Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept

Raheel Qader[1] Gwénolé Lecorvé[1] Damien Lolive[1] Pascalle Sébillot[2]
(1) IRISA/Université de Rennes 1, 6, rue de Kerampont, 22305 Lannion cedex, France
(2) IRISA/INSA Rennes, Campus de Beaulieu, 263, avenue du Général Leclerc, 35042 Rennes cedex, France
prenom.nom@irisa.fr

RÉSUMÉ

Cet article présente un travail exploratoire sur l'ajout automatique de disfluences, c'est-à-dire de pauses, de répétitions et de révisions, dans les énoncés en entrée d'un système de synthèse de la parole. L'objectif est de conférer aux signaux ainsi synthétisés un caractère plus spontané et expressif. Pour cela, nous présentons une formalisation novatrice du processus de production de disfluences à travers un mécanisme de composition de ces disfluences. Cette formalisation se distingue notamment des approches visant la détection ou le nettoyage de disfluences dans des transcriptions, ou de celles en synthèse de la parole qui ne s'intéressent qu'au seul ajout de pauses. Nous présentons une première implémentation de notre processus fondée sur des champs aléatoires conditionnels et des modèles de langage, puis conduisons des évaluations objectives et perceptives. Celles-ci nous permettent de conclure à la fonctionnalité de notre proposition et d'en discuter les pistes principales d'amélioration.

ABSTRACT

Automatic disfluency insertion towards spontaneous TTS : formalization and proof of concept.

This paper presents an exploratory work on the automatic insertion of disfluencies in text-to-speech systems. By inserting pauses, repetitions and revisions, the objective is to make synthetic speech more spontaneous and expressive. To achieve this task, we formalize the problem as a theoretical process, where transformation functions are iteratively composed. This is a novel contribution since most of the previous work either focus on the detection or cleaning of disfluencies in speech transcripts, or solely concentrate on pause insertion in text-to-speech. We present a first implementation of the proposed process using conditional random fields and language models, before conducting objective and perceptual evaluations. These experiments lead to the conclusion that our proposition is effective to generate disfluencies, and highlights perpectives for future improvements.

MOTS-CLÉS : Disfluences, synthèse de la parole, production automatique de langage naturel.

KEYWORDS: Disfluencies, text-to-speech, natural language generation.

1 Introduction

Les disfluences sont un phénomène qui interrompt le discours sans ajouter aucun contenu propositionnel (Tree, 1995). Ces interruptions apparaissent principalement lorsque l'élocution va plus vite que le processus de pensée, ce qui est particulièrement fréquent en parole spontanée, lorsque le locuteur n'a pas préparé son discours. En dépit de l'absence de contenu propositionnel, les disfluences jouent un rôle important dans le discours. Elles en améliorent la compréhension par des auditeurs, signalent la

complexité de propos à venir (Tree, 2001; Rose, 1998) (cité par Adell *et al.* (2012)) et, en situation de dialogue, facilitent la synchronisation interlocuteurs (Clark, 2002). Malgré cela, les systèmes de synthèse de la parole n'intègrent encore que très partiellement cette notion. À moins que l'utilisateur n'explicite la présence de disfluences dans le texte à synthétiser, les signaux de parole synthétique sont donc généralement peu expressifs et perçus comme relativement peu naturels lorsqu'il s'agit d'imiter un style spontané. En cela, l'ajout automatique de disfluences dans un énoncé en entrée d'un système de synthèse est un problème majeur à résoudre.

Cet article présente notre travail sur l'ajout automatique de disfluences dans le but de permettre la synthèse de parole spontanée. Précisément, nous proposons une formalisation novatrice qui modélise le mécanisme de production de disfluences, et non seulement son résultat. Cette formalisation permet de contrôler la nature et la proportion des disfluences produites. Nous étayons notre proposition par une preuve de concept sous la forme d'une première implémentation appliquée sur un corpus de langue anglaise. Nous nous appuyons sur des méthodes d'apprentissage statistique (les champs aléatoires conditionnels et des modèles de langue n-grammes) pour démontrer expérimentalement la capacité de notre approche à produire des énoncés disfluents plausibles. S'agissant d'un travail exploratoire, aucune expérience en synthèse n'a été conduite car la prise en compte des disfluences dans un système de synthèse nécessite des adaptations autres que la seule transformation du texte en entrée (corpus de parole sous-jacent, prédiction de la prosodie…). Il était donc préférable, comme nous l'avons fait, de s'en tenir dans un premier temps à une validation textuelle des énoncés produits.

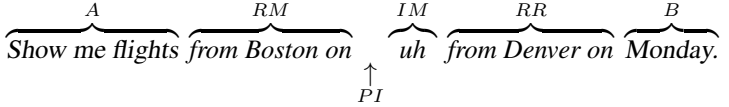
Après une revue du domaine et de nos motivations (section 2), nous présentons notre processus formel de production (section 3) et son implémentation (section 4). Enfin, nous présentons la validation expérimentale de notre travail sous la forme d'évaluations objectives et perceptives (section 5).

2 Revue du domaine et motivations

Historiquement, les disfluences ont été longtemps considérées comme des expressions désordonnées et n'ont ainsi reçu que peu d'attention. Certaines études ont néanmoins montré qu'elles présentent des regularités structurelles (Levelt, 1983; Shriberg, 1994; Clark, 1996). Parmi ces travaux, le schéma proposé par Shriberg, communément admis en traitement automatique du langage, décrit la structure d'une disfluence comme une suite de mots dont certaines sections jouent un rôle particulier. Ainsi, une portion de texte disfluente peut s'écrire comme $\langle A, RM, IM, RR, B\rangle$ où :

- — A et B sont des séquences de mots entourant la disfluence, ces séquences pouvant elles-mêmes comporter des disfluences ;
- — RM est une séquence de mots erronée appelée *reparandum* ;
- — RR est la séquence de mots corrigée correspondant à RM, cette nouvelle séquence étant dénommée *réparation* dans la suite ;
- — IM est l'espace dit *interregnum* qui marque l'interruption du flot de parole par un silence ou des mots dédiés.

Dans cette structure, la frontière entre le *reparandum* et l'espace *interregnum* est appelée *point d'interruption* (PI). Par exemple, la phrase « *Show me flights from Boston on uh from Denver on Monday.* » peut ainsi être découpée comme suit :



(Exemple 1)

En parallèle de cette structuration, de nombreux travaux se retrouvent – quoique la terminologie puisse différer – sur une catégorisation des disfluences en trois grandes familles : les pauses, les répétitions et les révisions (Shriberg, 1999; Tseng, 1999; Boula de Mareüil *et al.*, 2005). Les pauses ont des rôles multiples. Elles permettent par exemple de maintenir active la conversation lorsque le locuteur cherche ses mots ou d'instaurer une proximité avec un interlocuteur. Selon la situation, elles prennent la forme de silences, de mots de remplissage (par exemple, « *uh* » ou « *um* » en anglais) ou de marqueurs du discours (« *well* », « *you know* »…). Les répétitions peuvent, elles aussi, servir à gagner du temps mais elles peuvent également jouer un rôle rhétorique ou annoncer un problème à venir dans le discours. Enfin, les révisions sont des interruptions qui font revenir le locuteur en arrière dans ses propos suite à une erreur (syntaxique ou sémantique). Par cette définition, nous englobons ici également le cas parfois distingué des faux départs. Le rôle des révisions n'est autre que d'aider le locuteur à corriger son discours, comme le montre l'exemple ci-dessous :

$$\textit{I think } \overbrace{\textit{she will}}^{RM} \underset{\uparrow\, PI}{} \overbrace{\textit{I mean}}^{IM} \overbrace{\textit{he will}}^{RR} \textit{ not come today.} \qquad \text{(Exemple 2)}$$

Nous notons au passage que l'espace *interregnum* d'une révision peut contenir des marqueurs explicitant la présence d'une erreur et d'une correction à venir.

La majorité des travaux en TAL s'intéressant aux disfluences se situe dans le domaine de la reconnaissance automatique de la parole (Stolcke & Shriberg, 1996; Stolcke *et al.*, 1998; Liu *et al.*, 2006; Kaushik *et al.*, 2010; Hassan *et al.*, 2014). Ces travaux ont pour principal objectif d'intégrer ce phénomène dans le modèle de langage des systèmes ou de produire des transcriptions automatiques nettoyées de toutes éventuelles disfluences. Ainsi, ils se sont davantage intéressés à la forme de surface des disfluences qu'au processus qui conduit à leur production. Au contraire, les études en synthèse de la parole se sont faites plus rares. Adell *et al.* (2008) expliquent en partie ce décalage par le fait que les corpus de parole sur lesquels les système de synthèse s'appuient ne contiennent souvent pas de disfluences et que les briques de pré-traitement linguistique de ces systèmes (étiquetage morphosyntaxique, analyse syntaxique, prédiction de la prosodie…) ne sont généralement pas capables de fonctionner avec des disfluences, en dépit de progrès en TAL sur ce type de données (Honnibal & Johnson, 2014). Depuis une dizaine d'années, quelques travaux s'intéressent néanmoins aux disfluences en cherchant à insérer des pauses dans les énoncés à synthétiser. Dans (Sundaram & Narayanan, 2003), les auteurs proposent ainsi d'insérer des mots de remplissage (« *uh* » et « *um* ») entre les mots de phrases grâce à des automates à états finis. De leur côté, Adell *et al.* (2007), Dall *et al.* (2014) et Andersson *et al.* (2010) formalisent le problème comme la recherche de points d'interruption par un outil d'apprentissage automatique, puis la sélection parmi un ensemble de possibilités des meilleurs mots à insérer d'après les probabilités données par un modèle de langage, principe général que nous reprenons dans notre proposition. Le recours à une approche automatique s'appuie sur la prédictibilité par des humains des possibles points d'interruption et de la disfluence la plus appropriée pour chacun, tel que démontré dans (Dall *et al.*, 2014). Cette dernière étude montre néanmoins que l'insertion de disfluences n'est pas un phénomène systématique. Dans ces différents travaux, les modèles prédictifs s'appuient généralement sur les mots-formes et sur des étiquettes morphosyntaxiques. Le peu de profondeur de ces descripteurs, et notamment leur manque de lien direct avec la charge cognitive ou émotionnelle d'un locuteur, explique probablement la non-systématicité observée. Malgré cela, les résultats de ce type d'approches ont montré qu'elles permettent de feindre des traits de personnalité (Wester *et al.*, 2015).

Dans l'ensemble, l'étude des disfluences souffrent de certaines limites dans la perspective d'une utilisation en synthèse de la parole. Tout d'abord, les travaux sur l'ajout de disfluences ne s'intéressent

qu'à une unique famille de disfluences (les pauses) et ne proposent rien pour les autres. Récemment, Betz *et al.* (2015) ont proposé des modélisations distinctes pour plusieurs types de pauses. Dans le même esprit mais pour aller plus loin, nous introduisons une formalisation plus riche, capable d'intégrer des répétitions et des révisions. Ensuite, le schéma d'analyse des disfluences majoritairement repris n'est pas à même de décrire le processus de production des disfluences, processus qui permettrait une implémentation d'un algorithme d'ajout automatique. Par exemple, l'analyse de la phrase « *I want to to uh I mean I have to go.* » permet, certes, de reconnaître cette dernière comme disfluente mais elle n'explique pas si elle est le résultat d'une répétition, deux pauses et une révision, ou d'une répétition et une révision, voire uniquement d'une révision. Pour résoudre ce problème, nous proposons un processus, inspiré de ce schéma, qui clarifie ce genre de situation grâce à un mécanisme de composition de disfluences et est compatible avec une modélisation informatique déterministe. Enfin, il est important de noter que l'évaluation des travaux en production de disfluences, comme tous les travaux en production automatique de langage naturel, est difficile puisque plusieurs sorties sont généralement acceptables dans ces problèmes. Cela rend notamment délicat le calcul de mesures objectives lorsque les données étudiées, comme c'est le cas dans notre travail, ne contiennent qu'une unique référence à laquelle se comparer. En marge de notre contribution principale, nous discutons de ce problème dans la section 5.

3 Processus de production de disfluences

Le processus théorique de production de disfluences que nous proposons s'appuie sur un principe de composition de disfluences et sur des définitions de chaque famille de disfluences. Dans cette section, nous présentons les principe généraux de notre processus, traitons le cas de chaque famille, puis détaillons le mécanisme de composition.

3.1 Principes généraux

Dans le processus que nous proposons, une disfluence est vue comme le résultat d'une fonction de transformation d'une phrase fluide. Ainsi, un énoncé avec de multiples disfluences résulte d'une successions de transformations atomiques. Pour cela, nous divisons le schéma générique de Shriberg en sous-schémas, chacun dédié à une famille de disfluences et adapté ses spécificités structurelles, et définissions une fonction de transformation par famille. En supposons la famille T, la fonction de transformation f_T prend en entrée une séquence de n mots $\mathbf{w} \in V^n$, où V désigne le vocabulaire de la langue, et retourne une séquence de m mots, $m > n$, soit :

$$\begin{aligned} f_T : V^n &\to V^m \\ \mathbf{w} &\mapsto f_T(\mathbf{w}). \end{aligned} \tag{1}$$

De la sorte, plusieurs disfluences peuvent être générées par composition de leur fonctions respectives.

En nous inspirant des travaux de l'état de l'art, nous formalisons chaque fonction de transformation comme s'appuyant sur deux sous-fonctions : l'une, notée π_T, qui détermine la position du point d'interruption à considérer, l'autre, ω_T, insère les mots de la disfluence sur la base du résultat de π_T.

Mathématiquement, ces deux fonctions se posent comme suit :

$$\pi_T : \quad V^n \quad \rightarrow \quad [\![0, n]\!] \tag{2}$$

$$\text{et } \omega_T : \quad V^n \times [\![0, n]\!] \quad \rightarrow \quad V^m. \tag{3}$$

et $f_T(\mathbf{w})$ se calcule simplement comme $\omega_T(\mathbf{w}, \pi_T(\mathbf{w}))$.

À l'instar des fonctions f_T, chaque sous-fonction est spécifique à sa famille car elles modélisent le résultat de phénonèmes en partie différents. Ces particularités sont détaillés dans la suite.

3.2 Pauses

Syntaxiquement, les pauses sont de simples interruptions dans un énoncé. Elles ne contiennent aucun *reparandum* ni donc aucune réparation, et se réduisent alors à leur seul segment *interregnum*. Comme évoqué dans la section 2, ce segment peut se matérialiser de différentes manières. Dans notre travail, nous considérons possibles l'ensemble des expressions suivantes : « **silence** », « *uh* », « *um* », « *well* », « *you know* » et « *I mean* ». Cette liste est liée aux annotations présentes dans le corpus utilisé pour nos expériences mais rien n'empêche à terme de l'enrichir pour offrir plus de richesse à l'approche. L'exemple ci-dessous donne un exemple de transformation par une pause d'un énoncé fluide :

$\mathbf{w}$: *once you get to a certain degree of frustration you need to relieve,*

$f_{pause}(\mathbf{w})$: *once you get to a certain degree of* $\overset{IM}{\overbrace{uh}}$ *frustration you need to relieve,*

PI

(Exemple 3)

Pour illuster les sous-fonctions présentées plus tôt, le PI est ici déterminé par la fonction π_{pause} et le choix du/des mots à insérer est fait par la fonction ω_{pause}.

3.3 Répétitions

Étant donné que les répétitions sont la duplication d'une portion de texte, leurs *reparandum* et réparations sont indentiques. Par ailleurs, en raisons du mécanisme proposé de composition, nous considérons que l'espace *interregnum* entre ces deux régions est vide. Toute répétition est donc analogue à l'exemple suivant :

$\mathbf{w}$: *and also I think this happens to a lot of people,*

(Exemple 4)

$f_{répétition}(\mathbf{w})$: *and also* $\overset{RM}{\overbrace{I \quad think}}$ $\overset{RR}{\overbrace{I\ think}}$ *this happens to a lot of people.*

PI

Lors de la duplication des mots à répéter, la sous-fonction $\omega_{répétition}$ a pour rôle de déterminer le nombre de ces mots.

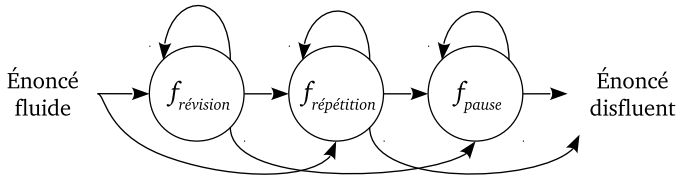


FIGURE 1: Processus complet de production des disfluences.

3.4 Révisions

Dans le même esprit, la fonction de révision $f_{révision}$ s'appuient sur $\pi_{révision}$ pour positionner le PI et $\omega_{révision}$ pour délimiter la zone de réparation, puis produire un *reparandum* lui correspondant, sans aucun espace *interregnum*. À la différence des répétitions, la production du *reparandum* est relativement complexe ici car il s'agit de produire un énoncé factice, éventuellement sous des contraintes de proximité linguistique avec la réparation. En ce sens, il est possible de modéliser les faux départs comme des révisions ou le *reparandum* est complètement différent de la réparation. Un exemple de révision est donné ci-dessous :

$$\begin{aligned} \mathbf{w} &: \textit{that is so that if whoever would get it,} \\ f_{révision}(\mathbf{w}) &: \textit{that is so that } \overbrace{\textit{if you}}^{RM} \underset{PI}{\uparrow} \overbrace{\textit{if whoever}}^{RR} \textit{ would get it.} \end{aligned} \qquad \text{(Exemple 5)}$$

3.5 Mécanisme de composition

La fonction de transformation de chaque famille de disfluences étant responsable de la production de certaines régions de texte, la seule manière de générer de multiples régions ou d'enchaîner l'insertion de différentes disfluences est de composer plusieurs fonctions. Par exemple, une phrase contenant une révision et une pause peut être vue comme le résultat de $f_{révision} \circ f_{pause}$. Néanmoins, cette phrase pourrait également être le résultat de $f_{pause} \circ f_{révision}$. Pour minimiser de telles ambiguïtés et rendre déterministe le processus de production, nous imposons l'ordre de précédence suivant entre familles de disfluences :

$$\text{révision} \prec \text{répétition} \prec \text{pause}. \tag{4}$$

Ainsi, sur l'exemple donné, la composition $f_{pause} \circ f_{révision}$ est interdite. Fondamentalement, cet ordre se justifie par le fait que connaître où sont les révisions et répétitions peut être utile pour déterminer où ajouter des pauses, et lesquelles. En pratique, il est également plus facile d'insérer une pause au milieu de mots répétés que de dupliquer certains mots autour d'une pause. Ensuite, prédire des répétitions avant des révisions pourrait casser une répétition. Au contraire, ajouter des répétitions au sein d'une révision peut permettre de renforcer l'effet produit par cette dernière.

Sur la base de cet ordre, notre processus de production de disfluences est présenté par le graphe de la figure 1. Partant d'une phrase fluide, les transformations de chaque famille peuvent s'appliquer zéro, une ou plusieurs fois. Par conséquent, les fonctions de transformations doivent être capables de traiter en entrée des énoncés fluides comme disfluents. Pour illustrer ce processus sur une phrase en français,

l'énoncé initial « Je souhaite que tu viennes. » pourrait être transformé de la manière suivante :

$$
\begin{aligned}
& \textit{Je souhaite que tu viennes.} \\
f_{révision} \quad & [\textbf{Je veux que } \textit{je souhaite que}]_{\text{rév.}} \textit{ tu viennes.} \\
\circ\ f_{répétition} \quad & \Big[\textit{Je veux } [\textbf{que } \textit{que}]_{\text{rép.}} \textit{ je souhaite que}\Big]_{\text{rév.}} \textit{ tu viennes.} \qquad \text{(Exemple 6)} \\
\circ\ f_{pause} \quad & \Big[\textit{Je veux } [\textit{que que}]_{\text{rép.}} [\textbf{euh}]_{\text{pause}} \textit{ je souhaite que}\Big]_{\text{rév.}} \textit{ tu viennes.} \\
\circ\ f_{pause} \quad & \Big[\textit{Je veux } [\textit{que que}]_{\text{rép.}} [\textit{euh}]_{\text{pause}} [\textbf{enfin}]_{\text{pause}} \textit{ je souhaite que}\Big]_{\text{rév.}} \textit{ tu viennes.}
\end{aligned}
$$

À chaque étape, les portions de textes traitées sont notées entre crochets et les mots insérés sont indiqués en gras. Cet exemple permet de souligner que notre processus ne désambiguïse pas totalement la production de disfluences puisque les deux pauses auraient tout aussi bien pu être insérées dans un ordre inverse. Malgré cela, notre formalisation s'avère plus compatible que le schéma traditionnel de Shriberg avec une utilisation en production. En particulier, les sous-fonctions π_T et ω_T peuvent être vues comme des problèmes d'apprentissage automatique et il est assez direct de fournir une implémentation du processus tel que présenté par la figure 1. C'est ce que décrit la prochaine section.

4 Implémentation

Nous proposons une implémentation possible de notre processus de production de disfluences. L'objectif de cette implémentation est de valider notre approche avant de pouvoir étudier des implémentations plus riches et performantes à l'avenir. Dans cette optique, nous nous limitons à l'ajout de pauses et de répétitions et laissons les révisions, plus difficiles, de côté. Cette situation représente un cas d'étude minimal mais fonctionnel car il permet de tester le principe de composition.

Pour chaque famille de disfluences, nous choisissons de modéliser la prédiction des PI (fonctions π_T) comme une tâche d'étiquetage automatique que nous traitons par des Champs Aléatoires Conditionnels (CAC). L'insertion de nouveaux mots (fonctions ω_T) est, quant à elle, traitée comme l'énumération de toutes les hypothèses disfluentes possibles et la sélection parmi celles-ci de la plus probable d'après un modèle de langage. En résumé, le processsus entier s'appuie donc sur deux CAC et deux modèles de langage. Cette section décrit tout d'abord l'algorithme général de production qui orchestre ces modèles, puis décrit les étapes de prédiction des PI et d'insertion des nouveaux mots.

4.1 Algorithme principal

L'algorithme 1 présente comment transformer une phrase en entrée en un énoncé disfluent. Chaque famille T de disfluences est examinée de la même manière en respectant l'ordre de précédence. L'algorithme tente alors de déterminer un potentiel PI (ligne 11). La validité de ce PI est vérifiée par un critère d'arrêt (l. 12) et, si le PI est valable, une nouvelle disfluence de type T est ajoutée dans la version courante de l'énoncé en cours de transformation (l. 13). Après la mise à jour de l'énoncé, un nouveau potentiel PI est proposé et une nouvelle itération démarre. Dès qu'un PI est rejeté par le critère d'arrêt, l'algorithme passe à la famille de disfluences suivante, ou à défaut retourne l'énoncé transformé. Ce mécanisme permet d'insérer de multiples fois des disfluences du même genre mais

entrée : ÉnoncéOriginal : un énoncé sans disfluence
sortie : énoncé d'entrée avec des disfluences automatiquement ajoutées

données :
Types : liste des familles de disfluences considérées
ÉnoncéTransformé : séquence de mots
PI : entier

Types $\leftarrow$ [*répétition*, *pause*]
ÉnoncéTransformé $\leftarrow$ ÉnoncéOriginal
pour chaque T $\in$ Types **faire**
    PI $\leftarrow \pi_T$(ÉnoncéTransformé)
    tant que $\neg$ *CritèreDArrêt(*T, ÉnoncéTransformé, PI*)* **faire**
        ÉnoncéTransformé $\leftarrow \omega_T$(ÉnoncéTransformé, PI)
        PI $\leftarrow \pi_T$(ÉnoncéTransformé)
retourner ÉnoncéTransformé

ALGORITHME 1**:** Algorithme principal de production de disfluences

aussi de n'en insérer aucune. Par défaut, le critère d'arrêt est défini de telle sorte qu'il stoppe les insertions lorsque, pour le type T en cours de traitement, la proportions de disfluences de ce type dans l'énoncé transformé atteint un seuil maximum fixé par l'utilisateur. En pratique, ces seuils ont respectivement été fixés à $1\,\%$ et $12\,\%$ pour les répétitions et les pauses, ces valeurs correspondant aux proportions observées dans notre corpus.

4.2 Prédictions des points d'interruption

La prédiction des PI est effectuée par un CAC prenant en entrée une séquence de mots (fluide ou disfluente) éventuellement accompagnés de descripteurs. Ce CAC vise à catégoriser les mots successifs sous deux étiquettes : les mots qui sont suivis d'un point d'interruption et les autres. Une fois appris pour une famille de disfluences, le CAC permet d'obtenir une liste de PI qui sont tour à tour examinés jusqu'à trouver un PI qui n'a pas encore été exploité pour insérer une disfluence du type considéré. Cet impératif de nouveauté a été fixé afin d'éviter que la méthode n'insère indéfiniment des disfluences à l'endroit jugé le meilleur par le CAC. Si aucun nouveau PI n'est trouvé, la fonction en informe l'algorithme principal qui passe à la famille suivante. Plusieurs étiquetages de la séquence d'entrée sont fournis par le CAC, par ordre de confiance. Pour chaque hypothèse, les PI éventuellement présents sont triés par ordre décroissant de leur probabilité *a posteriori*, puis ajoutés à la liste des PI potentiels. Ce mécanisme permet en pratique d'avoir un très grand choix de PI et de reporter principalement, comme souhaité, la décision de fin de l'algorithme sur le critère d'arrêt.

4.3 Insertion de nouveaux mots

Étant donné un PI, l'étape d'insertion de nouveaux mots cherche à produire une disfluence qui s'intègrent le mieux possible dans l'énoncé. Pour cela, notre implémentation des fonctions ω_T construit un ensemble de séquences possibles de mots, centrées sur le PI, puis détermine grâce à un modèle de langage n-gramme laquelle est la plus probable pour le type T examiné. Pour les

répétitions, il s'agit, en pratique, de chercher quel couple *reparandum*/réparation est le plus probable. Par exemple, en considérant des répétitions d'au maximum deux mots et l'exemple suivant :

$$\textit{I would like} \underset{\substack{\uparrow \\ IP}}{} \textit{to have a coffee,} \qquad \text{(Exemple 7)}$$

les répétitions possibles considérées sont :

I would like **to to** *have a coffee.*
I would like **to have to have** *a coffee.*

De manière similaire pour les pauses, les six types possibles de pauses considérés sont testés pour jouer le rôle d'espace *interregnum*.

L'évaluation d'une séquence candidate est effectuée en calculant la probabilité d'une fenêtre de mots autour des mots ajoutés. L'utilisation d'une fenêtre de mot au lieu de l'énoncé entier permet de comparer plus clairement les différentes disfluences testées. Dans notre travail, la taille de cette fenêtre est fixée à trois mots à gauche et trois mots à droite. Enfin, pour une séquence donnée, la probabilité retournée par le modèle de langage est pondérée par sa longueur afin de ne pas défavoriser les disfluences de plusieurs mots.

5 Validation expérimentale

Notre implémentation du processus proposé a été testée sur le corpus Buckeye (Pitt *et al.*, 2005), un corpus de parole conversationnelle en anglais. Ces données expérimentales représentent 20 heures de parole spontanée enregistrées auprès de 20 locuteurs lors d'interviews. Entre autres informations, le corpus fournit les transcriptions orthographiques de ces entretiens. Nous avons automatiquement annoté en disfluences ces transcriptions, puis les avons vérifiée manuellement. Les répétitions ont été repérées par leur *reparandum* et réparations, les pauses par leur *interregnum*. Les éventuelles erreurs d'annotation due à l'ambiguïté de certaines situations [1] ont été vérifiée manuellement. Au final, 2 714 répétitions et 20 264 pauses sont considérées. Les versions nettoyées des phrases disfluentes, utilisées notamment en entrée de nos tests, sont facilement obtenables à partir des annotations.

Pour chaque famille de disfluences, une version dédiée du corpus est dérivée de ces données en ne retenant que les phrases contenant au moins une disfluence de cette famille. Pour être cohérent avec l'ordre de précédence, le corpus dédiés aux répétitions a par ailleurs été nettoyé de toutes pauses. Enfin, les corpus ont été divisés en trois jeux de données : l'un pour l'entraînement des modèles (60 % des phrases), un autre de développement pour le réglage de certains hyper-paramètres (20 %) et un ensemble de test (20 %). Pour apporter de la variété à nos modèles, les CAC ont été appris sur les phrases disfluentes ainsi que sur des phrases partiellement nettoyées. Les CAC ont été appris avec Wapiti [2] et les modèles de langage sont des modèles trigrammes appris avec SRILM [3].

Le reste de la section présente les différentes évaluations objectives et perceptives conduites pour valider notre approche.

1. Par exemple, la proposition « *you know* » qui peut être une pause ou une séquence normale.
2. `http://wapiti.limsi.fr/`
3. `http://www.speech.sri.com/projects/srilm/`

5.1 Évaluation objective

Chaque étape de la production de disfluences est évaluée dans notre protocole expérimental. Les prédictions de PI le sont par l'intermédiaire de précisions, rappels et F-mesures comparativement à la référence de notre corpus, c'est-à-dire qu'un PI prédit est considéré comme un vrai positif s'il est placé à l'exacte même position qu'un PI du corpus de référence. En l'absence de références multiples, ce genre de mesures est difficile à interpréter. Ainsi, nous proposons également d'observer le Taux Comparé d'Interruption (TCI) entre les prédictions de nos modèles et la référence, mesure que nous introduisons. Ce taux est calculé comme le facteur d'échelle entre le nombre moyen de PI par phrase dans nos hypothèses et ce même nombre moyen dans la référence. Par exemple, un TCI de 1 indique ainsi une même proportion de PI, de 0,6 une sous-prédiction de $40\,\%$, et de 2,2 une sur-prédiction de $120\,\%$. Les résultats des étapes d'insertion de nouveaux mots sont, quant à eux, évalués par la perplexité attribuée à la séquence de mots produites par les modèles de langage appris sur nos jeux d'entraînement. Puisque ces modèles servent également à décider quels disfluences ajouter, il existe bien sûr un biais dans cette mesure mais celle-ci nous sert avant à analyser et comprendre le comportement général de notre proposition. En toute logique, il est donc attendu que les phrases disfluentes obtiennent des perplexités plus basses que les phrases d'origine.

Lors d'expériences préliminaires sur l'ensemble de développement, ce protocole d'évaluation nous a permis d'étudier les différentes configuration d'apprentissage des CAC de prédiction des PI. Précisément, deux facteurs ont été étudiés et réglés : l'ensemble des atttributs à prendre en compte et la taille du voisinage à considérer autour d'un mot pour décider s'il doit être suivi d'un PI ou non. À la suite de cette étude, il s'avère que nos meilleurs résultats sont obtenus avec très peu d'attributs, à savoir soit les mots seuls, soit accompagnés de leur catégorie grammaticale (POS). Quant au voisinage, il s'avère, assez logiquement, qu'une fenêtre de quelques mots autour du mot en cours d'examen apporte un bénéfice.

Sur l'ensemble de test, les configurations comparées sont : les énoncés nettoyés, disfluents de la référence et ceux produits par nos modèles sans ou avec réglage de l'apprentissage. Dans le cas des pauses, nous considérons également un attribut permettant de savoir si les mots traités par le CAC sont des mots présents dans la phrase originale ou ajoutés lors d'une des précédentes itérations sur les répétitions ou les pauses. Nous cherchons là à savoir si notre mécanisme de composition permet d'intégrer des dépendances entre transformations successives, comme c'est par exemple souhaité pour insérer des pauses au milieu de certaines répétitions. Notons finalement que, de par l'ordre de précédence, les références pour les répétitions ne contiennent aucune pause et que la version nettoyée pour les pauses peut contenir des répétitions.

Les tables 1 et 2 montrent les résultats obtenus respectivement pour les répétitions (R) et les pauses (P). Dans l'ensemble, les résultats sont tout d'abord globalement faibles, en particulier concernant les répétitions. Ces résultats sont à rapprocher de la relativement faible quantité de données d'apprentissage et de l'unicité de notre référence. Les TCI montrent également que, en raison de notre critère d'arrêt, nos productions comportent toujours moins de disfluences que la référence. Nos résultats semblent donc somme toute acceptables pour une première implémentation, à condition d'effectuer une sélection des attributs et un réglage de la taille du contexte comme le montrent les écarts de F-mesure, pour les répétitions comme pour les pauses. Sur le plan de la perplexité enfin, les résultats d'ensemble montrent que nos productions permettent d'obtenir une perplexité comparable aux références, ce qui semble montrer que nos productions se rapprochent de disfluences naturelles. Par ailleurs, les perplexités sur les pauses mettent en avant le biais précédemment évoqué puisque nous notons que la

	Attributs	Fenêtre	Rappel	Précision	F-mesure	TCI	Perpl.
($R_{net.}$)	Énoncés fluides					0,0	241
($R_{réf.}$)	Énoncés disfluents de référence					1,0	236
(R_A)	Mots	non	0,8 %	3,8 %	1,3	0,1	236
(R_B)	+ POS	oui	**6,2 %**	**17,1 %**	**9,2**	0,4	**231**

TABLE 1: Évaluation objective des répétitions sur l'ensemble de test.

	Attributs	Fenêtre	Rappel	Précision	F-mesure	TCI	Perpl.
($P_{net.}$)	Énoncés fluides					0,0	242
($P_{réf.}$)	Énoncés disfluents de référence					1,0	**172**
(P_A)	Mots	non	8,2 %	29,4 %	12,8	0,5	209
(P_B)	+ POS	oui	17,9 %	33,6 %	23,3	0,7	191
(P_C)	+ POS + disfl. préc. préd.	oui	**19,8 %**	**34,5 %**	**25,1**	0,7	188

TABLE 2: Évaluation objective des pauses sur l'ensemble de test.

perplexité est d'autant plus basses que la proportion de disfluences (interprétable à partir du TCI) est élévée. Enfin, les résultats sur les pauses confirment l'intérêt d'intégrer de l'information à propos des précédentes itérations de l'algorithme, comme le montre notamment l'augmentation d'environ 2 points de la F-mesure (P_B *versus* P_C).

5.2 Tests perceptifs

Afin de préciser la validité des phrases produites par notre approche, nous avons conduit deux séries de tests perceptifs. La première série consiste à étudier séparément les effets des répétitions et des pauses, alors que la seconde cherche à mesurer leurs incidences combinées. À partir de textes fluides, les testeurs doivent imaginer comment ceux-ci pourraient être énoncés lors d'une conversation spontanée et donner leur opinion sur plusieurs propositions, grâce à une échelle allant de 0 (énoncé impossible) à 10 (parfaitement possible). 24 adultes non natifs ont participé à ces tests et 40 mêmes textes en entrée sont utilisés pour toutes nos expériences [4]. Pour chaque test, nous rapportons l'opinion moyenne et son intervalle de confiance pour $\alpha = 0,05$.

Les résultats de la première série de tests sont présentées par les figures 2 et 3, respectivement pour évaluer les répétitions et pauses produites. Les noms des systèmes sont les mêmes que ceux indiqués dans les précédentes tables 1 et 2. Ces résultats font tout d'abord apparaître que les opinions moyennes sont globalement très proches d'une configuration à l'autre et que les écarts sont généralement peu significatifs, y compris entre les énoncés nettoyés et de référence. Ceci semble montrer que la perception des disfluences est une tâche difficile pour les testeurs, tout du moins lorsque présentées sous une forme textuelle. Sur les répétitions, il apparaît que les configurations avec aucune ou peu de répétitions ($R_{net.}$ et R_A) sont préférées à celles en contenant plus ($R_{réf.}$ et R_B). Ceci peut s'expliquer par l'absence de pauses dans les énoncés présentés. Ensuite, les résultats sur les pauses tendent

4. Ces textes ont été sélectionnés sur l'ensemble de tests de telle sorte que leur version disfluente de référence (c'est-à-dire telle que présente dans le corpus) contiennent un mélange de répétitions et de pauses. Par ailleurs, les textes très courts (< 4 mots) et trop longs (> 25 mots) ont été écartés.

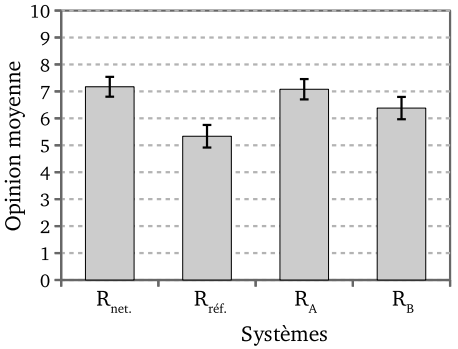


FIGURE 2: Opinions moyennes vis-à-vis des seules répétitions.

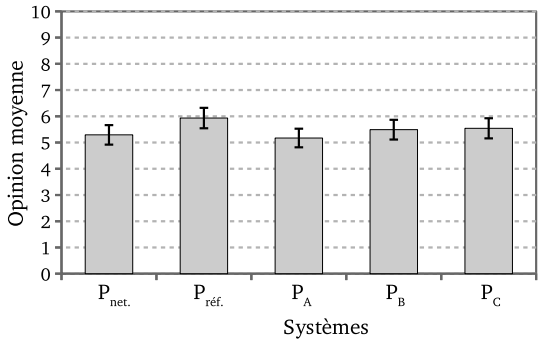


FIGURE 3: Opinions moyennes vis-à-vis des pauses sur la base des répétitions de la référence.

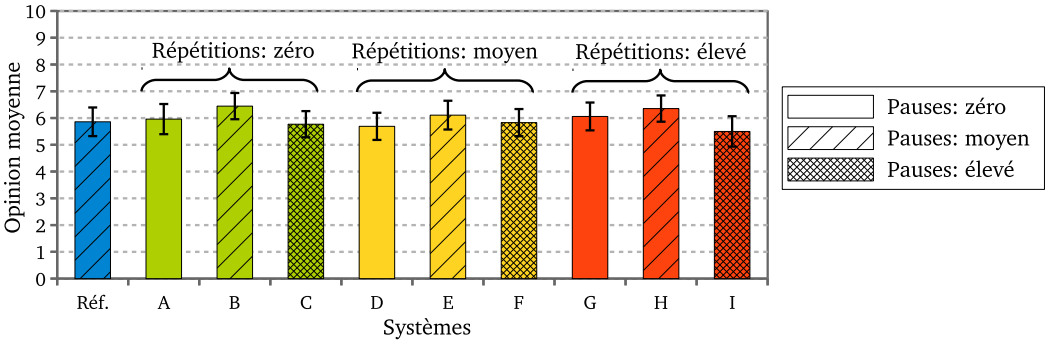


FIGURE 4: Opinions moyennes pour différentes combinaisons de répétitions et pauses.

au contraire à montrer que l'opinion croît avec la proportion de pauses, bien que les écarts soient ici majoritairement non significatifs. Dans l'ensemble, ces deux tests montrent malgré tout que les énoncés produits automatiquement par notre méthode ne dénotent pas par rapport à ceux de référence.

Dans la seconde série d'expériences, nous cherchons à étudier l'incidence des proportions de disfluences sur la perception par les testeurs. Pour cela, nous combinons la production de répétitions et de pauses et fixons pour chaque famille un objectif de production plus ou moins élevé en modulant la sévérité de notre critère d'arrêt (*cf.* algorithme 1). Trois niveaux sont considérés pour les répétitions et les pauses : *zéro*, c'est-à-dire qu'aucune disfluence de la famille considéré n'est produit ; *moyen*, une proportion cohérente avec notre corpus d'apprentissage ; ou *élevé*, de 2 à 4 fois plus de disfluences que dans le corpus. Les résultats de ces nouveaux tests perceptifs sont donnés par la figure 4. À nouveau, dans l'ensemble, les résultats des différentes configurations sont tous très similaires. Nous en déduisons, d'une part, que les compositions de disfluences produisent des énoncés plausibles et, d'autre part, que d'autres questions, plus directes, par exemple sur le degré de spontanéité perçue, pourraient être posées. Il semble néanmoins se dégager deux tendances : l'absence de pauses ou leur forte présence sont moins bien perçues que la situation intermédiaire (B, E, H) et les énoncés très disfluents (I) sont les moins bien perçus.

Globalement, ces tests perceptifs nous montrent donc que les énoncés produits par notre méthode sont

acceptables en comparaison à des textes nettoyés ou disfluents tels que prononcés en situation réelle. Ces résultats tendent donc à valider l'approche que nous proposons. Les faibles différences entre configurations nous incitent néanmoins à améliorer son implémentation ainsi qu'à nous questionner sur une manière plus discriminante de conduire les tests perceptifs.

6 Conclusion

Dans cet article, nous avons présenté une formalisation novatrice des phénomènes de révisions, répétitions et pauses afin de permettre l'ajout automatique de disfluences dans les énoncés en entrée d'un système de synthèse de la parole. L'objectif applicatif de ce travail est de rendre les signaux de parole synthétique plus spontanés et expressifs. Pour cela, nous avons introduit un processus théorique de compositions de disfluences et en avons fourni une première implémentation fondée sur des CAC et des modèles de langage. Les expériences menées sur cette implémentation montrent que le processus proposé est fonctionnel, bien que perfectible.

Parmi les perspectives, la première est donc de désormais produire une implémentation plus performante, si possible apprise sur davantage de données et incluant les révisions. Notons sur ce dernier aspect qu'une difficulté majeure de la production de révisions consiste en la construction de séquences de mots plausibles pour le *reparandum*, c'est-à-dire au minimum en cohérence sémantique ou syntaxique avec l'ensemble de la phrase et idéalement dont l'une des propriétés laisse implicitement comprendre la cause de la disfluence (confusion phonétique, intention du locuteur...). Ensuite, la question de l'évaluation des énoncés est un problème connu dans la littérature. Nos expériences montrent qu'il est nécessaire de s'y intéresser si nous cherchons à l'avenir à quantifier des améliorations entre diverses implémentations. La meilleure piste d'amélioration sur ce point nous semble être l'enregistrement de productions orales des différentes phrases considérées afin de placer le testeur dans des conditions plus réalistes. L'emploi de la synthèse de la parole à cette fin d'évaluation paraît, quant à lui, prématuré car la prédiction de la prosodie accompagnant les disfluences n'est pas encore un sujet maîtrisé. Le risque serait de biaiser la perception des variantes de phrases par des rythmes ou intonations parfois inadéquats. La synthèse automatique des phrases disfluentes reste donc un objectif ambitieux dont la modélisation de la prosodie est l'une des pistes prioritaires.

Références

ADELL J., BONAFONTE A. & ESCUDERO D. (2007). Filled pauses in speech synthesis : towards conversational speech. In *Proceedings of Text, Speech and Dialogue (TSD)*.

ADELL J., BONAFONTE A. & MANCEBO D. E. (2008). On the generation of synthetic disfluent speech : local prosodic modifications caused by the insertion of editing terms. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

ADELL J., ESCUDERO D. & BONAFONTE A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, **54**.

ANDERSSON S., GEORGILA K., TRAUM D., AYLETT M. & CLARK R. A. (2010). Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Proceedings of Speech Prosody*.

BETZ S., WAGNER P. & SCHLANGEN D. (2015). Micro-structure of disfluencies : Basics for conversational speech synthesis. *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

BOULA DE MAREÜIL P., HABERT B., BÉNARD F., ADDA-DECKER M., BARRAS C., ADDA G. & PAROUBEK P. (2005). A quantitative study of disfluencies in french broadcast interviews. In *Proceedings of Disfluency in Spontaneous Speech Workshop*.

CLARK H. H. (1996). *Using Language*. Cambridge University Press.

CLARK H. H. (2002). Speaking in time. *Speech Communication*, **36**.

DALL R., TOMALIN M., WESTER M., BYRNE W. J. & KING S. (2014). Investigating automatic & human filled pause insertion for speech synthesis. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

HASSAN H., SCHWARTZ L., HAKKANI-TÜR D. & TÜR G. (2014). Segmentation and disfluency removal for conversational speech translation. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

HONNIBAL M. & JOHNSON M. (2014). Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, **2**.

KAUSHIK M., TRINKLE M. & HASHEMI-SAKHTSARI A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. In *Proceedings of the Australasian International Conference on Speech Science and Technology (SST)*.

LEVELT W. J. (1983). Monitoring and self-repair in speech. *Cognition*, **14**.

LIU Y., SHRIBERG E., STOLCKE A., HILLARD D., OSTENDORF M. & HARPER M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, **14**.

PITT M. A., JOHNSON K., HUME E., KIESLING S. & RAYMOND W. (2005). The Buckeye corpus of conversational speech : labeling conventions and a test of transcriber reliability. *Speech Communication*, **45**.

ROSE R. L. (1998). *The communicative value of filled pauses in spontaneous speech*. PhD thesis, University of Birmingham.

SHRIBERG E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California.

SHRIBERG E. E. (1999). *Phonetic consequences of speech disfluency*. Rapport interne, DTIC Document.

STOLCKE A. & SHRIBERG E. (1996). Statistical language modeling for speech disfluencies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

STOLCKE A., SHRIBERG E., BATES R. A., OSTENDORF M., HAKKANI D., PLAUCHE M., TÜR G. & LU Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*.

SUNDARAM S. & NARAYANAN S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

TREE J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, **34**.

TREE J. E. F. (2001). Listeners' uses ofum anduh in speech comprehension. *Memory & cognition*, **29**.

TSENG S.-C. (1999). Grammar, prosody and speech disfluencies in spoken dialogues. *Unpublished doctoral dissertation. University of Bielefeld.*

WESTER M., AYLETT M. P., TOMALIN M. & DALL R. (2015). Artificial personality and disfluency. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.