MEMOIRE

En vue de l'obtention du

MASTER RECHERCHE

Informatique et Télécommunication

Année universitaire 2015/2016

Délivré par :

La Faculté des Sciences de l'Université Libanaise et l'Université Paul Sabatier – Toulouse 3

Présenté et soutenu par :

Sandy Elias Aoun

Le mercredi 28 septembre 2016

Titre

Constitution optimale d'une voix de synthèse pour livres audio

---

Jury : M. Ali Choumane (co-encadrant)
M. Bilal Chebaro (examinateur)
M. Ali Awada (examinateur)

Travail effectué à l'IRISA / équipe Expression sous la direction de :
Mme Nelly Barbot
M. Jonathan Chevelu
M. Damien Lolive

« Ce qui embellit le désert, dit le petit prince, c'est qu'il cache un puits quelque part… »

Antoine de Saint-Exupéry

Remerciements

Je remercie l'Université de Rennes 1, présidée par David Alis, pour sa collaboration avec l'Université Libanaise et l'équipe de recherche Expression de l'IRISA dirigé par Jean-Marc Jézéquel pour la gratification de stage des deux mois que j'ai passé à Lannion en France.

Je tiens à remercier l'ENSSAT, dirigée par Jean-Christophe Pettier, qui a mis à ma disposition un bureau, les matériels informatiques et l'accès à des ressources et de programmes nécessaires à l'étude de mon stage. Je remercie particulièrement l'assistance administrative de Joëlle Thépault et de Vincent Chevrette.

Je tiens tout d'abord à adresser mes remerciements à Ali Choumane, professeur en Informatique à l'Université Libanaise, qui m'a beaucoup aidée au début du stage pour m'initier au domaine de la synthèse de la parole et qui a assuré le bon déroulement de mon stage.

J'adresse mes plus sincères remerciements à mes maitres de stage, Jonathan Chevelu et Damien Lolive. Merci à Jonathan qui m'a introduit au monde de la recherche, qui m'a communiquée tant de choses et m'a apportée de nouveaux motifs de programmation. Merci Jonathan de votre patience et de vos remarques très profitables. Merci à Damien qui a toujours répondu à mes questions et notamment de m'avoir intégrée dans la vie de l'équipe, et de ses différents conseils scientifiques.

Je remercie également toute l'équipe Expression pour son accueil, en particulier Nelly Barbot et David Guennec. Merci à Nelly pour ses explications sur le problème SCP et certains algorithmes pour résoudre SCP, pour les références bibliographiques et de ses relectures et corrections. Merci à David pour nos discussions sur le système de synthèse par sélection de l'équipe.

Pour finir, je tiens à remercier Ali Awada et Bilal Chebaro, professeurs en Informatique à l'Université Libanaise, qui m'ont donnée accès à une salle visioconférence quand j'en avais besoin, et pour leur assistance continue durant cette année.

Résumé

La réalisation de livres audio par synthèse de la parole est un défi qui préoccupe le monde de la recherche et le monde industriel. Une voix de haute qualité avec des types d'expressivité adaptés aux contextes n'est actuellement pas atteignable en synthétisant des livres textuels avec un corpus acoustique généraliste. Ainsi ce stage consiste à choisir un sous-ensemble de phrases extrait du livre, le plus petit possible, afin de maximiser la qualité globale du livre audio une fois complété par synthèse. Dans ce cadre, notre première expérience repose sur un livre audio déjà enregistré. Pour étudier le compromis entre le coût d'enregistrement et la qualité acoustique finale, nous avons mis en œuvre un algorithme cracheur et une fonction d'évaluation objective. Nous concluons, selon les résultats partiels obtenus, que la spécification d'un ensemble de phrases du livre à synthétiser en priorité est réalisable. De plus, il serait intéressant d'élaborer des méthodes d'apprentissage automatique en étudiant les caractéristiques des phrases qui composent les partitions jugées optimales par l'algorithme. Ainsi nous pouvons espérer résoudre ce problème d'optimisation pour un livre textuel non enregistré à l'avance.

Mots-clés : Synthèse de la parole expressive ; optimisation combinatoire ; traitement automatique des langues.

Table des matières

1 Introduction

La synthèse de la parole est la production artificielle de la parole humaine. Le but de la synthèse de parole est de produire à partir d'un texte en langage naturel en entrée un signal de parole en sortie. Pour convertir automatiquement un texte en un signal de parole, le système de synthèse réalise deux types de traitements : les traitements linguistiques et les traitements acoustiques.

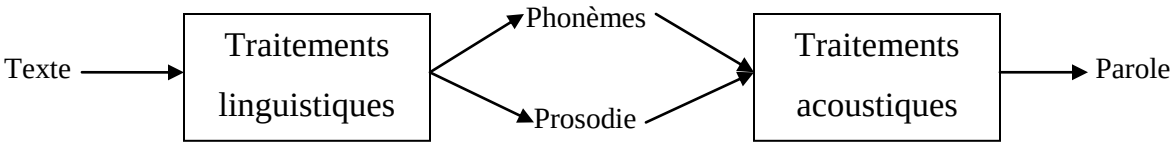

Figure 1 : Schéma général d'un système de synthèse de parole à partir du texte.

Les traitements linguistiques visent à fournir la suite des phonèmes et la prosodie associée au texte à synthétiser. Ces informations sont ensuite utilisées pour fabriquer un signal de parole correspondant à la vocalisation du texte donné.

L'une des méthodes fréquemment utilisée et qui s'inscrit dans notre travail de recherche est la synthèse par sélection d'unités. Ces moteurs de synthèse requièrent une base de parole construite en enregistrant un locuteur lisant un script de lecture précis. Le signal de parole est obtenu en sélectionnant et en juxtaposant des segments de parole de cette base.

Au cours des dernières années, la synthèse vocale a fait de nombreux progrès du point de vue de l'intelligibilité et de la qualité acoustique, cependant [Govind et al. 2013] expriment le manque de naturel et de qualité dans la parole synthétique expressive. Cette qualité de parole finale dépend bien du système de synthèse mais surtout du corpus acoustique utilisé.

L'étape d'enregistrement d'un script de lecture spécifique (c'est-à-dire construction de la base de parole) étant complexe et coûteuse, plusieurs travaux adressent son optimisation [Kawai et al. 2000; François 2002; François et Boëffard 2002; Chevelu et al. 2007; Cadic 2011; Barbot

et al. 2015]. Celle-ci consiste à maximiser la couverture d'événements acoustiques désirés pour une application précise tout en minimisant la durée des enregistrements.

Notre étude concerne la construction automatique de livres audio par synthèse vocale. La stratégie classique consiste à créer une voix de synthèse en se basant sur un script de lecture qui ne dérive pas du livre audio à synthétiser, puis à synthétiser l'ensemble des phrases du livre avec cette voix. Puisque en utilisant une voix généraliste la production du livre avec une voix expressive et de très bonne qualité demeure un verrou scientifique, nous aspirons à examiner une approche qui consiste à choisir un sous-ensemble du livre à enregistrer, puis à l'utiliser pour constituer une voix de synthèse consacrée à générer les phrases manquantes du livre audio par synthèse vocale.

Ce cas spécial offre une nouvelle perspective au problème de création de voix : lorsqu'on sait le texte à synthétiser à l'avance, est-ce que l'on peut optimiser le processus de création de voix traditionnel ? Alors il s'agit de déterminer une partie minimale à enregistrer des livres concernés, afin de produire une voix de synthèse finale la mieux adaptée à la partie restante à vocaliser. Notre première tentative à traiter ce problème repose sur l'utilisation d'un livre audio dont l'enregistrement naturel est disponible au préalable, ainsi nous utiliserons des mesures de qualité fondées sur le signal acoustique.

Ce rapport est organisé comme suit. Dans la section 2 nous développons les différentes étapes qui se déroulent dans un système de synthèse de parole à partir du texte. La section 3 porte sur le processus de création de voix pour les systèmes de synthèse par corpus. Nous expliquons les différentes approches pour évaluer les systèmes de synthèse dans la section 4. L'expérience que nous avons menée est formalisée dans la section 5 et la section 6 se charge de la présentation et la discussion des résultats. Enfin, dans la section 7 nous concluons le travail présenté dans ce rapport et nous proposons des perspectives sur des travaux à réaliser**.**

2 Synthèse de parole à partir du texte

La problématique de notre étude se situe dans le domaine de la synthèse de parole à partir du texte. Comme illustré sur la figure 2, la transformation d'un énoncé en un signal acoustique est opérée par le système de synthèse en deux blocs principaux : les traitements linguistiques de l'énoncé suivis des traitements acoustiques qui fabriquent le signal.

Cette section débute par la description des différentes phases qui constituent les traitements linguistiques. Les deux stratégies principales distinctives des traitements acoustiques sont ensuite présentées, en se concentrant sur la synthèse par corpus.

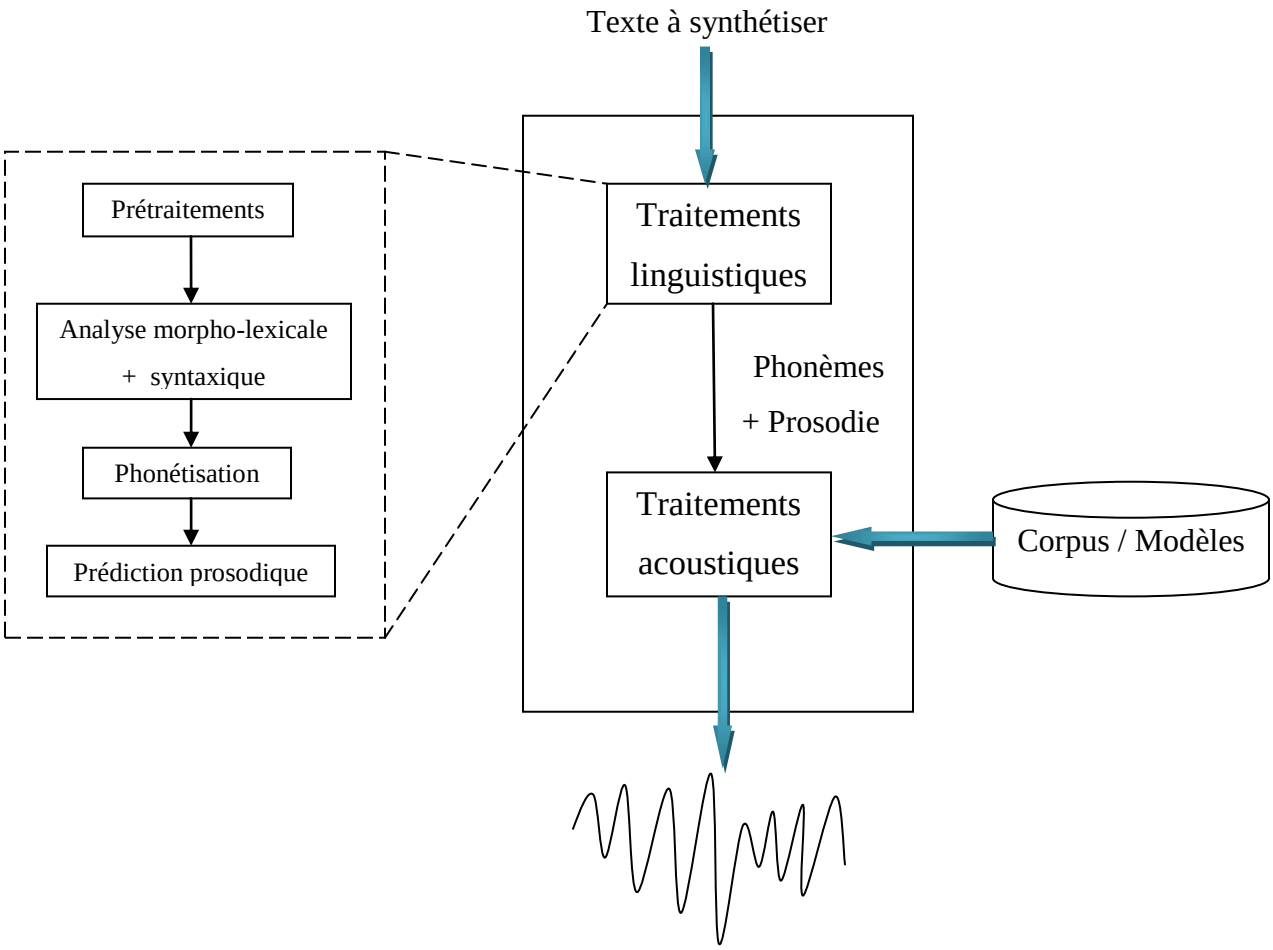

Figure 2 : Illustration plus détaillée du processus de synthèse de parole à partir du texte. Figure inspirée de [Le Maguer 2013].

2.1 Traitements linguistique

L'analyse linguistique et prosodique du texte d'entrée aboutit à la construction d'une chaîne phonétique et à la spécification des marqueurs prosodiques. Nous présentons dans la suite les différentes étapes du traitement linguistique et prosodique du texte.

Prétraitements

Tout d'abord, certains prétraitements du texte d'entrée sont effectués. Ils comportent le développement des abréviations, la reconnaissance et la réécriture des acronymes et des nombres, etc.

Analyse linguistique

Une pré-catégorisation grammaticale des mots présents dans le texte d'entrée sera établie suite à une analyse morpho-lexicale, qui utilise des lexiques et des règles de décomposition des mots inconnus en morphèmes (préfixes, racines, suffixe). Cependant certaines incertitudes grammaticales persistent. L'analyse syntaxique qui repose sur l'application d'un ensemble de règles grammaticales enlève ces ambigüités, et associe à chaque mot une étiquette morpho-syntaxique, comme le montre l'exemple de la table 1.

Phonétisation

La tâche de la phonétisation automatique est de produire la transcription phonétique du texte de départ. Elle consiste à convertir les graphèmes du message à synthétiser en phonèmes. Cette phase doit tenir compte de plusieurs contraintes comme :

- variation de la prononciation d'un graphème en fonction du contexte lexical ;
- présence des mots homographes hétérophones ;
- prononciation des noms propres et des mots nouveaux ;
- phénomènes d'assimilation de phonèmes ;
- ...

Plusieurs techniques peuvent être adoptées pour la conversion graphème-phonème. Parmi ces techniques nous pouvons distinguer les approches à base de règles [Béchet 2001; Claveau 2009] des approches statistiques [Bisani et al. 2008; Illina et al. 2011].

Prédiction prosodique

Les paramètres acoustiques souvent associés à la prosodie sont la fréquence fondamentale, la durée et l'intensité du signal. Le rôle de cette phase est la génération des marqueurs prosodiques tels que l'intonation et le rythme de la phrase en entrée. Elle se fait en déterminant la structure prosodique - découpage de la prosodie en groupes prosodiques - et la place des accents dans la phrase [Lolive 2008; Cadic 2011].

Nous	PRONOM personnel 1ère personne masculin pluriel nominatif
avons	VERBE auxiliaire indicatif présent 1ère personne pluriel
tout	PRONOM indéfini masculin singulier
rétréci	VERBE principal participe passé singulier masculin
de	PRÉPOSITION
nos	DÉTERMINANT possessif 1ère personne masculin pluriel
jours	NOM commun masculin pluriel
.	PONCTUATION

Table 1 : Un exemple d'étiquetage morpho-syntaxique de la phrase « Nous avons tout rétréci de nos jours. » en français extrait de [Paroubek 2006].

2.2 Traitements acoustiques

Le but des traitements acoustiques est de générer, étant donnés une chaîne phonétique et des marqueurs prosodiques, un signal acoustique de qualité qui correspond à la vocalisation du texte initial. Parmi les techniques de synthèse du signal existantes, deux approches notables prédominent actuellement le domaine de synthèse de parole à partir du texte : l'approche par concaténation et l'approche paramétrique.

La méthode par concaténation premièrement présenté par [Sagisaka 1988] a été adoptée par plusieurs systèmes de synthèse [Black et al. 1994; Hunt et al. 1996; Breen et al. 1998; Taylor et al. 1998; Clark et al. 2007; Alain et al. 2015]. Elle repose sur la concaténation des bouts[1] de parole tirés d'un corpus acoustique dédié à la synthèse vocale. La méthode paramétrique dont HTS est le système majeur [Yamagishi et al. 2008], porte sur la modélisation du signal par des paramètres acoustiques en vue de l'utiliser par la suite pour produire la forme d'onde du signal acoustique.

Nous allons décrire maintenant la synthèse par sélection d'unités[2] qui est la variante utilisée dans cette étude.

Synthèse par sélection d'unités

La synthèse par sélection d'unités consiste à concaténer des unités acoustiques extraites d'une base de données pré-enregistrée de parole.

Au début, l'unité élémentaire utilisée était le diphone : c'est l'unité acoustique qui s'étale du milieu d'un phone[3] au milieu du phone suivant, parce que ses frontières - les parties centrales des deux phones - se rapportent à des zones de stabilité acoustique, ce qui simplifie les

[1]Un bout de parole est une partie de longueur variable du signal acoustique.
[2]La synthèse par sélection d'unités est souvent appelée synthèse par corpus.
[3]Phone : Selon le contexte et différents facteurs comme l'origine, le sexe et l'âge du locuteur, chacun des phonèmes d'un système phonologique peut être prononcé de différentes manières. Ces variantes sont les phones, c'est-à-dire les réalisations acoustiques de ces phonèmes.

concaténations en minimisant les risques de discontinuités entre les unités consécutives. Le concept de synthèse vocale par concaténation d'unités de longueur variable a été introduit pour minimiser les effets de concaténation et la taille des enregistrements. À présent, une base de parole continue annotée sur plusieurs niveaux avec plusieurs types d'unités[4] est généralement utilisée dans ce cadre.

Pour chaque phrase d'entrée, les unités acoustiques à concaténer sont choisies par un algorithme de sélection d'unités. Chaque unité peut être un diphone, une séquence de diphones, une syllabe, un mot, un ensemble de mots, ainsi que la phrase complète si elle se trouve dans la base. La fonctionnalité de cet algorithme est de sélectionner la séquence d'unités de la base de parole qui convient le plus à la séquence d'unités cible[5]. Il s'agit en général de minimiser une fonction de coût constituée d'un coût cible et d'un coût de concaténation [Black et al. 1994] :

- Le coût cible sert à déterminer les unités du corpus qui ont des caractéristiques phonétiques, syntaxiques et prosodiques les plus proches de celles de la séquence cible ;

- Le coût de concaténation sert à mesurer les distorsions acoustiques provenant de la concaténation de deux unités candidates consécutives (unité courante et celle qui la précède) sur le chemin exploré.

La recherche de la meilleure séquence d'unité se ramène à la recherche du meilleur chemin dans un graphe. Cette problématique est typiquement résolue par l'exécution d'un algorithme de Viterbi [Viterbi 1967; Hunt et al. 1996; Conkie et al. 2000; Clark et al. 2007] ou d'un algorithme A^* [Hart et al. 1968; Guennec et al. 2014].

Enfin, un post-traitement, tel qu'un lissage spectral du signal généré peut être effectué aux endroits de concaténation des unités sélectionnées de manière à rendre les transitions plus continues [François 2002].

[4]Différents types d'unités comme les triphones, quadriphones, syllables, mots, etc.
[5]Une séquence cible d'unités acoustiques en contextes de la phrase d'entrée est définie à ce stade.

La qualité du signal de parole produit est certainement liée à la richesse du corpus (diversité de son contenu), ainsi que du contenu de la phrase à synthétiser et de l'algorithme de sélection utilisé. Plus le corpus est vaste, plus les unités acoustiques choisies peuvent être longues et adaptées au contexte.

L'équipe Expression de l'IRISA dispose d'un moteur de synthèse [Guennec et al. 2014; Alain et al. 2015] fondé sur la technique par sélection d'unités, que nous utiliserons par la suite dans notre travail (voir section 5.3).

3 Création de voix pour les systèmes de synthèse par sélection

Nous nous intéressons maintenant à la création d'une base de données vocale qui est indispensable au fonctionnent d'un système de synthèse par sélection d'unités. La construction d'un tel corpus est traditionnellement accomplie en trois étapes principales qui sont : la constitution du script de lecture, l'enregistrement du script et le post-traitement des données. Cependant une technique tout à fait distincte, qui consiste à collecter des enregistrements non dédiés à la synthèse de la parole [Cadic 2011; Chalamandaris et al. 2014], peut être appliquée dans le but de créer cette base.

Nous détaillerons ci-dessous les différentes étapes de la création classique d'une voix de synthèse, en se focalisant sur la constitution du script de lecture qui est au sein de notre travail de recherche.

3.1 Constitution du script de lecture

Le script de lecture correspond à l'ensemble des phrases qui seront lues par un locuteur pendant l'étape d'enregistrement du script. La constitution du script de lecture est une étape cruciale pour tout système de synthèse par corpus, puisque la variabilité contextuelle des unités disponibles dans la base de parole dépend principalement du contenu du script, donc celui-ci influence directement la qualité de parole générée par le système.

Au cours des années antérieures, la qualité de la voix synthétisée a été améliorée suite à l'augmentation de la quantité d'enregistrements vocaux. Pourtant [Kawai et al. 2004] perçoivent que la qualité de synthèse ne progresse plus, quand on dispose d'un corpus de parole de plus de 30 heures.

Néanmoins la minimisation de la durée d'enregistrements de parole est également un point critique pour minimiser le coût financier, pour assurer une qualité uniforme de la voix, pour réduire la tâche pénible d'enregistrement de script et pour restreindre le travail manuel de post-traitement des données [Barbot et al. 2015]. Il est alors fondamental d'optimiser le script de

lecture de façon à couvrir un ensemble d'unités requises pour une qualité acceptable de voix de synthèse fournie par le système.

Les unités plausibles à couvrir sont les phonèmes, les diphonèmes, les triphonèmes, les n-phonèmes en contexte (c.-à-d. diphonème si n=2, triphonème si n=3, etc.) : c'est le cas où une unité est distinguée par des caractéristiques contextuelles qui clarifient son environnement linguistique, prosodique, ou phonétique (exemple de contexte d'une unité : position dans la syllabe, le mot ou la phrase), les mots, l'étiquette POS[6], etc.

Les critères d'optimisation de script se traduisent par un taux de couverture d'un type d'unités ou d'une combinaison de plusieurs types d'unités. Pour mesurer ce taux, des approches en largeur et en fréquence peuvent être utilisées. Les approches en largeur tâchent à maximiser le nombre d'unités distincts du type d'unités recherché dans le script de lecture, tandis que les approches en fréquences tâchent à favoriser l'insertion des événements fréquents (c.-à-d. pourcentage d'apparition de l'événement dans un corpus de référence) dans le script de lecture en affirmant qu'ils seront plus utiles que ceux qui sont exceptionnels.

Pour couvrir les critères choisis pour un système de synthèse spécifique, plusieurs méthodes sont réalisables. Dans quelques systèmes, les phrases du script de lecture sont sélectionnées aléatoirement d'un immense ensemble de textes. Mais cette stratégie nécessite des investissements très importants parce que la distribution naturelle des phénomènes linguistiques suit la loi de Zipf [Zipf 1932], ce qui signifie que très peu de phénomènes sont assez fréquents alors que beaucoup de phénomènes sont très rares. Donc la couverture de plusieurs variantes des différents phénomènes linguistiques devient très complexe à atteindre [François 2002; Cadic 2011; Barbot et al. 2015].

Une autre méthode utilisée fréquemment dans ce cadre est la stratégie de condensation de corpus. Cette approche suppose l'existence d'un corpus initial formé d'un ensemble très vaste de phrases et qui dispose d'une large gamme d'attributs linguistiques, phonétiques et prosodiques. Les phrases du script de lecture sont sélectionnées du corpus initial de façon à assurer une

[6]POS : étiquette morpho-syntaxique de l'anglais POS.

couverture complète (ou pour le moins maximale) de l'ensemble d'unités à couvrir. La condensation de corpus peut être vue comme un problème de couverture d'ensemble [François 2002].

3.1.1 Problème de couverture d'ensembles pour la réduction de corpus

Nous adoptons les notations suivantes :

- Soit A le corpus initial composé de n phrases $s_1, \dots, s_n$. Chaque phrase du corpus est associée à une collection d'unités de différents types.
- Soit U l'ensemble des unités composant les phrases de A noté : $U = \{u_1, \dots, u_m\}$.
- Soit $B = (b_1, \dots, b_m)^T$ un vecteur colonne d'entiers donné qui est aussi nommé le vecteur de contraintes, où b_i correspond au nombre de représentants de l'unité u_i à couvrir.

A peut être représenté par une matrice $A = (a_{ij})$, avec a_{ij} le nombre d'instances de l'unité u_i dans la phrase s_j.

$$A = \begin{array}{c} \begin{matrix} s_1 & \cdots & s_j & \cdots & s_n \end{matrix} \\ \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix} \end{array} \begin{array}{l} \\ unité\ u_1 \\ \\ unité\ u_i \\ \\ unité\ u_m \end{array}$$

Une réduction $\mathcal{X}$ de A, appelée aussi une couverture de l'ensemble d'unités U, est définie comme un sous-ensemble de A qui contient pour tous $i \in \{1, \dots, m\}$, au moins b_i instances de u_i. Ceci peut être représenté par un vecteur colonne binaire $X = (x_1, \dots, x_n)^T$ avec $x_j = 1$ si la phrase $s_j \in \mathcal{X}$ et $x_j = 0$ sinon. C'est pourquoi une couverture de U est une solution $X \in \{0,1\}^n$ de la formule $AX \geq B$:

$$\forall i \in \{1, \dots, m\}, \sum_{j=1}^{n} a_{ij} x_j \geq b_i$$

L'objectif est d'optimiser une couverture de manière à minimiser une fonction de coût. Et comme nous visons à réduire la longueur totale de la couverture, nous définissons le coût d'une phrase en considérant une de ses caractéristiques de longueur. Ainsi dans ce cadre, le coût d'une phrase est équivalent à son nombre de phones. Soit $C = (c_1, \dots, c_n)$ le vecteur de coût, avec c_j le coût de la phrase s_j. Le coût d'une couverture est calculé en additionnant le coût de toutes les phrases qui sont incluses dans la couverture :

$$CX = \sum_{j=1}^{n} c_j x_j$$

Le problème de couverture d'ensembles correspond au problème d'optimisation formulé comme suit :

$$X^* = \arg \min_{\substack{X \in \{0,1\}^n \\ AX \geq B}} CX$$

Ce problème algorithmique est NP-difficile [Karp 1972]. Face à un tel problème, le recours à des heuristiques ou à des algorithmes sous-optimaux doit être envisagé pour le traiter en un temps de calcul raisonnable.

3.1.2 Algorithmes d'optimisation de script

Les stratégies itératives (appelées aussi méthodes gloutonnes) construisent, étape par étape, une solution sous-optimale au problème de couverture d'ensembles [Buchsbaum et al. 1996]. Dans chacun des algorithmes que nous présenterons dans la suite, une couverture initiale est déterminée en premier, et à chaque étape de sélection de phrase une modification sur la couverture en cours est envisagée : un ajout, un retrait ou un échange de phrases selon l'algorithme.

Durant l'étape de sélection de phrase, la capacité de couverture μ_j de la phrase s_j est définie comme le nombre de représentants de ses unités qui sont requis dans la couverture, ainsi ce nombre est déterminé en vue du vecteur de contraintes B :

équation 1 : $$\mu_j = \sum_{i=1}^{m} \min\{a_{ij}, b_i\}$$

Nous observons que si la phrase s_j contient $a_{ij} = 12$ instances de u_i et que nous avons besoin d'au moins $b_i = 4$ instances de u_i dans la couverture, alors nous compterons juste 4 instances de u_i qui sera ajouté à μ_j [Barbot et al. 2015].

Algorithme glouton

L'algorithme glouton (agglomerative greedy en anglais) est la stratégie la plus utilisée pour les problèmes de couverture d'ensembles [Van Santen et al. 1997]. Celui-ci part d'une couverture vide. La solution est construite en ajoutant à chaque itération une phrase du corpus initial A qui maximise une fonction de score [François 2002; Chevelu et al. 2007; Cadic 2011; Barbot et al. 2015].

Soit $\tilde{A}$ la matrice associée à l'ensemble des phrases candidates qui est fixé à A, soit $\widetilde{\mu_j} = \mu_j$ la capacité de couverture actuelle de la phrase s_j, et $\tilde{B} = B$ le vecteur de contrainte courant. Le score de la phrase s_j est définit comme suit :

équation 2 : $$\sigma_j = {\widetilde{\mu_j}}/{c_j}$$

À chaque itération une phrase s est insérée dans la couverture. En prenant en considération le contenu de la phrase s, $\tilde{B}$ est mis à jour : $\tilde{B} = \max\{\tilde{B} - \tilde{A}\Delta, 0_{\mathbb{R}^m}\}$ où la j$^{\text{ème}}$entrée de $\Delta = 1$ si $s_j = s$ et $\Delta = 0$ sinon. Après ça, la colonne associée à s dans $\tilde{A}$ est fixée à $0_{\mathbb{R}^m}$. Pour toute phrase s_j avec $\widetilde{\mu_j} \neq 0$, $\widetilde{\mu_j}$ sera mise à jour en utilisant $\tilde{A}$ et $\tilde{B}$ dans l'équation 1 [Barbot et al. 2015].

L'algorithme glouton s'arrête lorsque toutes les contraintes sont satisfaites (toutes les unités sont couvertes), ce qui signifie que $\tilde{B} = 0_{\mathbb{R}^m}$.

Nous obtenons alors l'algorithme suivant :

- $\mathcal{X} = \emptyset$: la couverture qui est la solution à construire
- $\mathcal{F} = A$: le corpus initial
- Tant que ($\exists i, \mathcal{X}$ contient moins que b_i instances de u_i)
    - Choisir la phrase $s_j \in \mathcal{F}$ qui a le score σ_j le plus élevé (équation 2) ;
      Si plusieurs phrases correspondent au même score σ_j, choisir la première d'entre-elles dans le corpus ;
    - $\mathcal{X} \leftarrow \mathcal{X} \cup \{s_j\}$ et $\mathcal{F} \leftarrow \mathcal{F} \setminus \{s_j\}$;
- Répéter

Algorithme cracheur

L'algorithme cracheur (ou glouton inversé) [François 2002; Chevelu et al. 2007; Cadic 2011; Barbot et al. 2015] commence avec une couverture complète : le corpus complet de phrases A. La solution est formée en retirant à chaque itération la phrase de la couverture qui minimise davantage une fonction de score. L'algorithme s'arrête lorsqu'on atteint un seuil où l'exclusion d'une phrase ferait obligatoirement perdre la couverture totale des unités souhaitées.

Algorithme d'échange par paire

L'algorithme d'échange par paire [Kawai et al. 2000; François 2002] débute avec une couverture constituée de phrases choisies aléatoirement du corpus complet et de sorte que sa taille est égale à la taille de la couverture finale souhaitée. À chaque itération deux phrases sont sélectionnées : une phrase qui fait partie de la couverture et une autre phrase qui appartient à son complémentaire dans le corpus complet. Les deux phrases sont permutées si l'échange est plus profitable, c'est-à-dire si la couverture modifiée couvre plus d'unités à moindre coût.

[François et Boëffard 2002] appliquent plusieurs combinaisons de méthodes gloutonnes (glouton, cracheur, échange par paire) à la constitution d'un corpus pour la synthèse vocale en français. Celui-ci contient au moins trois représentants des diphones les plus fréquents. Ils concluent que l'application d'un algorithme glouton suivie d'un algorithme cracheur fournit les meilleurs résultats.

Relaxation lagrangienne

La résolution du problème de couverture d'ensembles en utilisant une heuristique basée sur la relaxation lagrangienne [Caprara et al. 1999] peut fournir une solution exacte pour des problèmes de taille raisonnable. [Chevelu et al. 2007] ont adapté un algorithme appelé LamSCP (Lagrangian based Algorithm for Multi-represented SCP) qui constitue le script de lecture avec une technique de relaxation lagrangienne pour des critères d'optimisation spécifiques. Cet algorithme offre un minorant du coût de la couverture (solution) optimale. [Chevelu et al. 2007; Barbot et al. 2015] démontrent que la solution sous-optimale construite par l'algorithme glouton est au maximum 10.13% plus grande que la solution optimale, et ils envisagent que l'algorithme glouton reste la technique la plus satisfaisante en termes de performance et de temps de calcul pour résoudre le problème de couverture d'ensembles.

3.2 Enregistrement du script de lecture

En premier lieu, il faut choisir le locuteur ou la locutrice qui va lire le script de lecture pour l'enregistrer. Dans certains cas le locuteur est déterminé a priori (par exemple une célébrité, un amateur, etc.), dans d'autres cas une étape de casting serait inévitable pour sélectionner un locuteur parmi un panel de professionnels de la voix (des chanteurs, des comédiens, etc.). La plupart du temps la sélection de la voix se limite à effectuer un choix expert sur un ensemble d'enregistrements existants des différents candidats. Pourtant dans d'autres cadres, la réalisation d'enregistrements précis en conditions réelles est favorisée.

La chaine d'acquisition sonore (la chambre, le matériel d'acquisition, la position du locuteur relativement au microphone, etc.) doit rester invariable durant toute la période

d'enregistrement du script. Celui-ci est souvent réalisé dans une chambre anéchoïques qui sert à diminuer le bruit ambiant et la réverbération à des degrés absolument supportables. L'usage d'un électro-glottographe est répandu qui permet d'identifier les instants de fermeture de la glotte dans les parties voisées de la parole [François 2002; Cadic 2011].

3.3 Post-traitement des données

La phase de post-traitement des données consiste à segmenter et annoter la base de parole enregistrée auparavant. Ceci vise à fournir toutes les informations nécessaires au déroulement de l'algorithme de sélection.

En premier lieu, la base est segmentée généralement en phonèmes. La segmentation peut être effectuée manuellement pour des petites bases, ou automatiquement au moment où l'on dépasse une certaine taille [Boëffard et al. 1993; Ljolje et al. 1993; Nefti et al. 2001]. Des techniques de segmentation adoptées pour la reconnaissance de parole [Young et al. 05; Boëffard et al. 2012] sont notamment utilisées à cet effet. Pour une segmentation plus fine le recours à une verification manuelle est souvent considéré, mais ceci est extrêmement coûteux en termes de temps.

En second lieu, la base est annotée selon les divers critères de l'algorithme de sélection. À ce stade chaque unité acoustique de la base sera annotée conformément à ses caractéristiques contextuelles et acoustiques. L'analyse linguistique des phrases du script d'enregistrement sert à relever les informations contextuelles (phonétiques et linguistiques) des unités. La réalisation d'une analyse prosodique du signal sert à mesurer des paramètres acoustiques comme le pitch (estimation de la fréquence fondamentale F0), les durées des phones et les mesures spectrales.

4 Évaluation des systèmes de synthèse

Notre étude sur l'optimisation de la création de voix de synthèse sous un nouvel angle exige la mise en œuvre des méthodologies d'évaluation de la parole synthétisée. Dans cette section nous présentons ces différentes approches d'évaluation et nous nous concentrons sur l'évaluation objective des signaux acoustiques qui sera exploitée dans l'expérience menée dans la section suivante.

L'évaluation de la qualité de synthèse globale est une problématique délicate vu qu'elle touche à la perception humaine de la parole. Les approches d'évaluation peuvent être généralement classifiées en méthodes objectives ou subjectives.

4.1 Évaluation subjective

Les méthodes subjectives requièrent l'intervention d'évaluateurs humains pour juger les signaux acoustiques en fonction de leur intelligibilité, naturel, qualité, expressivité, agrément, etc. Puisque l'objectif principal de la synthèse vocale est de fabriquer de la parole destinée à des êtres humains, l'évaluation subjective est essentielle pour tout système de synthèse [Chen et al. 1999; Campbell 2007; Chevelu et al. 2015].

Parmi les méthodes subjectives nous pouvons différencier les tests de préférence comme AB et ABX [Kain 2001; Duxans et al. 2004] des tests de score comme MOS, DMOS [ITU-T 1996; Suendermann et al. 2005] et récemment MUSHRA [ITU-R 2015]. L'objectif commun de ces tests est de classer les systèmes de synthèse selon certains critères subjectifs.

Une campagne d'évaluation perceptive à grande échelle est utilisée pour le Blizzard Challenge [Black et al. 2005], mais à chaque fois la quantité d'énoncés à tester est réduite. Ceci est vrai pour la plupart des évaluations subjectives réalisées [Garcia et al. 2006; Hinterleitner et al. 2011; Sainz et al. 2014]. La faible quantité d'échantillons à évaluer est souvent due à la lourdeur (coûteux et prend beaucoup de temps) du processus d'évaluation perceptive.

4.2 Évaluation objective

Les évaluations subjectives nécessitent un grand nombre d'exemplaires à évaluer ainsi qu'un grand nombre d'auditeurs, tous les deux choisis en fonction du domaine d'application du système. Devant l'immensité de cette opération, des études ont été menées pour trouver des critères objectifs fortement corrélés à la perception humaine de la parole. Ainsi l'évaluation objective consiste généralement à calculer une distance entre deux signaux acoustiques : l'un correspond à la voix synthétisée et l'autre correspond à sa référence naturelle [Wouters et al. 1998; Chen et al. 1999; Donovan 2001; Vepa et al. 2002]. Pour le moment, aucune fonction de mesure objective ne permet de substituer complètement l'oreille humaine [Campbell 2007].

Afin de calculer la distance objective entre deux signaux de parole, une technique de paramétrisation est premièrement appliquée sur les deux signaux, puis un coût d'alignement DTW est calculé entre les vecteurs provenant de la paramétrisation en se basant sur une mesure de distance classique. Cette démarche dont nous détaillerons ses différents aspects par la suite est illustrée sur la figure 3.

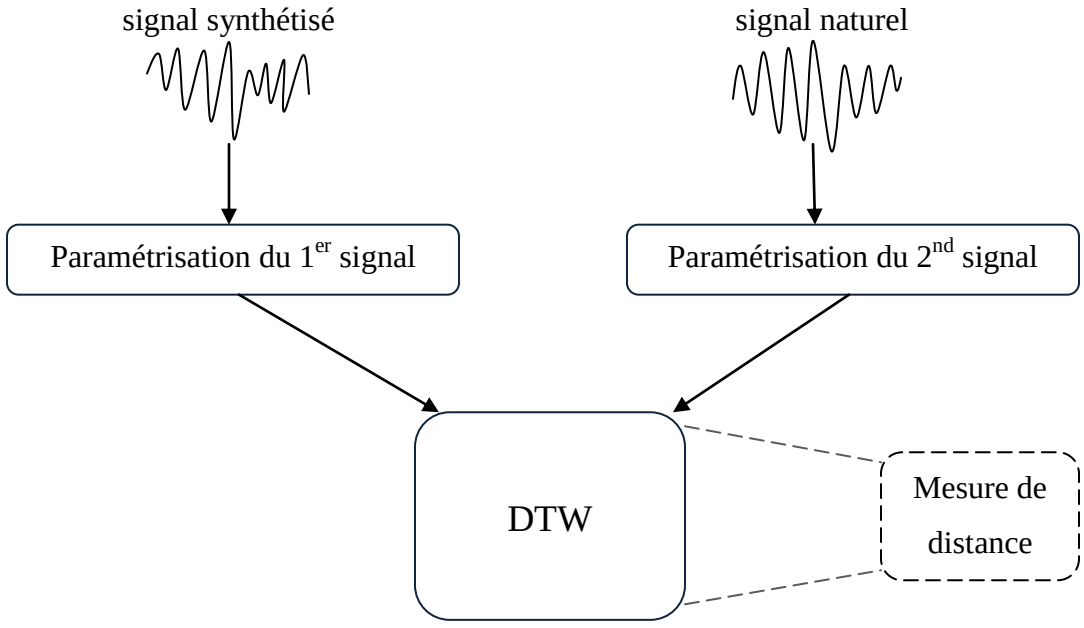

Figure 3 : Calcule de la distance objective entre deux signaux acoustiques.

Paramétrisation

La paramétrisation consiste à analyser un signal de parole dans un espace fréquentiel (spectral). Les propriétés spectrales d'un signal étant plus corrélées à la perception humaine de la parole que ses propriétés temporelles. Les deux signaux acoustiques sont d'abord segmentés en trames. Chaque trame est représentée par plusieurs coefficients de caractérisation du signal. Différentes techniques de paramétrisation du signal peuvent être utilisées dans ce contexte [Chen et al. 1999; Vepa et al. 2002]. Pour citer quelques exemples de paramètres acoustiques :

- Mel-Frequency Cepstral Coefficients (MFCC) ;
- Linear Prediction Coefficients (LPC) ;
- Mel-Generalized Cepstral (MGC) ;
- Multiple Centroid Analysis (MCA) ;
- Line-Spectrum Pairs (LSP) ;
- Line Spectral Frequencies (LSF) ;
- Bispectrum.

Alignement temporel

La voix synthétisée et sa référence naturelle peuvent être de durées inégales et par conséquent d'un nombre de trames différent, ainsi les suites de vecteurs issus de la paramétrisation ne seront pas de dimension identique. Le Dynamic Time Warping (DTW) sera utilisé pour l'alignement [Muda et al. 2010]. Le principe de base de DTW est de mesurer la similarité entre deux séquences temporelles (les deux suites de vecteurs) en cherchant le chemin optimal à traverser parmi l'ensemble des distances entre eux.

Mesures de distance

Des mesures de distance classiques peuvent être calculées entre deux vecteurs d'un paramètre acoustique, c'est-à-dire entre les paramétrisations des deux signaux en question

mentionnées ci-dessus. Nous décrivons brièvement quelques distances utilisées dans ce cadre [Vepa et al. 2002].

La distance absolue entre deux vecteurs :

$$L_1(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$$

La distance euclidienne entre deux vecteurs :

$$L_2(X,Y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}$$

La divergence de Kullback-Leibler [Kullback et Leibler 1951] est une mesure de dissimilarité entre deux distributions de probabilités $f(x)$ et $g(x)$:

$$KL(f,g) = \int \big(f(x) - g(x)\big) \log\left(\frac{f(x)}{g(x)}\right) dx$$

La distance de Mahalanobis [Mahalanobis 1936] est la distance euclidienne normalisée :

$$R(X,Y)^2 = \sum_{i=1}^{n} \left[\frac{x_i - y_i}{\sigma_i}\right]^2 \qquad \text{où } \sigma_i \text{ est l'écart type du i}^{\text{ème}} \text{coefficient des vecteurs}$$

Compte tenu des travaux antérieurs [Vepa et al. 2002], la distance entre les signaux synthétisés et naturels dans notre expérience sera le coût d'alignement DTW selon la distance euclidienne entre les séquences MGC de chaque signal. Ce coût est normalisé en le divisant par la longueur du chemin d'alignement. Plus la valeur de cette distance est faible, plus la parole synthétisée est proche de la voix naturelle et donc probablement de bonne qualité.

5 Expérience

Après avoir présenté le processus classique de création de voix pour les systèmes de synthèse par sélection et les méthodologies utilisées pour évaluer ses systèmes, nous développons ici le sujet principal de notre travail qui est l'optimisation du processus de création de voix de synthèse lorsque le texte à vocaliser est connu à l'avance.

Nous abordons ce point en étudiant la problématique de la création de livres audio à l'aide d'une voix de synthèse. Dans le cas de livres audio, nous connaissons tout à fait le texte du livre que nous voulons synthétiser. Si on enregistre un locuteur qui lit toutes les phrases du livre audio, cela coûte très cher mais on obtient une qualité de voix maximale, et si on n'enregistre aucune de ses phrases cela ne coûte rien mais on n'est pas capable de constituer la voix de synthèse ni de synthétiser les phrases du livre ensuite. Ainsi nous proposons de choisir un sous-ensemble du livre audio à enregistrer et de l'utiliser pour constituer une voix de synthèse. Celle-ci permet de produire le sous-ensemble complémentaire non enregistré du livre audio par synthèse vocale. Notre problème se ramène à l'optimisation d'un compromis entre la quantité de texte à enregistrer et la qualité des signaux acoustiques finaux.

Soit P l'ensemble des n phrases qui composent le livre : $P = \{p_1, p_2, \dots, p_n\}$. Nous voulons partitionner P en deux sous-ensembles V et $\overline{V}$: $P = V \cup \overline{V}$ (bien entendu $V \cap \overline{V} = \emptyset$). V est l'ensemble de phrases textuelles inclues dans P à enregistrer et $\overline{V}$ l'ensemble de phrases textuelles inclues dans P à synthétiser.

Soit C une fonction de coût des phrases à enregistrer : $C : \mathcal{P}(P) \longrightarrow \mathbb{R}$

$$V \longrightarrow C(V)$$

Dans notre cas, nous aspirons à minimiser la longueur totale de V. La longueur de prononciation d'une phrase v de V peut être approximée par son nombre de phones.

$$C(V) = \sum_{v \in V} longeur(v) \qquad avec\ longeur(v) = nombre\ de\ phones\ dans\ v$$

Dans un premier temps, nous tentons à résoudre la problématique formalisée ci-dessus en cherchant à reproduire un livre audio déjà enregistré. En d'autres termes $\forall p \in P$, nous disposons d'une réalisation acoustique de p en contexte produite par le locuteur visé. On note $N(V)$ l'ensemble des signaux naturels relatifs aux énoncés de V et $N(\overline{V})$ l'ensemble des signaux naturels relatifs aux énoncés de $\overline{V}$. Soit $S(N(V), \overline{V})$ l'ensemble des signaux relatifs aux énoncés de $\overline{V}$ qui sont synthétisés en utilisant $N(V)$ comme corpus de parole.

Soit $D\left(N(\overline{V}), S(N(V), \overline{V})\right)$ une fonction d'évaluation objective entre $N(\overline{V})$ et $S(N(V), \overline{V})$.

Nous cherchons la meilleure partition de P, la partition $\{V, \overline{V}\}$ qui permet de produire le livre audio de meilleure qualité et à moindre coût, autrement dit $D\left(N(\overline{V}), S(N(V), \overline{V})\right)$ et $C(V)$ minimales.

Si nous considérons toutes les partitions $\{V, \overline{V}\}$ possibles de P, nous aurions 2^n combinaisons à tester avec n de l'ordre de plusieurs milliers de phrases. L'évaluation d'une de ses partitions - création de la voix de synthèse, synthèse de $\overline{V}$ et évaluation de sa qualité par rapport à $N(\overline{V})$ - peut prendre un temps de calcul considérable (plusieurs dizaines de minutes), en conséquence l'évaluation de toutes les partitions possibles de P ne peut pas être traité en un temps de calcul raisonnable. C'est pourquoi nous évaluerons des partitions particulières (supposées proches d'un optimal) qui seront obtenues en appliquant un algorithme cracheur sur l'ensemble de phrases P du livre audio.

5.1 Algorithme proposé

V est initialisé à l'ensemble de toutes les phrases du livre, $\overline{V}$ est alors initialisé à vide. Les partitions sont formées en retirant à chaque itération une phrase de V et donc en l'ajoutant à $\overline{V}$.

Soit i le nombre de l'itération de l'algorithme cracheur. Soit $\{V_i, \overline{V_i}\}$ la partition de P lors de l'exécution de l'itération i de l'algorithme.

À l'itération i, $\forall v \in V_i$:

1. Constituer la voix de synthèse avec les phrases dans V_i en écartant v : $V_i \setminus \{v\}$;
2. Produire par synthèse vocale les phrases dans $\overline{V_i}$ y compris v : $\overline{V_i} \cup \{v\}$;
3. Mesurer la qualité des phrases synthétisées : $\overline{V_i} \cup \{v\}$.

La phrase v choisie est celle qui renvoie la meilleure qualité de signaux acoustiques finaux (qualité de $\overline{V_i} \cup \{v\}$).

L'algorithme s'arrête lorsque $\forall v \in V_i$, on n'est pas capable de synthétiser $\overline{V_i} \cup \{v\}$ à cause du manque d'unités acoustiques dans V_i, ou lorsque V_i est vide.

Nous obtenons alors l'algorithme cracheur suivant :

- $V = P$
- $\overline{V} = \emptyset$
- Tant que $\Big(\big(\exists v \in V, S\big(N(V \setminus \{v\}), \overline{V} \cup \{v\}\big)$ définie$\big)$ et $(V \neq \emptyset)\Big)$
    - Choisir la phrase $v_o \in V$ qui correspond à la dégradation la plus faible :

$$v_o = \arg\min_{v \in V} D\left(N\big(\overline{V} \cup \{v\}\big), S\big(N(V \setminus \{v\}), \overline{V} \cup \{v\}\big)\right);$$

    - $V \leftarrow V \setminus \{v_o\}$ et $\overline{V} \leftarrow \overline{V} \cup \{v_o\}$;
- Répéter

Dans notre cas, le calcul de la fonction $D\left(N\big(\overline{V}\big), S\big(N(V), \overline{V}\big)\right)$ est effectué comme suit :

$\forall v \in \overline{V}, D\big(N(\{v\}), S(N(V), \{v\})\big)$:

- Retirer l'en-tête des deux signaux $N(\{v\})$ et $S(N(V), \{v\})$;

- Appliquer une technique de paramétrisation[7] sur les deux signaux obtenus ;
- Calculer le coût d'alignement DTW entre les deux séquences de paramétrisation ;
- Normaliser le résultat par la longueur du chemin d'alignement.

$D\left(N(\overline{V}), S(N(V), \overline{V})\right)$ est la somme des résultats de la fonction d'évaluation objective appliquée à toutes les paires de signaux associées aux énoncés de $\overline{V}$:

$$D\left(N(\overline{V}), S(N(V), \overline{V})\right) = \sum_{v \in \overline{V}} D(N(\{v\}), S(N(V), \{v\}))$$

5.2 Mise en œuvre et heuristique d'amélioration

L'algorithme cracheur et la fonction d'évaluation objective ont été implémentés en Python. En ce qui concerne l'implémentation de la fonction d'évaluation objective, nous avons profité de SoX (Sound eXchange) pour supprimer l'en-tête des signaux, et de SPTK (Speech Signal Processing Toolkit) pour calculer les coefficients MGC, le score DTW et la normalisation de ce score.

À chaque étape de sélection de phrase de l'algorithme cracheur, nous devons examiner toutes les phrases incluses dans V. Le test d'une phrase peut prendre un énorme temps de calcul : la création de voix de $V \setminus \{phrase\}$ peut prendre quelque dizaines de minutes, la synthèse de $\overline{V} \cup \{phrase\}$ prend plusieurs dizaines de secondes par phrase et l'évaluation objective de $\overline{V} \cup \{phrase\}$ prend quelques secondes par phrase.

Nous observons que si V contient 300 phrases et $\overline{V}$ contient 50 phrases, le test d'une phrase dans V à cette itération prendra approximativement 24 minutes : 10 minutes pour la création de voix $+ 15 \times 51$ secondes pour la synthèse $+ 2 \times 51$ secondes pour l'évaluation. Alors une telle étape de sélection de phrase de l'algorithme cracheur prendra environ 120 heures (5 jours) : 300 phrases dans $V \times 24$ minutes.

[7]La paramétrisation utilisée ici est MGC.

Afin de minimiser ce temps de calcul, nous procédons en parallélisant l'étape de sélection de phrase et en mettant en place une heuristique d'élagage.

Parallélisation

L'étape de sélection de phrase de l'algorithme cracheur est parallélisée en utilisant du multiprocessus, pour fonctionner sur une machine à 64 cœurs avec 200 GB de RAM.

Heuristique d'élagage

En dépit de la parallélisation de l'étape de sélection de phrase, le temps de calcul reste très important. Une itération de l'algorithme cracheur s'étend à plusieurs heures et peut atteindre plusieurs dizaines d'heures. En partant de ce constat, nous avons mis en place une heuristique d'élagage pour accélérer davantage l'étape de sélection de phrase.

Si en testant une phrase dans V à une itération donnée le moteur de synthèse est incapable de synthétiser $\overline{V} \cup \{phrase\}$, le fait de la tester dans les itérations suivantes sera inutile. Ceci est vrai parce que l'incapacité de synthèse provient du manque d'unités acoustiques dans la voix $V \setminus \{phrase\}$, les unités absentes appartiennent évidemment à la phrase testée (la synthèse est possible avec la voix de V) ; et vu que l'algorithme cracheur vise à retirer à chaque itération une phrase de V, tester cette phrase dans les itérations suivantes signifie que ses unités acoustiques seront de nouveau éliminées, ce qui rend la synthèse infaisable. Ainsi l'heuristique d'élagage est formulée comme suit :

À n'importe quelle étape de sélection de phrase, si l'élimination d'une phrase v de V rend la synthèse de $\overline{V} \cup \{v\}$ impossible, ne plus examiner v dans toutes les étapes de sélection de phrase suivantes.

Soit $i \in \{1, \dots, k\}$ le numéro de l'étape de sélection de phrase de l'algorithme cracheur.
$\forall v \in V_i$, si $S\big(N(V_i \setminus \{v\}), \overline{V_i} \cup \{v\}\big)$ est irréalisable :

$\forall j \in \{i+1, \dots, k\}$, ne plus considérer v dans l'étape j de sélection de phrase.

5.3 Protocole expérimental

Tout d'abord nous introduisons le livre audio utilisé dans notre expérience. Ensuite nous présentons brièvement le moteur de synthèse vocale par corpus de l'équipe Expression de l'IRISA.

Corpus acoustique

Nous utilisons le livre « Albertine disparue » qui est le sixième tome de « À la recherche du temps perdu » de Marcel Proust. Ce livre a été enregistré avec une voix française masculine et expressive. Le corpus acoustique est composé de 3339 phrases et sa durée totale est de 10 heures 45 minutes. Il est annoté automatiquement en appliquant le processus détaillé dans [Boëffard et al. 2012] et géré en se servant du toolkit ROOTS [Chevelu et al. 2014]. La couverture des diphonèmes dans ce corpus est de 78%, pourtant tous les diphonèmes les plus couramment utilisés sont couverts. Ce corpus dispose aussi d'une variabilité acoustique ainsi qu'une haute qualité des signaux de parole. Notre expérience est réalisée sur 10% du corpus de parole introduit ci-dessus. 334 phrases sont tirées au hasard du livre audio complet. Donc le corpus utilisé dans l'expérience, appelé dès maintenant *AudioBook*, est composé de 334 énoncés totalisant 38460 segments (phonèmes et « Non Speech Sounds »[8]).

Système de synthèse de parole

Dans notre expérience, nous utilisons le système de synthèse par sélection d'unités de l'équipe Expression de l'IRISA [Guennec et al. 2014; Alain et al. 2015]. Le toolkit ROOTS [Boëffard et al. 2009; Chevelu et al. 2014] sert à extraire les informations linguistiques et prosodiques du message à synthétiser. Celui-ci est converti en une séquence de phonèmes à l'aide d'un phonétiseur français proposé par [Béchet 2001]. Les étiquettes « Non Speech Sound » peuvent être ajoutées à cette séquence. La sélection des segments acoustiques à concaténer du corpus de parole est ensuite faite en utilisant l'algorithme A^* expliqué dans [Hart et al. 1968; Guennec et al. 2014]. Le processus de sélection est réalisé en deux étapes.

[8] Non Speech Sounds : silences, souffles, événements para-verbaux, etc.

La première étape consiste à extraire pour chaque unité cible, l'ensemble des unités candidates du corpus qui satisfont les mêmes (ou les plus proches) caractéristiques (phonétiques, linguistiques et prosodiques). Des filtres de pré-sélection ont été implémentés pour extraire ces candidats. Une autre fonctionnalité de ces filtres est l'accélération de la recherche d'unités de taille variable, ce qui rend le processus de sélection plus rapide [Conkie et al. 2000]. L'ensemble des filtres de pré-sélection utilisés est le suivant :

1. Le phone est-il un « Non Speech Sound » (NSS) ?
2. Le phone est-il dans l'onset de la syllabe ?
3. Le phone est-il dans la coda de la syllabe ?
4. Le phone est-il dans la dernière syllabe de son groupe de souffle ?
5. La syllabe courante est-elle en fin de mot ?
6. La syllabe courante est-elle en début de mot ?

Étant donné tous les candidats (du point précédent), la deuxième étape consiste à rechercher le meilleur chemin en utilisant un algorithme d'optimisation (A^*) afin de fournir la séquence d'unités vocales finale. Pour évaluer la qualité des concaténations, l'algorithme A^* tâche à minimiser trois sous coûts, couramment utilisés dans les systèmes de synthèse basés sur la sélection d'unités, qui sont les distances calculées entre unité gauche et droite sur les MFCC, l'amplitude et le f0.

6 Résultats et discussion

Dans cette section, nous exposons et nous discutons les résultats de l'expérience décrite dans la section 5. Les résultats présentés ici correspondent aux 100 premières itérations de l'algorithme appliqué sur l'*AudioBook*. Compte tenu des temps de calcul, il manque encore les résultats des itérations ultérieures jusqu'à l'arrêt de l'algorithme. Ceci se produit lorsque n'importe quelle élimination de phrase de V rend la synthèse irréalisable. Compte tenu du fonctionnement du moteur de synthèse, nous estimons que l'arrêt devrait se produire lorsque V contiendra environ 50 phrases. Au bout des 100 itérations, V contenait 234 phrases.

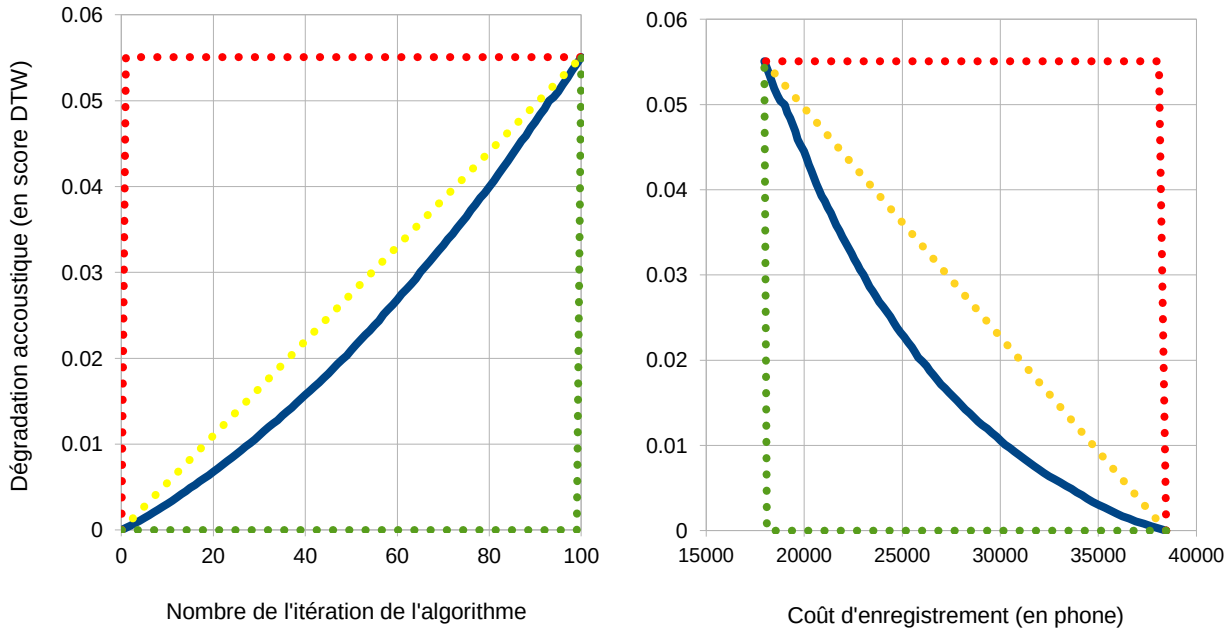

Figure 4 : Dégradation de la qualité de synthèse pour les 100 premières itérations de l'algorithme cracheur. Pour chaque graphique, la courbe rouge pointillée concerne le pire des cas théorique, la courbe verte pointillée concerne le meilleur cas théorique, la courbe jaune pointillée concerne le cas linéaire et la courbe bleue concerne les résultats de l'algorithme cracheur. À chaque itération de l'algorithme cracheur, le coût d'enregistrement diminue alors que la dégradation acoustique augmente.

La figure 4 montre les résultats de l'expérience selon deux graphiques où l'axe des ordonnées est le résultat de la fonction d'évaluation objective appliquée à $\overline{V}$: $D\left(N(\overline{V}), S(N(V), \overline{V})\right)$, et l'axe des abscisses du graphique à gauche est le nombre de l'itération de l'algorithme cracheur alors que celui du graphique à droite est le coût de V en nombre de phones : $C(V)$.

Les deux graphiques de la figure 4 comportent quatre courbes :

- Une courbe relative au pire des cas où l'élimination d'une phrase de V dégrade très fortement la qualité acoustique dans $\overline{V}$ (courbe 1) ;

- Une courbe relative au cas optimal où l'élimination d'une phrase de V, à l'exception de la dernière phrase, n'entraîne aucune dégradation acoustique dans $\overline{V}$ (courbe 2) ;

- Une courbe linéaire où chaque phrase enlevée de V à une étape de l'algorithme cracheur apporte la même dégradation acoustique dans $\overline{V}$ (courbe 3) ;

- Une courbe relative aux résultats des 100 premières itérations de l'algorithme cracheur (courbe 4).

Dans le cas de la courbe 1, la problématique de construction automatique de livres audio avec un sous-ensemble de phrases enregistré du livre visé ne serait jamais intéressante à étudier, parce que nous perdrons extrêmement de qualité dès que nous enlevions un petit nombre de phrases du livre afin de les synthétiser. En contrepartie dans le cas idéal (courbe 2), le sous-ensemble du livre à synthétiser sera celui qui réalisé par synthèse vocale reste de même qualité acoustique. La courbe 3 correspond au cas intermédiaire entre la courbe 1 et la courbe 2.

Nous observons que la courbe 4 (relative à la sortie de l'algorithme cracheur) est sous-linéaire, celle-ci est plus proche du cas idéal que du pire des cas. Alors nous pouvons prévoir que la réalisation d'une telle expérience soit en mesure de nous guider vers une partition optimale du

livre. Nous remarquons aussi que l'élimination des phrases de V durant les premières étapes de l'algorithme cracheur dégrade moins la qualité acoustique dans $\overline{V}$ que l'élimination des phrases durant les étapes de l'algorithme d'après.

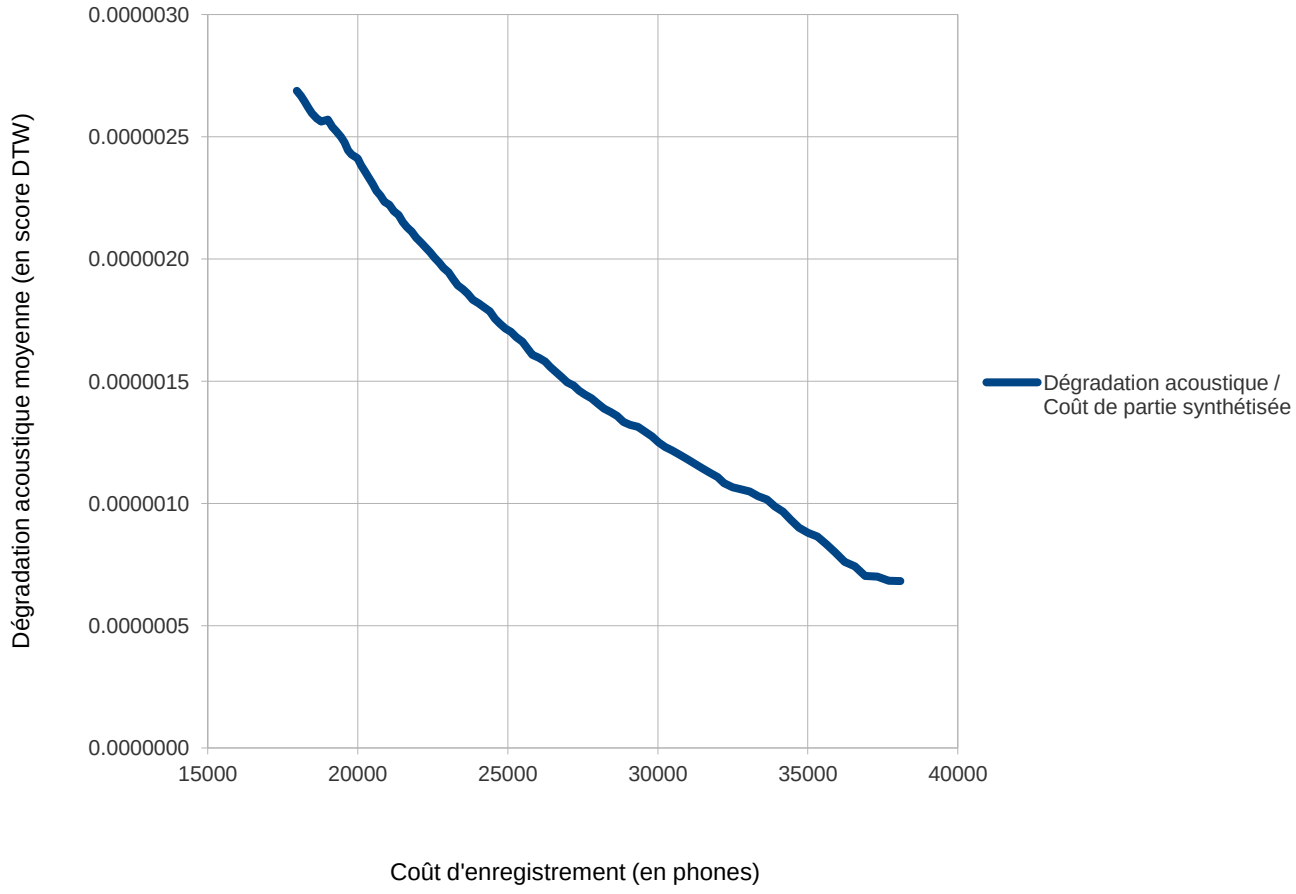

Figure 5 : Dégradation moyenne de la qualité de synthèse de chaque phone pour les 100 premières itérations de l'algorithme cracheur.

La figure 5 montre les résultats normalisés de l'expérience selon un graphique où l'axe des abscisses est le coût de V en nombre de phones : $C(V)$, et l'axe des ordonnées est le résultat de la fonction d'évaluation objective appliquée à $\overline{V}$ normalisé par le coût de $\overline{V}$ en nombre de phones : $D\left(N(\overline{V}), S\left(N(V), \overline{V}\right)\right)/C(\overline{V})$. Ainsi la figure 5 représente la dégradation acoustique moyenne de chaque phone dans $\overline{V}$ pour les différentes étapes accomplies par l'algorithme.

Nous apercevons que plus l'algorithme exclut des phrases de V en les ajoutant dans $\overline{V}$, plus la qualité acoustique moyenne de chaque phone dans $\overline{V}$ s'affaiblit. Nous notons qu'à la première itération de l'algorithme la qualité acoustique de chaque phone dans $\overline{V}$ dégrade de 7×10^{-7} en moyenne, alors qu'à la $100^{ème}$ itération celle-ci dégrade de quatre fois de plus (26×10^{-7} en moyenne).

En conséquent, les résultats illustrés sur la figure 5 viennent renforcer l'idée constatée de l'analyse de la figure 4. Cette expérience permet de mieux définir l'évolution de la qualité acoustique maximale atteignable en fonction du coût d'enregistrement. En effet, la courbe résultante de l'expérience (courbe 4) sert à séparer un espace de voix atteignable localisé au dessus de cette courbe d'un espace inatteignable localisé au dessous d'elle. Ainsi pour un coût d'enregistrement de 30000 phones une dégradation acoustique de 0.02 (en score DTW) des signaux synthétisés est une valeur de dégradation de qualité possible, alors qu'une dégradation acoustique de 0.006 est une valeur de dégradation de qualité impossible.

7 Conclusion

La réalisation de livres audio par synthèse de la parole constitue une problématique de recherche importante à l'heure actuelle. Étant donné que la construction automatique de livres audio avec une voix de synthèse généraliste ne satisfait toujours pas le rendu expressif souhaité dans ce cadre, nous cherchons à examiner une approche qui consiste à partitionner le livre audio en deux sous-ensembles, l'un sera à enregistrer et l'autre sera à synthétiser en créant la voix de synthèse à partir du sous-ensemble enregistré.

À cet effet, nous avons conduit une première expérience établie sur un livre audio déjà enregistré. Nous avons développé un algorithme cracheur pour obtenir des partitions spécifiques du livre, ainsi qu'une fonction d'évaluation objective pour mesurer la qualité de synthèse comparée à celle des réalisations naturelles en contexte. L'algorithme prenait un immense temps de calcul, c'est pourquoi nous avons mis en œuvre une heuristique d'élagage et une parallélisation de l'étape de sélection de phrase.

Malgré les techniques utilisées pour réduire le temps de calcul de l'algorithme, celui-ci reste encore conséquent. Dès lors nous avons choisi d'effectuer l'expérience préliminairement sur 10% du livre audio. Néanmoins la réalisation effective de l'expérience a pris plus de temps qu'escompté en raison des problèmes de stabilité de la machine. Ceux-ci ont nécessité la mise en place d'un mécanisme de relance des processus. De ce fait, nous avons maintenant les résultats des 100 premières itérations de l'algorithme.

Les résultats obtenus tendent à démontrer que la dégradation acoustique issue de la suppression des premières phrases pendant l'exécution de l'algorithme est plus faible que la dégradation acoustique issue de la suppression des phrases suivantes. Ceci est observé aussi bien au niveau phrase qu'au niveau phone. Il semble donc possible de définir un ensemble de phrases du livre à synthétiser en priorité pour obtenir un bon compromis entre le coût d'enregistrement et la qualité acoustique de synthèse.

De plus, l'application de l'algorithme cracheur sur l'*AudioBook* a permis de produire une partition jugée optimale à chacune de ses itérations. Ces partitions constituent une matière d'étude pour le problème générale de construction automatique de livres audio.

Perspectives

Comme signalé précédemment, l'expérience a été conduite sur 10% d'un livre audio et seulement les 100 premières itérations de l'algorithme ont été considérées. Il parait nécessaire de passer à l'échelle, notamment nous devons être capables de compléter les résultats de l'algorithme et de lancer l'expérience sur un livre audio complet. Pour cela il faut améliorer le temps de calcul de l'algorithme cracheur développé.

Puisque le temps de calcul du processus de création de voix de synthèse est de l'ordre de plusieurs dizaines de minutes, une possibilité d'amélioration consisterait à ne faire ce procédé qu'une seule fois[9]. Cela serait possible en introduisant une notion de liste d'unités proscrite qui sera fournie au système de synthèse. Ainsi en fournissant la liste des unités composant $\overline{V}$, on simule le processus de création de voix de $P \setminus \overline{V}$ en économisant l'étape de création de voix.

La solution proposée ci-dessus présente un inconvénient du fait que des paramètres de normalisation sont calculés au moment de création de voix et qu'ils ne pourront pas être mis à jour pour chaque liste d'unités proscrite. Donc ces paramètres utiliseront toujours des informations provenant des unités constituant $\overline{V}$, ce qui introduira un léger biais. Pour minimiser ce biais, nous envisageons d'effectuer le processus de création de voix de V une fois par étape de sélection de phrase.

Pour valider la courbe produite par l'algorithme, il sera nécessaire de mettre en relation la mesure objective de qualité acoustique avec des mesures perceptuelles. Ainsi il faut réaliser des tests d'écoute sur un échantillonnage de partitions fournies par l'algorithme de façon à obtenir les résultats des évaluations subjectives des signaux synthétisés pour des coûts d'enregistrement donnés.

[9]Créer une seule fois la voix de P.

En plus, la courbe validée permettra de positionner des solutions qui pourraient être produites par des algorithmes de réduction de corpus classiques [François 2002; Chevelu et al. 2007; Cadic 2011; Barbot et al. 2015] par rapport à la solution optimale estimée par l'expérience menée dans cette étude.

Des nouveaux travaux de recherche porteront sur l'analyse du contenu des partitions retournées par l'algorithme. À chaque itération, un ensemble de phrases du livre est considéré à éliminer en priorité. Quelles sont les caractéristiques linguistiques et phonologiques des phrases qui composent ces ensembles ? La qualification de leurs spécificités devrait contribuer à l'élaboration des méthodes d'apprentissage automatique (par exemple à base de réseaux de neurones). En conséquence, nous pouvons prévoir l'obtention de la partition optimale d'un livre textuel non enregistré à l'avance.

Références

[Alain et al. 2015] Pierre Alain, Jonathan Chevelu, David Guennec, Gwénolé Lecorvé and Damien Lolive. “The IRISA Text-To-Speech System for the Blizzard Challenge 2015”, In Blizzard Challenge Workshop, Berlin, Germany, Sep 2015.

[Barbot et al. 2015] N. Barbot, O. Boeffard, J. Chevelu, A. Delhay, Large linguistic corpus reduction with SCP algorithms, Computational Linguistics 41(3) : 355-383, 2015.

[Béchet 2001] Béchet, Frédéric. Liaphon: un système complet de phonétisation de textes. Traitement automatique des langues, 42(1):47–67, 2001.

[Bisani et al. 2008] M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, 2008.

[Black et al. 2005] Alan W. Black and Keiichi Tokuda. The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In Proc Interspeech 2005, Lisbon, 2005.

[Black et al. 1994] A. W. Black, and P. Taylor. CHATR: a generic speech synthesis system. 15th conference on Computational linguistics. Association for Computational Linguistics, pp. 983–986, 1994.

[Boëffard et al. 1993] O. Boëffard, B. Cherbonnel, F. Emerard, and S. White. Automatic Segmentation and Quality Evaluation of Speech Unit Inventories for Concatenative-Based, Multilingual PSOLA Text-To-Speech System. In Proceedings of the European Conference on Speech Communication and Technology, volume 1, pages 1449–1452, Berlin, Germany, 1993.

[Boëffard et al. 2009] Olivier Boëffard, Christophe d’Alessandro. “Speech Synthesis”, in J. Mariani, ed. “Spoken Language Processing”, ISBN: 9781848210318, ISTE & J. Wiley, p 99-168, 2009.

[Boëffard et al. 2012] Boëffard, Olivier, Laure Charonnat, Sébastien Le Maguer, Damien Lolive, and Gaëlle Vidal. Towards fully automatic annotation of audiobooks for TTS. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), pages 975–980, 2012.

[Breen et al. 1998] A. P. Breen, and P Jackson. Non-uniform unit selection and the similarity metric within BT’s Laureate TTS system. The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, 1998.

[Buchsbaum et al. 1996] Buchsbaum, A. L. et van Santen, J. P. H.. Selecting Training Inputs via Greedy Rank Covering. In Proceedings of the 7th Annual Symposium on Discrete Algorithms (SODA), pages 288–295, Atlanta, GA, USA, 1996.

[Cadic 2011] Didier Cadic. Optimisation du procédé de création de voix en synthèse par sélection, thèse de l'Université de Paris-Sud 11, 2011.

[Campbell 2007] N. Campbell, Evaluation of speech synthesis : From Reading Machines to Talking Machines, Evaluation of Text and Speech Synthesis, Eds. L. Dybkjær, H. Hemsen, W. Minker, Chapitre 2, 2007.

[Caprara et al. 1999] Caprara Alberto, Matteo Fischetti, and Paolo Toth. 1999. A heuristic method for the set covering problem. Operations Research, 47(5):730–743.

[Chalamandaris et al. 2014] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, S. Raptis. Using Audio Books for Training a Text-to-Speech System. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), pp 3076–3080, 2014.

[Chen et al. 1999] J.D. Chen & N. Campbell. Objective Distance Measures for Assessing Concatenative Speech Synthesis. In Eurospeech 99, éd., Proceedings of the 6th European Conference on Speech Communication and Technology, volume 2, pages 611–614, Budapest, Hungary.

[Chevelu et al. 2007] Chevelu Jonathan, N. Barbot, O. Boeffard and A. Delhay. Lagrangian relaxation for optimal corpus design. In Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6), pages 211–216, 2007.

[Chevelu et al. 2014] Chevelu Jonathan, Gwénolé Lecorvé, and Damien Lolive. ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections. In Proceedings of the 9th International Language Resources and Evaluation Conference (LREC), pp. 619–626, 2014.

[Chevelu et al. 2015] Chevelu Jonathan, D. Lolive, S. Le Maguer and D. Guennec. How to Compare TTS Systems: A New Subjective Evaluation Methodology Focused on Differences. Interspeech, 2015.

[Clark et al. 2007] R. A. Clark, K. Richmond, and S. King. Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication, 49 (4), pp. 317–330, 2007.

[Claveau 2009] V. Claveau. Letter-to-phoneme conversion by inference of rewriting rules. In Proceedings of Interspeech, Brighton, UK, 2009, pp. 1299–1302.

[Conkie et al. 2000] A. Conkie, M. C. Beutnagel, A. K. Syrdal, and P. E. Brown. Preselection of candidate units in a unit selection-based text-to-speech synthesis system. In ICSLP, Vol. 3, pages 314–317, 2000.

[Donovan 2001] Robert E. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesizers. The 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.

[Duxans et al. 2004] H. Duxans, A. Bonafonte, A. Kain, and J.P.H. Van Santen. Including dynamic and phonetic information in voice conversion systems. In International Conference on Spoken Language Processing, 2004.

[François 2002] H. François, Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue, thèse de l'Université de Rennes 1, 2002.

[François et Boëffard 2002] François, Hélène and Olivier Boëffard. 2002. The greedy algorithm and its application to the construction of a continuous speech database. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), volume 5, pages 1420–1426.

[Garcia et al. 2006] M.-n. Garcia, C. D'Alessandro, G. Bailly, P. Boula De Mareüil, and M. Morel. A joint prosody evaluation of french text-to-speech synthesis systems. In LREC, pp. 55–57, 2006.

[Govind et al. 2013] D. Govind, S. R. Mahadeva Prasanna. Expressive speech synthesis : a review. International Journal of Speech Technology, pages 1–24, 2013.

[Guennec et al. 2014] David Guennec and Damien Lolive. Unit selection cost function exploration using an A* based Text-to-Speech system. In Proceedings of the 17th International conference on Text, Speech and Dialogue, 2014.

[Hart et al. 1968] P. E. Hart, N. J. Nilsson, and B. Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on System Science and Cybernetics, vol. 4, pp. 100–107, 1968.

[Hinterleitner et al. 2011] F. Hinterleitner, G. Neitzel, S. Moller, and C. Norrenbrock. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In Proc. Blizzard Challenge Workshop, 2011.

[Hunt et al. 1996] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing Conference. Vol. 1. Ieee, pp. 373–376, 1996.

[Illina et al. 2011] I. Illina, D. Fohr, and D. Jouvet. Grapheme-to-Phoneme Conversion using Conditional Random Fields. In Proceedings of Interspeech, Florence, Italy, 2011, pp. 2313–2316.

[ITU-R 2015] ITU-R Recommendation BS.1534-3. Method for the subjective assessment of intermediate quality level of audio systems. 2015.

[ITU-T 1996] ITU-T Recommendation p.800: Methods for subjective determination of transmission quality. 1996.

[Kain 2001] A. Kain. High Resolution Voice Transformation. Ph.D. thesis, OGI School of Science and Engineering at Oregon Health and Science University, 2001.

[Karp 1972] Karp, Richard M. 1972. Reducibility among Combinatorial Problems. In Miller, R. E. et Thatcher, J. W., éditeurs, Complexity of Computer Computations, pages 85–103. Plenum Press, New York.

[Kawai et al. 2000] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu. A design method of speech corpus for text-to-speech synthesis taking account of prosody. In Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP), volume 3, pages 420–425, Beijing, China, 2000.

[Kawai et al. 2004] H. Kawai, T. Toda, J. Ni, M. Tsuzaki & K. Tokuda. XIMERA : A new TTS from ATR based on corpus-based technologies. In 5th ISCA Workshop on Speech Synthesis (SSW), 2004.

[Kullback et Leibler 1951] S. Kullback and R. Leibler, "On information and sufficiency", Annals of Mathematical Statistics, vol. 22, pp. 79–86, 1951.

[Le Maguer 2013] Sébastien Le Maguer. Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française. PhD thesis. Université de Rennes 1, 2013.

[Ljolje et al. 1993] A. Ljolje and M. D. Riley. Automatic Segmentation of Speech for TTS. In Proceedings of the European Conference on Speech Communication and Technology, pages 1445–1448, Berlin, Germany, 1993.

[Lolive 2008] Damien Lolive. Transformation de l'intonation : application à la synthèse de la parole et à la transformation de voix. PhD thesis, Université de Rennes 1, 2008.

[Mahalanobis 1936] P. C. Mahalanobis, "On the generalised distance in statistics", Proceedings of the National Institute of Sciences of India, vol. 2, n° 1, 1936, p. 49–55.

[Muda et al. 2010] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal Of Computing, Volume 2, Issue 3, March 2010.

[Nefti et al. 2001] S. Nefti and O. Boëffard. Acoustical and topological experiments for an hmm-based speech segmentation system. In Proceedings of the European Conference on Speech Communication and Technology, volume 3, pages 1711–1714, Aalborg, Denmark, 2001.

[Paroubek 2006] Patrick Paroubek, Etiquetage Morphosyntaxique, Technolangue.net, 10 octobre 2006.

[Sagisaka 1988] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. 679–682, 1988.

[Sainz et al. 2014] I. Sainz, E. Navas, I. Hernaez, A. Bonafonte, and F. Campillo. Tts evaluation campaign with a common spanish database. In LREC, pp. 2155–2160, 2014.

[Suendermann et al. 2005] D. Suendermann, A. Bonafonte, H. Duxans, and H. Hoege. Tc-star: Evaluation plan for voice conversion technology. In DAGA 2005, 31st German Annual Conference on Acoustics, Munich, Germany, March, 2005.

[Taylor et al. 1998] P. Taylor, A. W. Black, and R. Caley. The architecture of the Festival speech synthesis system. Proc. of the ESCA Workshop in Speech Synthesis, pp. 147–151, 1998.

[Van Santen et al. 1997] J.P.H. Van Santen & A.L. Buchsbaum. Methods for optimal text selection. In Eurospeech, volume 97, page 2, 1997.

[Vepa et al. 2002] J. Vepa, S. King, P. Taylor. Objective distance measures for spectral discontinuities in concatenative speech synthesis. Interspeech, 2002.

[Viterbi 1967] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 13(2), pp. 260–269, 1967.

[Wouters et al. 1998] J. Wouters and M. Macon. A Perceptual Evaluation of Distance Measures for Concatenative Speech Synthesis. Proc. ICSLP98, pp. 2747–2750, 1998.

[Young et al. 05] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK book, 2005.

[Yamagishi et al. 2008] J. Yamagishi, Z. Ling, and S. King. Robustness of HMM-based speech synthesis. In Proc. of the international conference on speech prosody, pages 581–584, 2008.

[Zipf 1932] G. K. Zipf. Selective studies and the principle of relative frequency in language. Harvard University Press, Cambridge, MA, 1932.