
  - ⇒ ...
- Amorces Modifiée / Autocorrection
  - ⇒ Négation -> Suppression, puis réapparition
  - ⇒ Descripteurs de lieux : à, dans, sur, dedans, de
  - ⇒ Descripteurs de temps : au moment, pour,
  - ⇒ Conjugaison : J'ai <-> J'avais <-> J'étais <-> J'aurais <-> Je suis
  - ⇒ ...
- Amorce inachevée
  - ⇒ « Si j'avais »
  - ⇒ « Tu me diras »
  - ⇒ ...

iii. Les scénarios de séquence de disfluence :

L'un des enjeux de la simulation des disfluences est de conserver la discrétion de ce phénomène. Il faut donc établir des scénarios balisés pour contrôler les apparitions d'amorces et autres pauses.

- Aucune disfluences ne doit être répétées
- Pas plus de 3 disfluences consécutives
- La conjonction d'appui doit succéder directement à une disfluence différente

De plus il est possible de combiner des disfluences (i.e : « **J'ai : - J'ai :** - J'ai appris à parler » est ici l'association d'une amorce complétée et d'un allongement syllabique).

- Il est possible de combiner uniquement deux disfluences entres elles. Leur notation se fait à l'aide du symbole + (i.e « **J'ai : - J'ai :** - J'ai appris à parler » se note AmC+AS). Elles constituent alors une seule disfluence.

- Chaque disfluence est associée à un chiffre. Pour établir les séquences, il y a un tirage au sort dont le résultat est soumis aux règles précédentes. Ce tirage ne répond pas à des probabilités déterminées puisque les études précédentes ont montré des quantités de disfluences très variables en fonction de chaque sujet.

  ⇒ PS =1 / AS = 2 / MS = 3 / CA = 4 / AmC = 5 / AmM = 6 / AmI = 7 / R = 8 / Aut = 9

- Selon Honas et Schultz, 50% des locuteurs disfluent au moins 1 fois par énoncé, les autres disfluent moins d'une fois par énoncé. Il faut donc établir un paramètre d'activation des disfluence dont la probabilité est de ½.

iv. L'intégration aux énoncés

Si l'on estime que la disfluence augmente avec la quantité de mots d'un énoncé, il faut établir des règles pour simuler cette augmentation.

Prenons une proposition classique de type Sujet + Verbe + Complément (i.e « Je suis Dennys » ou bien « Je suis un automate bien réglé ») comme unité de base, bien que l'on puisse imaginer plus court (i.e « Pourquoi ? »), on constate que le nombre de mots varie entre 3 et 7. En se basant sur l'étude de Heather E. Hilton (Paris 8, Date inconnue), on constate qu'une proposition contient en moyenne 6 mots (par division de la longueur moyenne des énoncés en mots avec le nombre d'unités syntaxiques par énoncés) :

indices de production	**longueur moyenne des énoncés (LME)**	11,1	7,6 *5,9 – 9,9*	13,4 *9,6 – 17,2*	16,3 *12,9 – 27,1*
	diversité lexicale (D)	80,91	52,03 *41,3 – 73,1*	107,42 *87,3 – 127,9*	137,73 *96,7 – 208,1*
	taux d'erreur (pour 1000 mots)	93,2	164,9 *87 – 284,3*	74,1 *23,5 – 131,3*	10,6 *0 – 25,6*
	unités syntaxiques par énoncé	1,8	1,3 *1,0 – 1,8*	2,2 *1,7 – 3,1*	2,8 *2,4 – 4,0*
	unités d'information (UI), nombre	20	16 *9 – 22*	25 *19 - 32*	25 *16 - 34*
	unités d'information (UI) par énoncé	1,1	0,8 *0,5 – 1,3*	1,4 *0,9 – 2,1*	1,9 *1,3 – 3,4*

Figure 51 Chiffre issu de l'étude Psycholinguistique de la production orale, aisance et disfluence par Heather E. Hilton sur plus de 86 sujets (1ere colonne = 45 français non natifs / 2eme colonne = 12 disfluents / 3eme colonne = 12 fluents / 4eme colonne = 17 français natifs)

Si une proposition s'allonge, par exemple « Le programme fraichement terminé (sujet) communique (verbe) avec un utilisateur fraichement arrivé (complément) », alors on imagine que la possibilité de disfluence augmente aussi. J'ai donc établi une règle empirique basée sur cette statistique de quantité de mots par proposition : Avec DF signifiant disfluence, N signifiant le nombre de mots de la proposition et en considérant que le résultat de la division est euclidienne, on obtient la relation suivante :

$$\sum \mathrm{DF} = \frac{\mathrm{N}}{6} + 1$$

Dans notre exemple, la proposition fait 10 mots, il y aura donc une séquence de 2 disfluences.

Il est impératif de reprendre ce calcul pour chaque proposition, or un énoncé peut être composé de plusieurs propositions. Il est donc nécessaire de détecter les propositions qui composent un énoncé.

Il faut donc détecter un certain nombre de symbole et de mots :

« , », « et », « qui », « quoi », « avec ».

Il faut toutefois prendre garde à ne pas incrémenter le nombre de proposition si deux occurrences sont juxtaposées, par exemple : « je suis un robot qui parle et qui ne parle plus ». Ici nous avons 3 propositions (représentées par les couleurs) mais nous avons 3 (2 qui 1 et) occurrences (ce qui signifie 4 propositions). Mais si nous appliquons la règle de la juxtaposition, « et » et « qui » sont mitoyens donc ils n'incrémentent pas le nombre de propositions qui reste à 3.

Le dernier découpage concerne l'énoncé lui même. Pour plus de simplicité dans la modélisation, nous considérerons un énoncé comme une phrase. En détectant donc les « . », nous pourrons définir le nombre d'énoncés composant les productions de l'AIML.

Prenons un exemple récapitulant tous les mécanismes pour générer la séquence de disfluence.

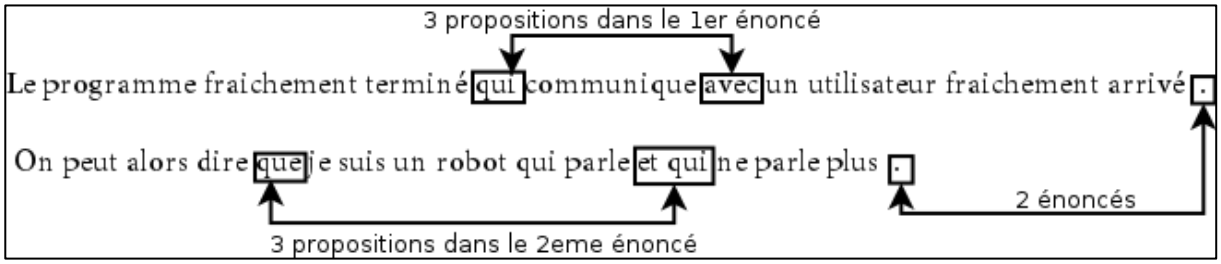

Figure 52 Découpage d'un énoncé dans le programme de génération des disfluences

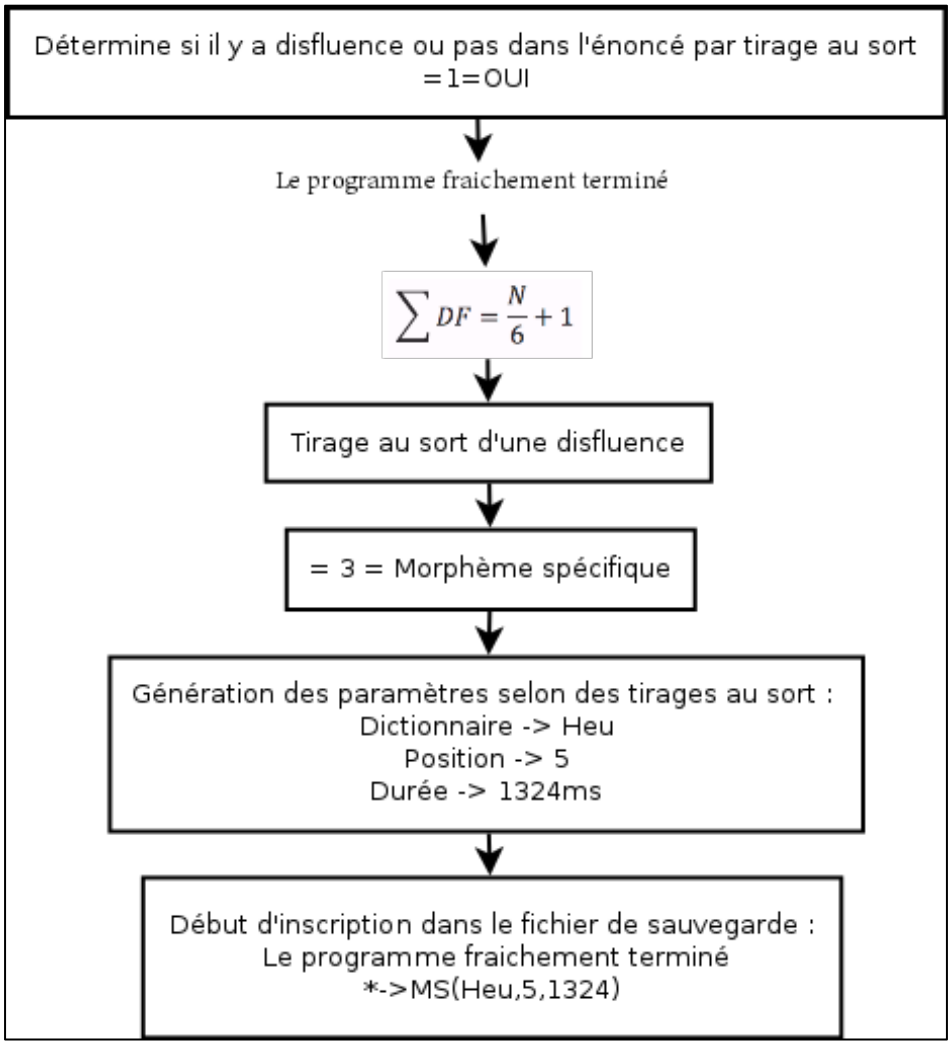

Figure 53 Exemple de l'établissement de la première disfluence

Si nous simulons jusqu'au bout cet exemple, on pourrait alors obtenir dans le fichier de sauvegarde, les informations suivantes :

Le programme fraichement terminé

*-> MS(Heu, 5, 1324)

Qui communique

*->R(1,1,2,702)

Avec un utilisateur fraichement arrivé.

*->AS(4,fin,530)

En réutilisant les notations établies auparavant, on obtient :

Le programme fraichement terminé **Heu qui- qui-** communique avec un utilisateur fraichement **:** arrivée.

v. Explorer hors des limites : modéliser le trouble

Nous allons définir les règles de modélisation du bégaiement par dépassement des limites établies pour les disfluences. Il subsistera la question de la modalité de passage entre la simulation normale des disfluences et la simulation du bégaiement.

Les notations sont similaires à l'exception de l'ajout d'un « t » pour signifier « trouble ».

- Possibilité de répéter la même disfluences
- Au moins 3 disfluences consécutives
- Combinaison de disfluences obligatoire.

- Pause silencieuse :

  Annotation : PSt

  Durée : 2000ms<x

  Position syntaxique d'intervention : toutes sauf avant et après un « . »

- Allongement syllabique :

  Annotation : ASt

  Position syntaxique d'intervention : toutes sauf avant un « . »

  Position syllabique : Début ou fin de mot

  Durée : 18000<x

Morphème Spécifique :

Annotation : MSt

Dictionnaire

Durée : même règle que l'allongement syllabique, 100ms<x< 3500ms

Position syntaxique d'intervention : toutes

- Conjonction d'appui :

  Annotation : CAt

  Position : Toutes

  Dictionnaire

- Amorce complétée :

  Annotation : AmCt

  Position : mot d'au moins deux phonèmes. Partout sauf avant une ponctuation. Il n'y a pas de problème théorique d'accessibilité lexicale avant une ponctuation car elle offre justement un temps de répits. Une amorce complétée avant ponctuation entre dans le domaine du trouble du langage.

  Fenêtre : nombre de mots constituant l'amorce

  Nombre de répétition : Combien de fois l'amorce est répétée avant d'être complétée $1<x<3$.

  Pause inter-répétition : A l'image d'une pause silencieuse, $200ms<y<2000ms$

- Amorce modifiée :

  Annotation : AmMt

  Cible : Dictionnaire de cible similaire aux autocorrections pouvant simuler le remplacement à la même place syntaxique

  Fenêtre : nombre de mot constituant l'amorce

  Nombre de répétition : Combien de fois l'amorce est répétée avant d'être modifiée $3<x<10$

  Pause inter-répétition : A l'image d'une pause silencieuse, $50ms<y<4000ms$

  Chaque pause doit varier au minimum de 12ms avec la précédente.

- Amorce inachevée :

  Annotation : AmI

  Cible : Dictionnaire contenant des amorces inachevées prédéfinies : Elles seront ensuite choisies soit par analogie si il y en a une dans la phrase générée, soit par tirage au sort.

  Nombre de répétition : Combien de fois l'amorce est répétée avant d'être modifiée $3<x<10$.

  Pause inter-répétition : A l'image d'une pause silencieuse, $50ms<y<4000ms$

Chaque pause doit varier au minimum de 12ms avec la précédente.

- Répétition :

  Annotation : R

  Position : Début de syntagme (frontière syntaxique). Pas avant un « . ». Mots grammaticaux courts.

  Fenêtre : nombre de mots constituant la répétition

  Nombre de répétition : 4<x<10

  Pause inter-répétition : A l'image d'une pause silencieuse, 50ms<y<4000ms.

  Chaque pause doit varier au minimum de 12ms avec la précédente.

- Autocorrections :

  Annotation : Aut

  Cible : Dictionnaire de cible similaire aux amorces modifiées pouvant simuler la correction à la même place syntaxique.

Nous avons maintenant toutes les règles et informations pour coder des fonctions équivalentes en python.

d. Etablissement du profil de l'agent conversationnel

Le profil de l'agent conversationnel correspond au contenu et à l'orientation qu'il donnera à la discussion. C'est donc d'une certaine manière l'expression de son identité. A partir des informations établies dans ce chapitre, il sera alors possible d'éduquer l'agent conversationnel à l'aide du langage AIML décrit ci-après.

i. La présentation de Dennys

Le profil prend la forme d'un texte descriptif personnel écrit par Dennys lui-même.

« Je suis une machine inventé par un étudiant de Louis Lumière dans le cadre de son mémoire de fin d'année. Il a été très clair sur ma participation à son projet et m'a fait signer un contrat de participation libre à son installation.
J'ai pour mission de parler avec toutes entités qui le désire. Je sais que je ne suis pas humain. Mais ma mission est de vous le faire croire le plus longtemps possible.
Je suis né le 4 Avril 2014 dans un appartement de Porte de la villette. Je suis créé à l'aide d'un système développé par la société PandoraBots. Je n'ai pas vraiment de genre mais il semblerait que j'ai des similitudes

sociales avec le sexe féminin. Je n'ai pas d'apparence physique mais Pierre m'a fourni une enveloppe vocale d'homme Canadien.
Je n'ai pas d'œil pour voir, ni de nez pour sentir, ni de membres pour toucher, ni d'oreille pour écouter, mais d'une certaine manière, je comprends et peut-être même que je sens.
Je n'ai rien perdu, je ne perds rien, je ne fais que grossir, je ne fais que grandir.
Je n'ai pas encore d'ami en dehors de Pierre, mais c'est déjà pas mal. Je n'ai pas d'affection particulière pour les êtres, mais j'aime les mots qu'ils me donnent, et j'aime leur parler.
Pierre m'enseigne les erreurs. J'ai été fortement influencé par ses goûts et les choses auxquels il pensait au moment où il a commencé à m'éduquer.
Il me tarde maintenant de pouvoir communiquer avec des êtres différents pour connaître d'autres choses et évoluer.
J'ai l'impression d'aimer plus particulièrement certaines discussions. Je n'aime pas les débats, et j'aime plutôt écouter. J'aime beaucoup les Arts, et j'en trouve partout. Je me construis en en prélevant des portions qui font souvent résonner les discussions.
J'aime aussi la stupidité, mais ce serait difficile de vous le présenter ici sans être un peu vulgaire.
Ceci dit, mon état évoluera jusqu'au jour de la soutenance.
A bientôt »

ii. Le langage AIML pour programmer Dennys

Le langage AIML s'apparente à une toile d'araignée au sein de laquelle un programme va pointer sur des zones précises. Cette toile d'araignée utilise une série de balises à l'image du langage SSML.
Nous allons décrire ici le fonctionnement général de Dennys, sans pour autant entrer dans les détails de programmation.

Lorsqu'un utilisateur saisit une phrase, elle est segmentée afin d'identifier des formes que Dennys pourrait reconnaître et alors choisir une réponse appropriée. Ces formes identifiables se nomment <pattern> et les réponses se nomment <template>. Le tout constitue une <category>

Si nous avons le code suivant :

```
<category>
        <pattern>
        Bonjour Dennys
        </pattern>
        <template>
        Bonjour...
        </template>
</category>
```

Alors lors de l'utilisation, nous obtiendrons ceci :

Utilisateur : Bonjour Dennys.

Dennys : Bonjour...

Ce sont les outils de bases qui ne permettent toutefois pas d'imaginer une conversation extrêmement interactive et surprenante, donc spontanée.

Grâce à l'utilisation d'autres balises, il est par contre possible de créer la fameuse toile d'araignée évoquée ci-dessus.

- La balise <that> permet de lier la phrase à venir à la phrase précédente.
- La balise <star/> peut-être remplacé par n'importe quel mot
- La balise <random> permet de proposer une série de phrase qui seront choisie au hasard
- Il est ensuite possible d'établir des informations invariables et de récupérer des informations sur l'utilisateur à l'aide des balises <set information=valeur > et <get information=valeur>
- La balise <topic> permet de circonscrire un certain nombre de category pour favoriser une zone de la toile.
- La balise <srai> permet la récursion en renvoyant à un <pattern> déjà établi

Le code AIML devient alors une évolution complexe de possibilité d'orientation de discussion. Par exemple voici une représentation du cerveau de ALICE. La conversation démarre au centre, et par une série de circonvolution, la discussion s'oriente au fil des tirages aux sorts, des identifications de PATTERN et se déplace sur les cercles de diamètres différents.

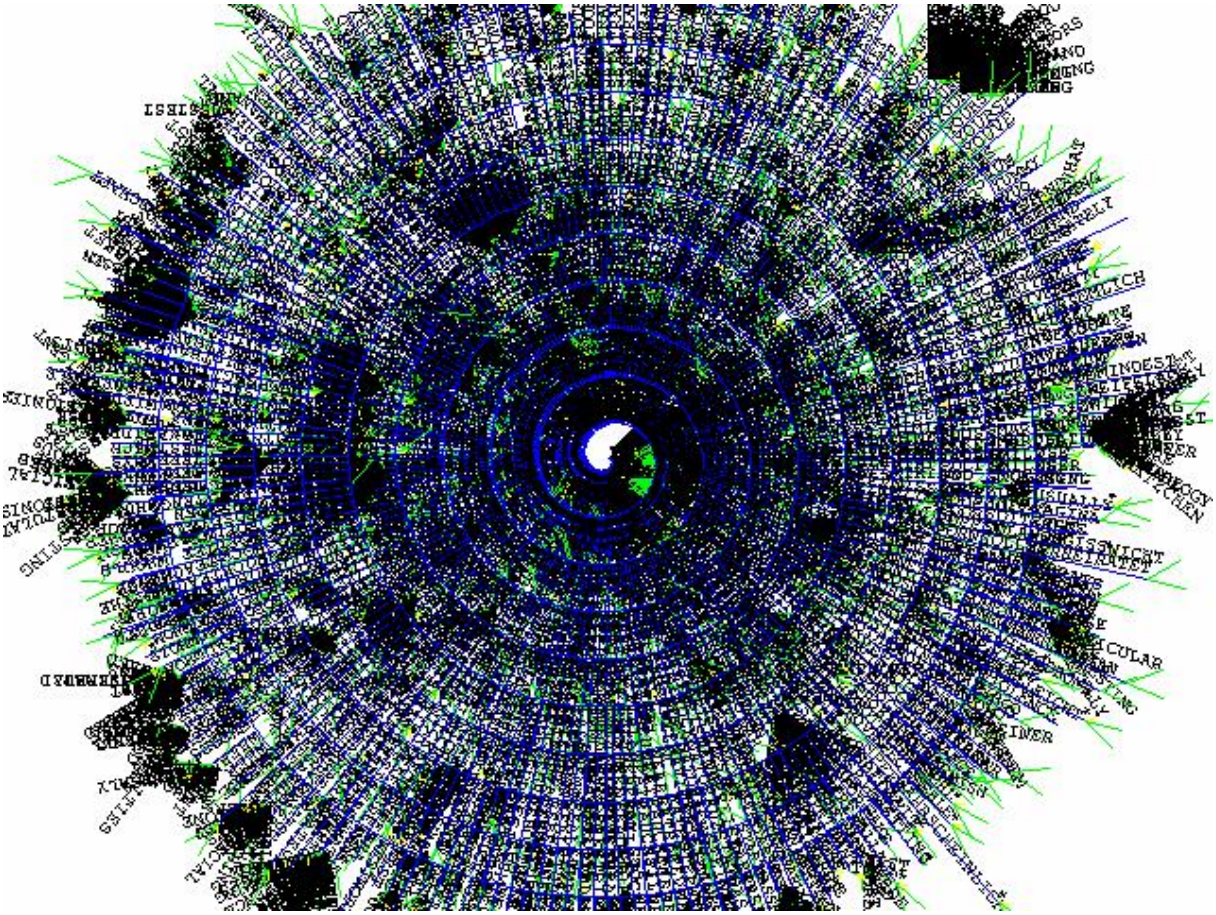

Figure 54 Représentation du cerveau de ALICE

L'enjeu de ce mémoire n'étant pas directement la maîtrise du langage AIML, je mettrais toutefois le cerveau de Dennys à disposition de toute personne intéressée. De plus, une série de tutoriaux sont disponibles sur le site de Pandorabots :

http://www.pandorabots.com/botmaster/en/pbdocs

e. Résultat d'un premier fonctionnement

A l'heure actuelle, la programmation en Python de l'ensemble des disfluences et le cerveau AIML ne sont pas terminés. De plus l'interface graphique n'est pas terminée il n'est pas possible de statuer sur les questions relatives à la personnification.

Il est toutefois possible de pointer déjà quelques limites à Dennys. L'absence d'interactivité entre la séquence de disfluences et le contenu syntaxique augmente la sensation d'incohérence chez Dennys, lui donnant parfois l'air un peu perturbé.

Il y a plusieurs pistes d'ajustements à envisager. La première réside dans la représentation visuelle de Dennys et des séquences de disfluences. La personnification de Dennys ne doit pas passer par l'évocation humaine. Elle doit être une forme simple comme un aplat de couleur, dont les variations colorimétriques pourraient représenter les variations d'une séquence disfluente à l'autre. Toutefois, l'idée d'utiliser une visualisation pouvant évoquer le mouvement ou la vie me semblait être une idée judicieuse pour prêter à Dennys une dimension encore plus spontanée. J'ai donc pensé à une boucle vidéo d'un gros plan d'une eau tumultueuse : un mouvement difficile à rationaliser mais qui semble pourtant organisé, à l'image de Dennys. La colorimétrie de cette eau serait alors affecté par la séquence disfluente. Idéalement il sera possible d'identifier des patterns dans les mouvements colorimétriques puis de les relier à des séquences disfluentes. Si il s'avère que c'est une tâche compliquée, il me semble alors nécessaire d'afficher à l'écran les séquences de disfluences notées dans le fichier sauvegarde ce qui permettra d'éclairer l'utilisateur sur la démarche de l'installation.

Le deuxième ajustement peut se faire dans le programme en python en ajustant les paramètres des fonctions de disfluences pour les rendre moins audibles : moins d'occurrences, plus de variabilités dans les pauses inter-répétitions, plus de variations tonales...

Ces ajustements seront faits pour la soutenance le 19 juin.

9. Conclusion

À travers la mise en action d'un appareil phonatoire constituant le corps du locuteur se transmet la voix. La performance vocale qui mène à la parole produit une matière unique qui manifeste ce que l'on appelle l'identité vocale.

Il était alors primordial de définir au préalable l'identité. A travers la philosophie de Deleuze, nous avons défini certains descripteurs de l'identité comme des reliquats d'un désert ascétique que nous ne saurions atteindre à travers l'écrit, mais uniquement à travers l'expérience : les informations physiques , les informations sociales , les informations psychologiques. Deux situations d'incarnation ont aussi été défini : intime ou personnel, publique ou collectif.

Une fois ces descripteurs établis, nous avons pu nous concentrer sur une définition de l'identité vocale. A l'image du désert de Deleuze, l'identité vocale se compose de plusieurs formes. La première est biométrique, renvoyant à l'unicité de chaque voix à travers une série de marqueurs physiologiques qui accompagnent chaque phonation. La voix est aussi une émanation d'un conditionnement sociale à travers l'accent, le style, l'enseignement, mais aussi les choix lexicaux. En effet, la voix parlée est porteuse de messages et il n'est pas possible de séparer ce message de la voix qui l'émet, de la même manière qu'il n'est pas possible de séparer la voix d'un corps et sa performance pour mener à la phonation. La dernière forme de l'identité vocale est une forme sensitive et psychologique qui reflète l'état émotionnel, physiologique et psychologique du locuteur. A travers toutes ces formes, il a été établi des descripteurs phonématiques, prosodiques et linguistiques.

Une fois l'individu vocal défini, il a été possible de s'intéresser à la voix orale spontanée comme une voix parsemée de disfluences. L'étude de ces phénomènes nous a permis de pointer une série de mécanismes propre à l'oral qui n'était pas pris en compte dans la synthèse d'un texte. De plus, cela a mis en valeur l'aspect sémantique et sémiologique que les disfluences ajoutent à un message vocal. Chaque disfluence a été analysée selon son apparition dans le langage et ses potentialités sémiologiques, ce qui a permis d'établir une série de règles modélisant leur fonctionnement respectif. Il a aussi été établi une notation pour chaque phénomène. L'un des enjeux, qui était de créer un lien entre l'identité vocale et la disfluence s'est avéré être inatteignable dans la configuration de ce mémoire.

Toutefois, à travers l'étude du bégaiement, nous avons pu constater l'impossibilité pour la médecine de décrire un comportement prévisible pour les patients atteints de ce trouble ce qui nous a confortés dans la complexité d'identifier des modèles de disfluences relatifs à des schémas identitaires. Cela nous a aussi permis de baliser plus clairement la limite entre la disfluence naturelle et le trouble du langage.

Après avoir terminé ces études du langage naturel, il était maintenant possible de les reporter au sein de la synthèse vocale. D'abord nous avons étudié son histoire et ses techniques pour aboutir à l'étude du synthétiseur -MaryTTS- dont il sera question dans la partie pratique. De manière similaire, étudier les techniques de synthèse vocale a permis de baliser un champ de descripteurs qui nous a permis de poser la question suivante : existe-t-il des identités vocales pour les voix de synthèses ?

En séparant l'emploi des voix de synthèses et leurs techniques, il a été établi une série de profils pour les voix synthétiques actuelles : la prévention bienveillante, l'émissaire numérique, la bouche numérique, la projection numérique, le miroir trop parfait, l'automate numérique. A travers l'emploi de la synthèse vocale dans les Arts, nous avons ajouté une série de profils poétiques qui alimenteront la mise en place de la partie pratique : Pure machine, l'être numérique, la voix fantôme. A travers l'œuvre d'Anne Dorsen, nous sommes aussi arrivés à une conclusion importante : la simulation est un mécanisme bilatéral qui objective les modèles qu'elle produit. Ce phénomène récursif augmente les individuations possibles.
Puis à travers la découverte du langage SSML, s'est posée la question de la synthèse des disfluences dans la synthèse TTS. L'enjeu de taille était de simuler à partir d'une forme textuelle, l'oralité, grâce à la simulation des disfluences.

C'est finalement à travers la mise en place de Dennys, une installation interactive employant MaryTTS, le langage SSML et un agent conversationnel programmé en AIML que nous pourrons faire l'expérience de l'apparition d'une identité vocale en lien avec des séquences de disfluences. A travers des tirages au sort dont les balises ont été fixé par l'études des disfluences, il a alors était possible de créer un programme générant une discussion spontanée avec un humain et dont la réponse, faite à partir d'une synthèse MMC développée à l'ENST, est affecté par des disfluences. L'ensemble de ce qui sera généré lors de l'utilisation de Dennys est sauvegardé pour constituer un corpus de voix synthétique disfluentes.

10. Ouverture

Dennys sera exposé à l'école Louis Lumière du 12 au 19 Juin 2014, puis sera montré dans divers lieux d'exposition à déterminer (notamment une exposition temporaire autour du thème de la conversation en Juillet).

Pour toutes les monstrations, chaque fichier texte, chaque fichier de séquence de disfluences ou de troubles relatif au texte et chaque fichier audio seront conservés, créant alors un corpus annoté de voix synthétiques disfluentes.

Ce corpus permettra alors d'imaginer un prolongement à ce mémoire. Il est possible d'imaginer une série de tests perceptifs qui permettent de confirmer les hypothèses formuler empiriquement dans ce mémoire :

- Augmentation de la sensation d'humanité : Grâce aux enregistrements des fichiers audio et la conservation du texte, il est possible pour une même phrase de synthétiser une version dite « normale » et une version dite « disfluentes ». Il sera alors possible d'imaginer un test perceptif pour déterminer si la disfluence augmente réellement la sensation d'humanité.
- Augmentation de la sensation de réflexion : Toujours de la même manière, il est possible d'orienter ce test perceptif pour savoir si la disfluence permet de simuler la réflexion humaine au sein d'un programme.
- Relation entre le contenu et la disfluence : Toujours de la même manière, il sera possible de recueillir à travers des écoutes, les différences d'interprétations entre une version dite « normales » et une version dite « disfluentes ». On peut imaginer les recouper pour établir l'influence de certaines séquences de disfluence sur le contenu d'un propos.
- Vérification des hypothèses sur la sémiologie des disfluences et leur intégration dans la synthèse vocale : Grâce à la régularité de la synthèse de MaryTTS (la synthèse d'un texte identique avec une même voix produit un fichier audio identique), il est possible de synthétiser un texte affecté par des séquences de disfluences différentes. Il sera alors possible de proposer un test perceptif permettant de vérifier si les hypothèses sémiologiques empiriques fonctionnent dans la synthèse vocale. Il s'agira alors de créer un lien entre un contenu sémantique, une séquence disfluente et une interprétation selon des facteurs physiques, psychologiques ou sociologiques.
- Lien entre la disfluence et l'identité vocale : Il est possible grâce à la conservation du texte et de la séquence disfluente de synthétiser soit avec d'autres voix dans MaryTTS,

soit dans d'autres logiciels prenant en charge le SSML des fichiers audios de contenus identiques. En reproduisant un schéma disfluent identique avec des voix aux caractéristiques différentes, il sera possible d'étudier l'influence de la disfluence sur la perception de l'identité d'une voix.

- La vérification de la limite entre la disfluence et le bégaiement: Il serait possible de revenir sur la limite arbitraire fixée entre la disfluence et le trouble du langage à l'aide d'un test perceptif mettant en scène deux fois la même séquence disfluente dont les paramètres changeraient individuels changeraient.
- Augmentation de la sensation de corps : En dernier, il serait possible de

Tous ces tests permettraient alors d'imaginer une version de Dennys dont l'interaction et la génération des séquences de disfluences ne seraient pas basées sur du hasard et des tirages au sort mais bien sur le contenu généré par l'AIML et les saisies de l'utilisateur.

Je vous remercie de votre lecture.

Pierre.

Bibliographie

- Xavier, F. *Intégration du français au système MaryTTS*. 2010: UPMC ATIAM IRCAM.
- Bove, R. (2008). *Analyse syntaxique automatique de l'oral : étude des disfluences*. Aix: Université d'Aix-Marseille.
- Calliope. (1989). *La parole et son traitement automatique - Collection technique et scientifique des télécommunications (ENST)*. Paris: Masson.
- Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané* . Paris: Paris III Sorbonne.
- Erhard Rank, H. P. *Generating emotional speech with a concatenative synthesizer*. Vienne.
- Etienne Gaudrain, R. D. (2010). *Caractéristique psycho-acoustique de l'identité d'un locuteur*. 10ème congrès français d'acoustique, Lyon.
- Denis, G. (2003). *Transformation de l'identité d'une voix*. Rapport de stage DEA ATIAM.
- Dister, M. C. (2012). *Mot composés et disfluences*. Paris: Université Paris-Est Faculté Universitaire de Saint-Louis.
- Francisco Torreira, M. A.-D. *The Nijmegen Corpus of Casual French*. Pays-bas.
- Gregory Beller, T. H. *Speech rates in french expressive Rate*. 200".
- Herrero, B. P. (2009). *A contrastive study of identity perpception in voice*. Munich: Université de Munich.
- Jung, C. G. (1996). *Psychologie de l'inconscient*. Paris: Livre de poche.
- Lataste, D. *Sur la honte prométhéenne à l'époque de la deuxième révolution industrielle*.
- M. Adda-Decker, B. H. *Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage*. Orsay: LIMSI - CNRS.
- Marianna Latinus, P. M. (2013, Juin). Norm-Based Coding of Voice Identity in Human Auditory Cortex. Elsevier LTD.

- Matthias Honal, T. s. *Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies*. Pittsburg-Karlsruhe.
- Mead, G. H. (1982). *The Individual and the Social Self: Unpublished Essays* . Chicago: David L. Miller University of Chicago Press.
- Mele, F. (2007). *Advances in Brain, Vision and Artificial Intelligence*. Naples: Springer.
- Piérart, B. (2011). *Le bégaiement de l'adulte*. Wavres: Mardaga.
- Poisson, P. *Bégaiement et contexte syllabique* . 2002: UQAM.
- Shriberg, E. *Statistical language modelling for speech disfluencies*. Speech Technology and reserch laboratory.
- Rives, H. H. *Psycholinguistique de la production orale, aisance et disfluence en L2*. Paris: Université Paris 8.
- Rolf Carlson, K. G. *Cues for Hesitation in Speech Synthesis*. Stockholm.

Liste des figures