Pauzee : Prédiction des pauses dans la lecture d'un texte

Marion Baranes, Karl Hayek, Romain Hennequin et Elena V. Epure[1]
(1) Deezer Research, Paris, France
research@deezer.com

RÉSUMÉ

Les pauses silencieuses jouent un rôle crucial en synthèse vocale où elles permettent d'obtenir un rendu plus naturel. Dans ce travail, notre objectif consiste à prédire ces pauses silencieuses, à partir de textes, afin d'améliorer les systèmes de lecture automatique. Cette tâche n'ayant pas fait l'objet de nombreuses études pour le français, constituer des données d'apprentissage dédiées à la prédiction de pauses est nécessaire. Nous proposons une stratégie d'inférence de pauses, reposant sur des informations temporelles issues de données orales transcrites, afin d'obtenir un tel corpus. Nous montrons ensuite qu'à l'aide d'un modèle basé sur des transformeurs et des données adaptées, il est possible d'obtenir des résultats prometteurs pour la prédiction des pauses produites par un locuteur lors de la lecture d'un document.

ABSTRACT

Pauzee : Pauses Prediction in text reading.

Silent pauses play a crucial role in text-to-speech synthesis, where they help make the text reading sound more natural. In this work, our goal is to predict these silent pauses from texts to improve automatic reading systems. As this task has not been extensively studied for French, it is necessary to build training data dedicated to the prediction of pauses. We propose a strategy for inferring pauses, based on temporal information from transcribed speech, in order to obtain such a corpus. We then show that with the help of a model based on transformers and appropriate data, it is possible to obtain promising results for the prediction of pauses produced by a speaker during text reading.

MOTS-CLÉS : pauses silencieuses, prédiction des pauses, annotation pour la synthèse vocale.

KEYWORDS: silent break, break prediction, speech synthesis annotation.

1 Introduction

La conversion de textes en contenu audio permet de multiples applications telles que la production de nouveaux médias sonores (*e.g.* livres audio) (Steinhaeusser *et al.*, 2021) ou l'amélioration de l'accessibilité, pour les personnes malvoyantes ou non-voyantes, à des contenus textuels (Freitas & Kouroupetroglou, 2008). Toutefois, obtenir une lecture réaliste et fluide du texte reste un défi en synthèse vocale. Par exemple, sans variations d'emphase, de ton, de rythme ou encore sans pauses de respiration, un texte lu automatiquement semblera monotone et peu naturel (Székely *et al.*, 2020).

Dans ce projet, nous nous intéressons aux pauses silencieuses réalisées par les lecteurs en français, dans le but d'améliorer à terme la lecture automatique d'histoires et d'articles. Nous cherchons notamment à identifier les endroits où les humains marquent une pause silencieuse, consciemment ou non, lors de leur lecture d'un texte à voix haute. Ces pauses peuvent être produites pour des raisons

respiratoires, stylistiques ou encore syntaxiques (Grosman *et al.*, 2018). Elles peuvent être de durée variable. Campione & Véronis (2002b) ainsi que Goldman *et al.* (2010) distinguent par exemple les pauses brèves, moyennes et longues. Grosman *et al.* (2018) définissent les pauses silencieuses comme "une interruption de la phonation". Dans ce papier, nous considérons qu'une interruption de la phonation est une pause silencieuse dès lors qu'elle peut être perçue à l'oreille humaine. Cette notion de perception est délicate à prendre en compte. Pour ce faire, certain travaux choisissent par exemple d'écarter les pauses trop brèves en utilisant des seuils prédéfinis en millisecondes (Grosjean & Deschamps, 1975; Candea, 2000). Toutefois, étant donnée la variabilité des débits de paroles et des durées de pauses, utiliser des seuils fixes peut avoir des conséquences sur les résultats (Campione & Véronis, 2002b; Grosman *et al.*, 2018). Utiliser la ponctuation comme marqueurs de pauses et de respiration est aussi une solution proposée (Wang *et al.*, 2021) pour prédire où un lecteur fera des pauses. Néanmoins, elle peut ne pas convenir à tous les types de textes, notamment les phrases longues et peu ponctuées comme dans les articles de Wikipédia. Définir quand une pause silencieuse peut apparaître est ainsi un problème non trivial.

Comme nous le verrons en section 2, les stratégies utilisées ont beaucoup évolué au fil du temps. Toutefois, ces études ont majoritairement été effectuées pour la langue anglaise (Székely *et al.*, 2019). D'autres travaux se sont penchés sur des tâches similaires, telles que la prédiction de ponctuation (Rei *et al.*, 2021) avec des approches multilingues. Chacun de ces systèmes de prédiction requiert des données d'apprentissage pertinentes dans la langue traitée. Ces données, très spécifiques, peuvent se révéler complexes à trouver pour des langues moins dotées que l'anglais, telles que le français. Or, les construire de toutes pièces n'est pas une tâche aisée. Comme expliqué plus haut, un système se reposant uniquement sur des symboles de ponctuation ou sur des seuils fixés manuellement ne semble pas optimal. Pour pallier cela, nous proposons d'inférer un tel corpus, à partir de données temporelles intermots provenant de transcriptions de données orales, et de l'utiliser pour prédire des pauses. Le système de prédiction de pauses proposé, *Pauzee*, est inspiré des systèmes de prédiction de ponctuation. L'adaptation de tels systèmes à notre tâche permettrait en effet de prédire, dans un texte, les endroits où des pauses pourraient être réalisées en lecture. Dans un second temps, nous nous intéressons au niveau de détail qu'il est possible d'obtenir en tentant de prédire la longueur d'une pause. En synthèse vocale, si toutes les pauses générées sont de la même durée, le résultat restera peu naturel voire peu intelligible. Disposer d'informations à ce sujet est donc primordial. Les contributions principales réalisées dans ce travail sont les suivantes [1]

— La mise en place d'un système d'inférence dynamique des pauses silencieuses à partir d'informations temporelles prenant en considération les variations de ces pauses compte-tenu des différents locuteurs et types de corpus, et permettant ainsi la production d'un corpus annoté.
— Pauzee : l'implémentation d'une nouvelle approche s'appuyant sur des transformeurs pour la prédiction de pauses réalisées lors de la lecture d'un texte.

La suite du papier est organisée comme suit : la section 2 offre un aperçu des travaux déjà faits dans le domaine. La section 3 détaille la construction des données utilisées pour l'apprentissage et l'évaluation du système de prédiction Pauzee. Ce système est ensuite décrit en section 4. Enfin, les résultats obtenus sont détaillés en section 5 et sont suivis d'une conclusion en section 6.

1. Le code développé pour la création du jeu de données utilisé ainsi que pour la prédiction de pauses silencieuses est disponible ici : https://github.com/deezer/pauzee_taln23.

2 État de l'art

Dans le domaine du traitement automatique des langues, la prédiction de pauses a donné lieu à plusieurs travaux. Les premiers sur le sujet s'appuyaient sur des systèmes par règles (Sorin *et al.*, 1987; Bachenko & Fitzpatrick, 1990; Atterer, 2002) et sur des arbres de décision (Ostendorf & Veilleux, 1994; Apel *et al.*, 2004). Puis, l'évolution des techniques en apprentissage automatique a influencé le domaine. Certaines études ont ainsi opté pour des modèles de Markov cachés (Taylor & Black, 1998), des champs aléatoires conditionnels (Keri *et al.*, 2007) ou encore des réseaux de neurones récurrents (Pascual & Bonafonte, 2016). Plus récemment, Székely *et al.* (2019) a choisi d'utiliser un système de classification qui s'appuie sur un réseau de neurones avec deux couches convolutives suivies d'une couche récurrente bidirectionnelle de type bloc LSTM. Cette stratégie permet de prendre en compte un plus long contexte temporel et d'obtenir de meilleures performances. Les auteurs ont notamment repris cette approche dans des travaux ultérieurs (Székely *et al.*, 2020; Alexanderson *et al.*, 2020; Wang *et al.*, 2021).

Toutes ces approches nécessitent des données d'apprentissage dans la langue étudiée. Ces données ne sont pas systématiquement disponibles dans toutes les langues. Pour répondre à ce problème, différentes stratégies sont mises en place. Certaines études prennent le parti d'utiliser des corpus existants, il en existe notamment pour l'anglais. Nous pouvons par exemple citer le Spoken English Corpus qui est annoté en pauses (Taylor & Black, 1998). D'autres, choisissent d'annoter manuellement leurs données (Ostendorf & Veilleux, 1994; Székely *et al.*, 2019). Enfin, une dernière stratégie est de réaliser cette annotation de manière plus automatique. C'est par exemple le cas de Keri *et al.* (2007) qui considèrent tous les silences de plus de 150 ms comme des pauses silencieuses. À notre connaissance, à l'exception de Sorin *et al.* (1987), rares sont les travaux fait sur le français et, par conséquent, rares sont les corpus annotés disponibles.

Comme dit précédemment, les symboles de ponctuation, tels que les points et les virgules, sont souvent associés à une pause dans la lecture (Wang *et al.*, 2021). Campione & Véronis (2002a) montrent d'ailleurs que près de 88% des pauses lors de la lecture apparaissent en présence d'une ponctuation (Campione & Véronis, 2002a). Cette même étude constate que 18,7% des symboles de ponctuations ne provoquent pas de pauses. Bien que non systématique, cette corrélation entre ponctuation et pauses reste observable et suggère que la prédiction de pauses pourrait utiliser des méthodes similaires à celles de la prédiction de ponctuation. En prédiction de ponctuation, les travaux les plus récents utilisent majoritairement des transformeurs. C'est par exemple le cas de Sunkara *et al.* (2020) qui prédit la ponctuation en alignant des caractéristiques lexicales et prosodiques. Plus récemment, SEPP-NLG 2021 (Tuggener & Aghaebrahimian, 2021), la première tâche partagée sur la prédiction de fin de phrase et de ponctuation, a été organisée afin de développer des solutions dans ce domaine. Trois systèmes se sont retrouvés gagnants : OnPoint (Michail *et al.*, 2021), HTW+t2k (Guhr *et al.*, 2021) et Unbabel (Rei *et al.*, 2021). Ces trois systèmes avaient pour point commun d'utiliser des transformeurs.

Pauzee, le système de prédiction de pauses silencieuses que nous voulons mettre en place à besoin de données d'apprentissage. Nous proposons dans ce travail de développer une nouvelle manière d'inférer si un silence annoté en millisecondes peut être considéré comme une pause silencieuse. Pour ce faire nous faisons cet apprentissage sur des données orales transcrites qui ont l'avantage d'être disponibles. Contrairement à Keri *et al.* (2007), nous ne souhaitons pas utiliser un seuil fixe prédéfini. Un telle stratégie risquerait d'ignorer les variations de durée des pauses observables d'un locuteur à un autre (Grosman *et al.*, 2018). Pour éviter cela, nous proposons d'apprendre ces seuils

de manière dynamique, les rendant ainsi adaptables au style de narration du locuteur. À la vue des travaux réalisés dans le domaine, nous avons choisi de nous inspirer des travaux les plus récents et d'évaluer l'intérêt des transformeurs pour une telle tâche. Des travaux en prédiction de ponctuation ayant déjà été réalisés avec des transformeurs, nous proposons de reprendre cette stratégie pour l'adapter à la prédiction de pauses silencieuses.

3 Données

Afin de constituer nos données d'entraînement et d'évaluation, nous proposons d'inférer les pauses produites par les locuteurs à partir de deux corpus de transcriptions d'histoires parlées et lues en français, contenant des données temporelles indiquant le début et la fin de chaque mot. Le premier, le corpus SynPaFlex (Sini *et al.*, 2018), contient des extraits de livres datant du 19ème siècle lus par une seule et unique locutrice. De ce corpus, seuls les textes contenant des informations temporelles ont été conservés [2]. Le second corpus, contenu sur la plate-forme ORFEO (Outils et Ressources sur le Français Ecrit et Oral) (Benzitoun & Debaisieux, 2020), est le French Oral Narrative Corpus (Carruthers *et al.*, 2013). Ce jeu de données regroupe 87 contes oraux narrés en français par des conteurs professionnels et semi-professionnels. Chaque conte n'apparaît qu'une fois dans le corpus, nous ne disposons pas de contes identiques racontés par deux locuteurs différents. Notons que ce corpus est un corpus de parole spontanée ce qui le différencie de SynPaFlex. Bien que le format conté de mémoire des contes se distingue d'un point de vue prosodique du format lu (Levin *et al.*, 1982), il se rapproche toutefois de ce dernier d'un point de vue intentionnel. En effet dans ces deux cas, il est important de pouvoir partager une histoire de manière compréhensible et les pauses y jouent un rôle clé, bien que leur durée et leur fréquence différent. Il est à préciser qu'un prétraitement a été réalisé sur ces deux corpus : tous les symboles de ponctuation ont été retirés. Ce pré-traitement avait pour but d'aligner les deux corpus ensemble, le corpus conté n'étant pas doté de ponctuation. Par ailleurs, retirer la ponctuation de nos corpus présente aussi un intérêt pour notre étude puisque nous ne souhaitons pas développer un système dépendant des signes de ponctuation.

Pour annoter ces corpus, nous proposons d'utiliser les informations temporelles disponibles pour inférer la présence des pauses. Ces informations nous indiquent notamment la durée (en millisecondes) de chaque silence produit entre deux mots. Nous appellerons ici cette information "silence intermot". Précisons que tous les silences intermots, trop courts, ne peuvent être considérés comme des pauses (*e.g.* les occlusives). Toutefois, fixer un seuil strict pour la prédiction de pauses (tel que 150, 200 ou 300ms) n'est pas recommandé en raison de la variabilité temporelle des pauses et de la façon de s'exprimer du locuteur (Campione & Véronis, 2002a; Grosman *et al.*, 2018). Sans seuil, Campione & Véronis (2002a) observent d'ailleurs manuellement des pauses descendant jusqu'à 60ms. Pour définir ce que nous pouvons considérer comme pause ou non, plusieurs analyses de corpus ont été réalisées. Bien que SynPaFlex se distingue du corpus French Oral Narrative Corpus (pauses plus courtes et moins nombreuses), nous ne pouvons savoir si ces différences sont dues au type de corpus ou à la manière de s'exprimer de la locutrice. Pour ces raisons, les observations sur les pauses, décrites ci-dessous, se concentrent sur le corpus Oral Narrative.

La figure 1 présente l'évolution de la durée des silences intermots observés dans le French Oral Narrative, percentile par percentile, pour chaque locuteur. Les courbes représentent les différents

2. Les textes conservés sont : *la fille du pirate*(1878) de Chevalier ; *la vampire* (1865) de Feval, *Madame Bovary* (1857) de Flaubert, *Carmen* (1845) et la vénus d'Ille (1835) de Mérimée, *les mystères de Paris* (1842) de Sue et, *Contes Sénégal et du Niger* (1913) de Zeltner.

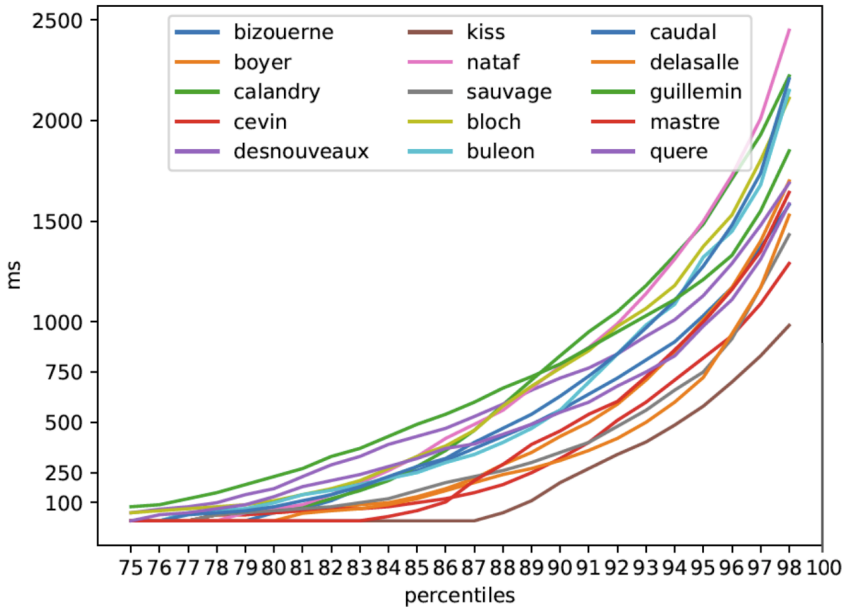

FIGURE 1 – Répartition des pauses faites dans le corpus Oral Narrative

locuteurs, l'ordonnée correspondant au nombre de millisecondes et l'abscisse indiquant les percentiles. Cette visualisation est limitée aux percentiles supérieurs à 75, les durées intermots en dessous étant trop courtes pour être considérées comme des pauses silencieuses. En effet, à l'exception d'un locuteur, à 75 percentiles tous les silences intermots produits sont en dessous de 60ms. Cette illustration offre deux constats : 1) les locuteurs ne font pas le même pourcentage de pauses dans leurs silences intermots (ce pourcentage varie entre 22% et 12%) et 2) la durée d'une pause peut varier d'une personne à l'autre. Par exemple, les pauses les plus longues constatées pour chaque locuteur varient entre 1000ms et 2500ms. Ces observations doivent être interprétées avec prudence : les différences peuvent également provenir du contenu de l'histoire racontée par chaque locuteur, chaque récit étant différent des autres. Suite à cette étude, nous avons choisi de définir un seuil dynamique pour chaque locuteur et chaque histoire. Utiliser un seuil ne dépendant pas d'un nombre de millisecondes prédéfini pourrait offrir une meilleur prise en compte de la fluctuation des pauses et ainsi conduire à une lecture plus fluide et naturelle. Pour cela, nous appuyons sur deux critères : la taille de la pause en millisecondes et le percentile. Suite à notre étude et à la lecture de différentes études citées précédemment, nous considérons qu'une pause doit durer au minimum 80 ms et qu'un narrateur peut faire jusqu'à 22% de pauses silencieuses parmi les silences intermots observés. Pour chaque corpus, nous avons observé la longueur de toutes les pauses apparaissant à partir du percentile 78, puis avons incrémenté ce percentile jusqu'à trouver des longueurs de pauses correspondant à ces critères. Cette stratégie nous permet d'avoir un corpus annoté en pauses prenant en considération leur variabilité compte-tenu du locuteur et du corpus.

Parmi les pauses réalisées en texte lu, se distinguent les pauses brèves, moyennes et longues (Campione & Véronis, 2002b; Goldman *et al.*, 2010; Grosman *et al.*, 2018). Cependant, les seuils qui délimitent

ces pauses varient selon les études : celle de la pause courte est souvent comprise entre 100 ms et 250 ms (Campione & Véronis, 2002b; Goldman *et al.*, 2010; Bailly & Gouvernayre, 2012), celle de la pause moyenne entre 500 ms et 1000 ms (Goldman *et al.*, 2010; Campione & Véronis, 2002b) et à la pause longue correspond aux pauses de durées supérieures (Campione & Véronis, 2002b; Goldman *et al.*, 2010). De même que le seuil qui sépare un silence d'une pause peut varier, nous pensons que les seuils qui distinguent une pause courte, moyenne ou longue peuvent varier selon les locuteurs. On constate sur la figure, il n'y a pas de séparation nette entre les différentes longueurs de pauses, un constat qu'avait également fait Campione & Véronis (2002b). Pour chaque locuteur, nous déterminons les percentiles parmi lesquels se trouvent les pauses, puis nous divisons ces percentiles en trois groupes de tailles égales afin d'obtenir les seuils correspondant aux pauses courtes, moyennes et longues. En moyenne, les seuils séparant les pauses courtes des moyennes sont de 350ms et celles séparant les moyennes des longues est de 950ms. Cette seconde stratégie permet de compléter le corpus annoté en pause par une seconde annotation en longueur de pauses.

La table 1 illustre le nombre de pauses courtes (C), moyennes (M) et longues (L) ainsi que leur total dans chaque corpus. À noter, les corpus ont été divisés au préalable en deux afin de constituer un corpus dédié à l'entraînement (train) et un autre dédié à l'évaluation (test). Pour SynPaFlex, cette division a été effectuée en extrayant 30% des lignes de chaque corpus pour constituer le corpus test et en conservant les autres pour le corpus train. Pour le French Oral Narrative Corpus, une division par ligne n'était pas possible, chaque fichier contenant un mot par ligne. Ainsi, nous avons procédé à la séparation des fichiers de contes en veillant, tant que possible, à ce que les mêmes locuteurs apparaissent à la fois dans les corpus de test et d'entraînement. Pour ce dernier cas, nous avons tenté d'avoir ici aussi un corpus train représentant 70% du corpus initial et un corpus test représentant 30% du corpus initial.

TABLE 1 – Nombre de pauses et de mots contenus dans chaque corpus.

Datasets	# Pauses C	# Pauses M	# Pauses L	# Pauses	# Mots
SynPaflex train	2 702	3 164	3 047	8 913	54 426
Oral Narrative train	6 272	7 481	7 072	20 825	115 912
Total train	8 974	10 645	10 119	29 738	170 338
SynPaflex test	874	1 062	1 017	2 953	19 453
Oral Narrative test	2 659	3 101	2 960	8 720	48 747
Total test	3 533	4 163	3 977	11 673	68 200

4 Système d'apprentissage

Pour cette tâche, nous souhaitons reprendre un système faisant de la prédiction de ponctuation pour l'adapter à notre tâche : la prédiction de pauses silencieuses et de leur longueur. La sélection du système qui nous a servi de point de départ s'est faite sur deux critères : ce système devrait être au niveau de l'état de l'art pour la prédiction de ponctuation et devait être très performant pour le français. Notre choix s'est arrêté sur Unbabel (Rei *et al.*, 2021) qui utilise un modèle multilingue (anglais, français, allemand et italien) proposant de prédire dans un texte les fins de phrases et les signes de ponctuation. Son système a été retenu gagnant à SEPP-NLG 2021 (Tuggener & Aghaebrahimian, 2021) et a obtenu de bons résultats pour le français.

Le système d'Unbabel étend le travail de Miguel Guerreiro *et al.* (2021). Il fonctionne comme suit : un encodeur pré-entraîné reposant sur un transformeur est tout d'abord mis en place. Cela permet la création d'embeddings pour chaque sous-mot et chaque couche du transformeur. Puis ils encapsulent toutes les couches de ce transformeur en un seul embedding à l'aide d'un mécanisme d'attention. Enfin, la dernière étape est dédiée aux têtes de classification qui concatènent les embeddings obtenus. Cela permet de les utiliser comme paramètres afin de prédire d'une part la fin d'une phrase et d'autre part le signe de ponctuation. Pour entraîner leur modèle, Rei *et al.* (2021) utilisent un système d'affinage. Ils combinent ainsi le modèle XLM-Roberta large (Conneau *et al.*, 2020) à un jeu de données annoté pour la prédiction de fin de phase et ponctuation (corpus fourni par les organisateurs de SEPP-NLG 2021 (Tuggener & Aghaebrahimian, 2021)). Les paramètres, bien décrits dans le papier, sont divisés en deux parties : ceux pour XLM-Roberta large et ceux pour les têtes classification. Les paramètres de l'encodeur sont gelés durant les étapes de 0,1% de la première époque. Cela permet aux paramètres pour la classification de s'ajuster à l'objectif de la tâche avant de modifier, et ainsi affiner, ceux pré-entraînés. Entre chaque époque une évaluation est effectuée sur 50% des données. Si aucune amélioration n'est constatée pendant deux époques consécutives, l'entraînement est interrompu.

Notre modèle, Pauzee, reprend ce système et l'adapte à la prédiction de pause. Pour ce faire, nous avons principalement modifié les paramètres qui concernaient la tâche de classification. Nous avons ainsi remplacé les paramètres propres à la ponctuation par des paramètres propres aux pauses. Nous prenons ainsi en compte la prédiction de pauses avec deux valeurs possibles : "absence de pause" et "présence de pause" et la prédiction de la longueur d'une pause avec quatre valeurs possibles : "absence de pause", "pause courte", "pause moyenne" et "pause longue". Concernant l'entraînement nous continuons à utiliser le modèle de langue XLM-Roberta large que nous affinons, non plus au jeu de données fourni par SEPP-NLG 2021, mais aux deux jeux de données (French Oral Narrative Corpus et SynPaFlex) automatiquement annotés en pauses décrits dans la section 3, plus adaptés à notre tâche.

5 Résultats

Ce travail tente de comprendre dans quelle mesure des pauses silencieuses, produites par un humain à partir d'un texte lu, ainsi que leurs longueurs peuvent être prédites. L'évaluation de ces tâches de classification est réalisée avec des métriques classiques : la précision, le rappel et la F-mesure. Ces métriques sont calculées pour chacune des classes étudiées, leur macro-moyenne et leur moyenne pondérée.

Prédire la présence d'une pause à un endroit donné est une tâche de classification binaire. Nous ne disposons malheureusement pas des modèles et des corpus utilisés par les autres travaux décrits dans l'état de l'art et nous ne pouvons donc comparer notre système, Pauzee, aux leurs. Par conséquent, nous proposons d'utiliser deux approches de base :

— La première, "Syllabe", est naïve. Elle repose sur l'idée que les locuteurs font des pauses régulières. Grosjean & Deschamps (1973) montrent que la longueur médiane des espaces entre les pauses sont de 6 syllabes dans une description et de 15 syllabes dans une interview. Nous proposons donc d'ajouter une pause de manière régulière toutes les 7 syllabes. Pour déterminer le nombre de syllabes pris en compte, nous avons testé cette approche de base en allant de 5 à 15 syllabes. Plus le nombre de syllabes était élevé, meilleurs étaient les résultats pour la prédiction d'absence de pauses et moins bons étaient ceux pour la prédiction de pauses.

Le choix concernant nombre de syllabes s'est donc fait sur la macro-moyenne.

— La seconde, "Unbabel", propose de s'appuyer sur le système Unbabel initial entraîné pour de la prédiction de ponctuation. L'idée sous-jacente étant de mesurer l'apport des données d'entraînement que nous avons choisi d'utiliser pour affiner le modèle appris. Unbabel prédit non pas des pauses, mais de la ponctuation. Ainsi, pour prendre en compte ses résultats dans notre évaluation, tous les marqueurs de fin de phrases (".", " !", " ?") ainsi que les virgules et point virgule prédits par Unbabel sont considérés comme une prédiction de pause.

Le tableau 2 montre, pour chaque classe prédite et en moyenne, les résultats obtenus par les approches de base et par Pauzee. La dernière colonne indique le nombre d'éléments concernés.

TABLE 2 – Résultats pour la prédiction de pauses.

Système		Précision	Rappel	F-Mesure	# éléments
Syllabe	classe Absence Pause	0,84	0,81	0,83	56 489
	classe Pause	0,23	0,27	0,25	11 673
	macro-moyenne	0,53	0,54	0,54	68 162
	moyenne pondérée	0,74	0,72	0,73	68 162
Unbabel	classe Absence Pause	0,92	0,93	0,93	56 489
	classe Pause	0,65	0,63	0,64	11 673
	macro-moyenne	0,79	0,78	0,79	68 162
	moyenne pondérée	0,88	0,88	0,88	68 162
Pauzee	classe Absence Pause	0,93	0,95	0,94	56 489
	classe Pause	0,73	0,64	0,68	11 673
	macro-moyenne	0,83	0,79	0,81	68 162
	moyenne pondérée	0,89	0,90	0,90	68 162

Plusieurs études ont montré que les pauses apparaissaient aux niveaux des frontières syntaxiques (Grosman *et al.*, 2018). Dès lors, les résultats obtenus pas la "Syllabe" ne sont pas surprenant et ce d'autant plus sans prise en compte des débuts et fins de phrases (indication absente de nos corpus). En effet, les transformeurs utilisés par les systèmes Unbabel et Pauzee apprennent des représentations vectorielles de mots et de phrases à partir de grandes quantités de textes. Cela leur permet de prendre en compte de manière implicite de nombreuse informations syntaxiques telles que ces frontières. Concernant la tâche de prédiction de l'absence de pause, Unbabel et Pauzee obtiennent des résultats similaires. Concernant la tâche de prédiction d'une pause, les résultats obtenus sont plus éloquents au niveau de la précision obtenue par les deux tâches. Celui obtenu par Pauzee monte à 0,73 alors que celui d'Unbabel reste à 0,65. Le rappel obtenu montre que plusieurs pauses ne sont toujours pas prédites et que ce système d'apprentissage pourrait être encore amélioré. Cela s'explique notamment par le fait que certaines pauses sont plus propres au locuteur qu'au texte lu. Notons toutefois que le fait que le système Unbabel, appris sur de la ponctuation, et le notre, appris sur des pauses silencieuses, obtiennent des résultats similaires illustre bien la corrélation existante entre pause et ponctuation.

Quelques exemples de résultats des systèmes Unbabel et Pauzee sont partagés dans la table 3. Les exemples proviennent des corpus SynPaFlex ("*Carmen*" (1) et "*Madame Bovary*" (2)) et French Oral Narrative ("*Mélusine*" (3) et "*La pierre barbue*" (4)). Chacun de ces passages montrent les différentes découpes du texte en pauses <P> obtenues par Unbabel et par Pauzee. Pour certains passages, les deux propositions semblent acceptables et dépendent plutôt de la manière de raconter du locuteur. C'est par exemple, le cas des extraits 1 et 3. L'exemple 4 est intéressant. Il propose deux manières de raconter un même passage. Unbabel propose une version très ponctuée et contée. Notre système propose

TABLE 3 – Exemples de pauses prédites .

Unbabel	1) *je me sentais près de pleurer* <P> *je lui dis que je reviendrais et je me sauvai* <P> 2) *le soir* <P> *après le maigre diner de son propriétaire* <P> *il remontait à sa chambre et se remettait au travail dans ses habits mouillés qui fumaient sur son corps* <P> *devant le poêle rougi* <P> 3) *on était samedi soir* <P> *Raymondin a reçu la nouvelle tout seul et il a eu toute la nuit pour y penser* <P> *et le dimanche matin* <P> *Mélusine est arrivée tout doucement vers lui* 4) *la hyène* <P> *elle* <P> *rôdait partout* <P> *affamée comme une bête* <P> *quand* <P> *un jour* <P> *elle est allée au fond d' une vallée* <P>
Pauzee	1) *je me sentais près de pleurer* <P> *je lui dis que je reviendrais* <P> *et je me sauvai* <P> 2) *le soir après le maigre diner de son propriétaire* <P> *il remontait à sa chambre* <P> *et se remettait* <P> *au travail dans ses habits mouillés* <P> *qui fumaient sur son corps* <P> *devant le poêle rougi* <P> 3) *on était samedi soir* <P> *Raymondin a reçu la nouvelle tout seul* <P> *et il a eu toute la nuit pour y penser* <P> *et le dimanche matin* <P> *Mélusine est arrivée tout doucement vers lui* <P> 4) *la hyène elle rôdait partout affamée comme une bête* <P> *quand un jour elle est allée au fond d' une vallée* <P>

une version plus épurée, peut être trop pour paraître naturel. L'exemple 2 est plus problématique : Unbabel ne le segmente pas assez et nous fait prononcer 19 mots d'affilée sans aucune pause. Pauzee, lui, le segmente trop. Il propose par exemple une pause après le mot "remettait", ce qui gène la compréhension. Pour être acceptable cette pause devrait être à peine perceptible. Ces exemples nous montrent ainsi les limites des deux systèmes.

La seconde question à laquelle nous souhaitons répondre est relative à la longueur des pauses. Elle tend à préciser notre compréhension des résultats présentés plus haut. Il s'agit, cette fois-ci, d'une tâche de classification en classes multiples. Le tableau 4 illustre les résultats obtenus pour cette seconde tâche. Aux vues des résultats obtenus ici, on constate que la prédiction de l'absence de pauses concorde avec les résultats précédents mais que la prédiction des différents types de pauses est plus complexe à interpréter. La matrice de confusion proposée en figure 2 illustre de manière plus précises les résultats obtenus pour chaque classe. Cette matrice a été normalisée par ligne donc en fonction des pauses attendues.

TABLE 4 – Résultats pour la prédiction de longueur des pauses.

Système	Précision	Rappel	F-Mesure	# éléments
classe Absence Pause	0,90	0,98	0,94	56 489
classe Pause C	0,25	0,00	0,00	3 533
classe Pause M	0,33	0,15	0,21	4 163
classe Pause L	0,48	0,63	0,55	3 977
macro-moyenne	0,49	0,44	0,42	68 162
moyenne pondérée	0,81	0,86	0,82	68 162

On y constate tout d'abord que l'absence de pauses demeure bien prédite. C'est lors de la prédiction des pauses que Pauzee rencontre plus de difficultés. Les pauses courtes ne sont ici jamais prédites. Les pauses moyennes ne sont correctement prédites que pour 15% d'entre elles. Toutefois, on note que 48% d'entre elles sont bien reconnues en tant que pauses. Enfin, les pauses longues sont bien prédites pour 63% d'entre elles et 73% d'entre elles sont reconnues comme pauses. Ainsi, plus une pause est longue, plus elle semble évidente à prédire. Cela s'explique principalement par la variabilité des pauses. Les pauses longues sont probablement celles qui sont les plus communes à tous les

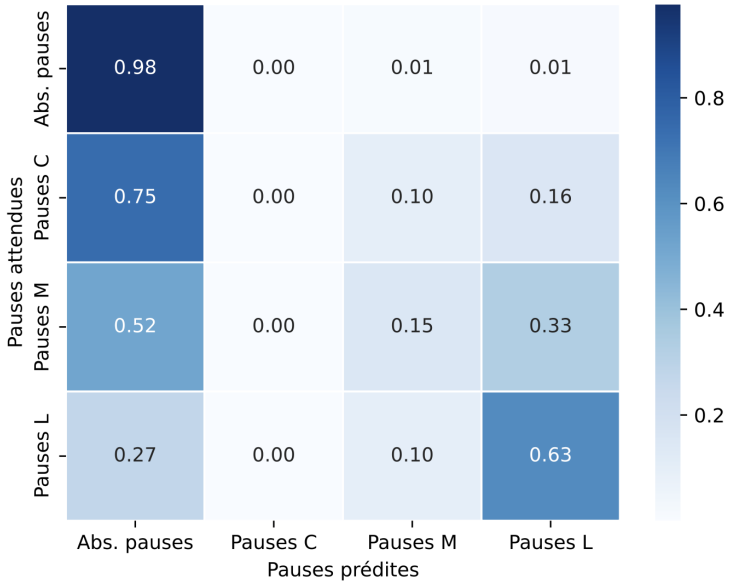

FIGURE 2 – Matrice de confusion des résultats normalisés

locuteurs. Il s'agit de pauses marquant souvent des fins de phrases ou de paragraphes. Leurs contextes d'apparition sont, par conséquent, plus réguliers. Les pauses courtes, quant à elles, semblent plus propres au locuteur. Elles apparaissent généralement au milieu des phrases et ont tendance à être optionnelles, ce qui rend leur prédiction moins claire. Par ailleurs, il est important de noter que plus une pause est brève moins elle est audible. On peut donc se demander si chaque pause courte de notre corpus serait annotée comme telle lors d'une tâche d'annotation manuelle. Les résultats obtenus dans la table 4 sont ainsi dus à cette variabilité mais aussi au fait que notre système d'évaluation est trop strict. Il considère de même importance une erreur portant sur la prédiction d'une pause et une portant sur une confusion entre deux longueurs de pauses. Avec un système plus flexible, capable de considérer que ne pas prédire une pause doit être plus sanctionné qu'une erreur portant sur sa longueur, la F-Mesure de la macro-moyenne peuvent monter à 0,59 et celle de la moyenne pondérée à 0,86.

6 Conclusion

Dans cet article, nous présentons un système de prédiction des pauses silencieuses qui peuvent survenir lors de la lecture à voix haute d'un texte. Cette prédiction peut être utilisée pour améliorer les outils de synthèse vocale. Notre système, Pauzee, utilise un modèle de prédiction de ponctuation basé sur des transformeurs. Une méthode qui, à notre connaissance, n'avait pas été proposée auparavant dans la littérature. Nous avons entraîné Pauzee sur des données inférées plutôt que manuellement annotées car il n'existe pas de données spécifiques pour la prédiction de pauses en français. Ce système d'inférence offre une meilleure prise en compte de la fluctuation des temps de pauses qui peuvent apparaître d'un

locuteur à l'autre et entre différents type de corpus. Cela nous permet d'obtenir des données annotées pour chaque locuteur. Peu bruyant, Pauzee produit des résultats encourageants. Nous pouvons noter les pauses prédites sont en bonne partie corrélées à de la ponctuation, les autres restent toutefois un défi. Nos résultats mettent par ailleurs en évidence la variabilité des pauses silencieuses. Étant donné que chaque élément de notre corpus est lu ou narré par une seule personne, il est difficile de distinguer les pauses spécifiques au locuteur de celles communes à tous les locuteurs. On constate cependant que plus une pause est longue plus elle est prévisible et probablement attendue de tous.

Nos résultats pourraient être améliorés à plusieurs niveaux. Apprendre à prédire plus finement les pauses communes à tous les locuteurs et celles qui sont plus optionnelles est une première piste. Comme nous avons pu le constater, parfois plusieurs annotations semblent acceptables à l'oreille humaine. Par conséquent, une évaluation humaine de ce travail serait pertinente. Même si nos résultats ne sont pas identiques à ceux attendus dans le corpus d'évaluation, ils pourraient finalement être perçus comme valides par des humains. En outre, les seuils dynamiques de notre système d'inférence pourrait être mieux adaptés. La définition de pauses courtes demeure ici encore trop ambiguë et trop permissive. Enfin, les caractéristiques du locuteur ne sont pas prises en compte dans ce travail, les ajouter dans notre système pourrait le rendre plus précis et réduire le nombre d'erreurs.

Références

ALEXANDERSON S., SZÉKELY É., HENTER G. E., KUCHERENKO T. & BESKOW J. (2020). Generating coherent spontaneous speech and gesture from text. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, p. 1–3.

APEL J., NEUBARTH F., PIRKER H. & TROST H. (2004). Have a break ! modelling pauses in german speech. In *KONVENS*, p. 5–12.

ATTERER M. (2002). Assigning prosodic structure for speech synthesis : a rule-based approach. In *Speech Prosody 2002, International Conference*.

BACHENKO J. & FITZPATRICK E. (1990). A computational grammar of discourse-neutral prosodic phrasing in english. *Computational Linguistics*, **16**, 155–170.

BAILLY G. & GOUVERNAYRE C. (2012). Pauses and respiratory markers of the structure of book reading. In *Interspeech 2012-13th Annual Conference of the International Speech Communication Association*, p. Thu–O9d.

BENZITOUN C. & DEBAISIEUX J.-M. (2020). Orféo : un corpus et une plateforme pour l'étude du français contemporain. HAL : hal-03011344.

CAMPIONE E. & VÉRONIS J. (2002a). Etude des relations entre pauses et ponctuations pour la synthèse de la parole à partir de texte. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 177–186, Nancy, France : ATALA.

CAMPIONE E. & VÉRONIS J. (2002b). A large-scale multilingual study of silent pause duration. In *Speech prosody 2002, international conference*.

CANDEA M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits" d'hésitation" en français oral spontané. Etude sur un corpus de récits en classe de français.* Thèse de doctorat, Université de la Sorbonne nouvelle-Paris III.

CARRUTHERS J. *et al.* (2013). French oral narrative corpus. *Oxford Text Archive Core Collection*.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

FREITAS D. & KOUROUPETROGLOU G. (2008). Speech technologies for blind and low vision persons. *Technology and Disability*, **20**, 135–156. DOI : 10.3233/TAD-2008-20208.

GOLDMAN J.-P., FRANÇOIS T., ROEKHAUT S., SIMON A. C. *et al.* (2010). Étude statistique de la durée pausale dans différents styles de parole. *Journées d'Etude sur la Parole (JEP)*.

GROSJEAN F. & DESCHAMPS A. (1973). Analyse des variables temporelles du français spontané. *Phonetica*, **28**(3-4), 191–226.

GROSJEAN F. & DESCHAMPS A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, **31**(3-4), 144–184.

GROSMAN I., SIMON A. C. & DEGAND L. (2018). Variation de la durée des pauses silencieuses : impact de la syntaxe, du style de parole et des disfluences. *Langages*, **211**, 13–40. DOI : 10.3917/lang.211.0013.

GUHR O., SCHUMANN A.-K., BAHRMANN F. & BÖHME H.-J. (2021). Fullstop : Multilingual deep models for punctuation prediction. In *Swiss Text Analytics Conference*.

KERI V., PAMMI S. C. & PRAHALLAD K. (2007). Pause prediction from lexical and syntax information. In *Proceedings of International Conference on Natural Language Processing (ICON)*.

LEVIN H., SCHAFFER C. A. & SNOW C. (1982). The prosodic and paralinguistic features of reading and telling stories. *Language and speech*, **25**(1), 43–54.

MICHAIL A., WEHRLI S. & BUCKOVÁ T. (2021). Uzh onpoint at swisstext-2021 : Sentence end and punctuation prediction in nlg text through ensembling of different transformers (short paper). In *Swiss Text Analytics Conference*.

MIGUEL GUERREIRO N., REI R. & BATISTA F. (2021). Towards better subtitles : A multilingual approach for punctuation restoration of speech transcripts. *Expert Systems with Applications*, **186**, 115740. DOI : 10.1016/j.eswa.2021.115740.

OSTENDORF M. & VEILLEUX N. M. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, **20**(1), 27–54.

PASCUAL S. & BONAFONTE A. (2016). Prosodic break prediction with rnns. In *Advances in Speech and Language Technologies for Iberian Languages : Third International Conference, IberSPEECH*, p. 64–72. DOI : 10.1007/978-3-319-49169-1_7.

REI R., BATISTA F., GUERREIRO N. M. & COHEUR L. (2021). Multilingual simultaneous sentence end and punctuation prediction (short paper). In *Swiss Text Analytics Conference*.

SINI A., LOLIVE D., VIDAL G., TAHON M. & DELAIS-ROUSSARIE É. (2018). SynPaFlex-corpus : An expressive French audiobooks corpus dedicated to expressive speech synthesis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, p. 328–333, Miyazaki, Japan : European Language Resources Association (ELRA). HAL : hal-01826690.

SORIN C., LARREUR D. & LLORCA R. (1987). A rhythm-based prosodic parser for text-to-speech systems in french. *XIème Congrès International des Sciences Phonétiques*, p. 125–128.

STEINHAEUSSER S. C., SCHAPER P. & LUGRIN B. (2021). Comparing a robotic storyteller versus audio book with integration of sound effects and background music. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, p. 328–333, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3434074.3447186.

SUNKARA M., RONANKI S., BEKAL D., BODAPATI S. B. & KIRCHHOFF K. (2020). Multimodal semi-supervised learning framework for punctuation prediction in conversational speech. In *Interspeech 2020*, p. 4911–4915. DOI : 10.21437/Interspeech.2020-3074.

SZÉKELY É., HENTER G. E., BESKOW J. & GUSTAFSON J. (2020). Breathing and speech planning in spontaneous speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7649–7653 : IEEE.

SZÉKELY É., HENTER G. E. & GUSTAFSON J. (2019). Casting to corpus : Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6925–6929 : IEEE.

TAYLOR P. & BLACK A. W. (1998). Assigning phrase breaks from part-of-speech sequences. *Comput. Speech Lang.*, **12**(2), 99–117. DOI : 10.1006/csla.1998.0041.

TUGGENER D. & AGHAEBRAHIMIAN A. (2021). The sentence end and punctuation prediction in nlg text (sepp-nlg) shared task 2021. In *Swiss Text Analytics Conference*.

WANG S., ALEXANDERSON S., GUSTAFSON J., BESKOW J., HENTER G. E. & SZ'EKELY E. (2021). Integrated speech and gesture synthesis. *Proceedings of the 2021 International Conference on Multimodal Interaction*.