Disfluences et erreurs d'alignement au niveau du phonème : le cas des consonnes de liaison en français

Mathilde Hutin[1] Caihong Weng[2] Martine Adda-Decker[1,3] Ioana Vasilescu[1] Lori Lamel[1]
(1) Université Paris-Saclay, CNRS UMR 9015, LISN, 91400, Orsay, France
(2) Université de Paris, 75013, Paris, France
(3) Université Paris 3 / Sorbonne Nouvelle, CNRS UMR 7018, LPP, 75005, Paris, France
{mathilde.hutin, ioana.vasilescu, lori.lamel}@lisn.up-saclay.fr,
caihong.weng@etu.u-paris.fr, martine.adda-decker@sorbonne-nouvelle.fr

RESUME

Les disfluences (pauses, hésitations, répétitions, certains marqueurs discursifs...) sont caractéristiques de l'oralité mais représentent un défi pour le traitement automatique des langues. Nous analysons ici plus de 150h de français en combinant alignement automatique avec variantes de prononciation et jugement humain pour identifier les causes des erreurs d'alignement sur des phonèmes particuliers du français : les consonnes de liaison. Les résultats montrent que les disfluences sont en effet la deuxième source d'erreurs. Parmi elles, la plus représentée est la répétition (chez les femmes et les hommes), qui favorise l'alignement erroné avec [t] ou [z], suivi de la pause silencieuse, qui favorise [p]. En comparant ces données mal alignées à un ensemble similaire de données bien alignées, nous montrons que la présence de disfluences dans un empan de quatre mots est corrélée aux erreurs d'alignement, mais que cet effet s'annule si la disfluence est à au moins un mot de distance du site de liaison.

ABSTRACT

Disfluencies and alignment errors at the phonemic level: The case of French *liaison* consonants. Dysfluencies (pauses, hesitations, repetitions, some discourse markers…) are highly frequent in language use yet remain a challenge to automatic speech processing. We analyze 150+ hours of spoken French combining automatic alignment with pronunciation variants and human judgement to identify the sources of alignment errors on a particular class of French phonemes: *liaison* consonants. Results show that disfluencies are indeed the second most represented source of error. Among them, the most represented type is repetition (in female like in male speech), favoring the alignment of erroneous [t] or [z], followed by silent pause, favoring the alignment of erroneous [p]. Finally, a complementary study compares this erroneously aligned data with a similar dataset of well-aligned data and shows that the presence of disfluencies in a four-word chunk is correlated with alignment error, but this effect is canceled when the disfluency is at least one-word distant from the *liaison* site.

MOTS-CLES : disfluences, erreurs d'alignement, alignement automatique, liaison, français.
KEYWORDS: dysfluencies, alignment errors, automatic alignment, *liaison*, French.

1 Les disfluences, défi pour la reconnaissance automatique

Les disfluences sont un groupe hétéroclite d'événements linguistiques tels que les pauses (ou pauses silencieuses), les hésitations (ou pauses remplies, par ex. *euh*, *hem*…), les faux départs, les répétitions,

10.21437/JEP.2022-48

certains marqueurs discursifs, etc. qui sont caractéristiques de l'oralité (Shriberg, 1994) et représentent jusqu'à 14% des unités prononcées dans un cadre dialogique (Adda-Decker et al., 2004). Cependant, elles présentent de remarquables régularités dans leurs proportions par rapport à la longueur de l'énoncé, leurs positions dans l'énoncé, leurs types, leurs cooccurrences (Shriberg, 1994) et même leurs caractéristiques phonétiques et prosodiques (Shriberg, 1999 ; Moniz et al., 2012 ; Christodoulides & Avanzi, 2014), et ce, y compris à travers les langues (Vasilescu et al., 2004 ; Lee et al. 2020). Longtemps considérées comme des accidents de performance sans intérêt, leurs fonctions psycholinguistiques sont en fait nombreuses : d'une part, elles signalent des problèmes de traitement lors de la construction incrémentale du message (Levelt, 1989 ; Vasilescu et al., 2010), d'autre part, elles font office de stratégies communicatives, par exemple pour gérer l'interaction dialogique (Moniz et al., 2009 ; Vasilescu et al., 2010) ou encore donner ou mettre en valeur de nouvelles informations (Arnold et al., 2003 ; Vasilescu et al., 2010).

Les disfluences présentent donc un intérêt majeur pour le traitement automatique des langues, mais restent problématiques : Adda-Decker et al. (2004) montrent qu'elles sont responsables de 20% du taux d'erreur-mot et Goldwater et al. (2009) confirment que les disfluences et les marqueurs discursifs (en particulier en début d'énoncé) représentent un défi à la reconnaissance automatique du langage, mais que le taux d'erreur-mot dépend du type et de la position des disfluences. La disfluence la plus représentée est le faux départ (37% chez Adda-Decker et al., 2004) ou son corolaire, la répétition non-finale (Goldwater et al., 2009). La pause remplie (ou hésitation) est la plus représentée des disfluences simples (Christodoulides & Avanzi, 2014) mais ne représente qu'un taux d'erreur minime (Adda-Decker et al., 2004 ; Goldwater et al., 2009).

La présente étude propose de contribuer à l'état de nos connaissances sur l'effet des disfluences sur la reconnaissance vocale en analysant les erreurs d'alignement d'un système relâché de traitement automatique de la parole sur plus de 100 heures de discours francophone. Nous nous intéressons en particulier aux erreurs d'alignement d'un type de liaison facultative en français, qui présente l'intérêt d'être un phénomène catégoriel (la liaison est réalisée ou non) mais non généralisé (la liaison n'est pas réalisée partout) qui nous permet d'explorer un terrain peu connu des erreurs d'alignement : les erreurs au niveau du phonème. Ce choix méthodologique est expliqué plus en détail en Section 2.1. Les Sections 2.2 et 2.3 décrivent les corpus et la méthodologie respectivement. La Section 3 présente une typologie des erreurs d'alignement de la consonne de liaison entre le mot précédant et le mot suivant le site de liaison, et détaille notamment quelles disfluences sont les plus impactantes et quelles consonnes elles favorisent. La Section 4 étend le scope de l'exploration à un empan de quatre mots afin de voir si la présence de disfluences peut impacter l'alignement à distance. La Section 5 est consacrée à la conclusion et à la discussion des résultats.

2 Corpus et méthodologie

Dans cette section, nous expliquons notre choix d'étudier les erreurs d'alignement de la liaison en français (Section 2.1) puis décrivons nos corpus (Section 2.2) et notre méthodologie (Section 2.3)

2.1 Cas d'étude : pourquoi la liaison facultative ?

La liaison est un phénomène de sandhi externe très fréquent en français. Les éléments impliqués dans le processus de liaison sont deux mots (Mot1 et Mot2) et une consonne de liaison. Cette consonne latente peut apparaître entre deux mots si Mot2 commence par une voyelle, comme dans *les + amis* = [le**z**ami], mais pas dans *les + copains* = *[le**z**kopɛ̃].

La cohésion syntaxique entre Mot1 et Mot2 est posée comme déterminante pour la réalisation de la liaison. Les études des catégories syntaxiques possibles des Mot1 et Mot2 ont permis une première classification (Delattre 1947) qui distingue trois catégories de liaison : « obligatoire » (ex. *un [n] ami*), « facultative » (ex. *amis [z/0] intimes*) et « interdite » (ex. *les [0] héros*). Dans ce projet, nous nous concentrons sur le phénomène de la liaison facultative. Elle se définit par le fait que le locuteur peut prononcer la séquence de deux mots soit avec liaison, soit sans liaison. Il y a donc de la variation dans la réalisation de la liaison, mais même pour un locuteur humain, production et non-production de la liaison sont également acceptables. La liaison est donc réalisée de façon catégorielle (elle est réalisée ou non) mais de façon non généralisée (pas dans tous les sites potentiels de liaison).

Étant donnée la méthode que nous avons adoptée, et que nous expliquons dans la Section 2.3 ci-dessous, le choix de la liaison facultative permet d'utiliser l'alignement automatique comme un proxy pour un locuteur humain et de mettre en lumière l'inadéquation entre traitement humain et traitement automatique, en particulier dans le traitement des disfluences. De plus, le choix d'étudier un phénomène phonotactique n'est pas anodin : notre étude est novatrice en ce que la plupart des études passées sur les erreurs de reconnaissance dues aux disfluences s'intéressent généralement à la reconnaissance du mot, pas du phonème.

Pour cette étude préliminaire, nous restreignons l'étude au contexte « verbe *être* conjugué suivi de l'article indéfini singulier *un* ou *une* ». Ce choix a été fait afin de restreindre les facteurs de confusion et de permettre de combiner l'analyse quantitative à une analyse qualitative des erreurs d'alignement qui soit humainement réalisable.

2.2 Les corpus

Outre que l'observation à grande échelle permet des résultats statistiquement significatifs (Coleman et al., 2016), l'avantage des grands corpus repose sur le fait que la parole y est plus naturelle que dans des enregistrements opérés lors d'expériences en laboratoire ou d'enquêtes sur le terrain, ce qui est essentiel dans l'observation des disfluences (Shriberg, 1994). Nous utilisons ici trois grands corpus représentatifs chacun d'un style de parole différent : ESTER (Galliano et al., 2005) comprend 80h de parole journalistique caractéristique du discours (semi-)préparé, soigné, potentiellement lu à haute voix ; ETAPE (Gravier et al., 2012) contient 13,5h de radio et 29h de télévision, notamment débats et entretiens, et est donc représentatif de monologues et dialogues semi-préparés à deux ou plusieurs interlocuteurs ; NCCFr (Torreira et al., 2010) contient 31h d'interaction *de visu* entre amis et est donc représentatif du discours informel spontané.

2.3 La méthodologie

La méthodologie se divise en deux phases : l'alignement automatique, et l'évaluation auditive.

L'alignement automatique avec variantes de prononciation

Les corpus étaient tous trois munis d'une transcription manuelle dans l'orthographe du français. Ils ont ensuite été traités suivant la méthode décrite dans Gauvain (2002) et Hallé & Adda-Decker (2011). Le système de reconnaissance automatique de la parole du LIMSI-LISN a utilisé la transcription en français pour opérer un alignement avec des transcriptions phonétiques. Pour cela, le système dispose (i) de modèles de phones pour chaque phonème du français, qui lui permettent d'identifier les phones sur un spectrogramme, et (ii) de son dictionnaire de prononciation, qui lui donne accès aux prononciations possibles de chaque mot de la langue française.

Dans le cadre de notre étude, le dictionnaire a été enrichi de façon à autoriser l'une des quatre consonnes de liaison facultative /t, n, z, p/ devant n'importe quel mot commençant par une voyelle. Le /r/ n'a pas été ajouté à l'étude car il présente beaucoup de variation à travers les variétés de français (voir Webb 2009 pour la revue de littérature) et ne représente de toute façon qu'une liaison extrêmement rare (0,04% des liaisons dans PFC selon Durand & Lyche, 2008). Ainsi, la séquence *était une* par exemple pouvait être alignée avec les transcriptions canoniques [etɛyn] ou [etɛ**t**yn], mais aussi avec les transcriptions non-canoniques [etɛ**z**yn], [etɛ**n**yn] ou [etɛ**p**yn], selon que la machine estimait qu'une consonne était réalisée ou non entre Mot1 et Mot2, et laquelle.

Nous avons ensuite trié les données afin d'assurer que la construction syntaxique était bien du type « verbe conjugué + objet » (ex. *était un homme*), et non « verbe conjugué + adverbe » (ex. *était un peu*). Au total, nous avons conservé 5049 occurrences du verbe *être* conjugué suivi de *un*/*une*.

Parmi ces 5049 occurrences, 731 présentent une inadéquation entre les consonnes de liaison alignées par le système et l'orthographe du mot (par exemple *es**t** une* aligné [ɛ**z**yn], [ɛ**n**yn] ou [ɛ**p**yn]). Parmi les 4320 restantes, l'écoute aléatoire de 964 occurrences a montré que 71 étaient mal alignées. Additionnées aux 731 déjà repérées, on obtient un total de 802 occurrences mal alignées, contre 893 bien alignées. Quelles sont les causes de ces erreurs, et surtout, quel rôle les disfluences jouent-elles ?

L'évaluation auditive

Lors de la deuxième étape, la seconde autrice a écouté 1695 occurrences : les 731 repérées préalablement par l'inadéquation avec l'orthographe et les 964 de contrôle. Pour chacun des 802 alignements erronés, elle a spécifié la cause probable de l'erreur :

- les causes techniques :
  - la bande passante est insuffisante : l'audio a parfois été enregistré sur bande téléphonique, coupée sous 4 kilohertz, ce qui est suffisant pour extraire les formants mais provoque une inadéquation avec les modèles acoustiques appris sur bande large (= 8 kHz),
  - l'enregistrement est inaudible : il arrive que le son n'ait pas été correctement capté et qu'il ne soit pas bien présent dans le fichier-son ;
- les causes liées à la situation discursive :
  - il y a du bruit de fond,
  - des disfluences (pauses, hésitations, répétitions, autocorrections, bégaiements et rires) se trouvent à l'intérieur de la séquence Mot1-Mot2,
  - deux personnes ou plus parlent en même temps,
  - le locuteur hypo-articule, et le Mot1 ou le Mot2 n'est pas clair,
  - la parole présente des caractéristiques de production particulières, en particulier une prosodie inhabituelle (chant, voix soudainement plus basse ou forte, voix craquée, accentuation…)

Notons ici que nous n'avons pas pris en compte le débit de parole, car ce dernier se confond aisément avec les réductions de prononciation (hypo-articulations).

Dans ce qui suit, nous présentons une typologie des erreurs d'alignement entre Mot1 et Mot2 et détaillons notamment le rôle des disfluences (Section 3), puis nous élargissons notre exploration à l'effet de certaines disfluences plus éloignées du site de liaison (Section 4).

3 Les disfluences dans l'environnement immédiat du site de liaison

Dans cette section, nous présentons une typologie des erreurs d'alignement et montrons le rôle joué par les disfluences (3.1). Nous affinons ensuite nos analyses en explorant quelle(s) disfluence(s) particulière(s) sont les plus liées aux erreurs d'alignement (3.2) et à quelles consonnes (3.3).

3.1 Les disfluences, deuxième cause d'erreurs d'alignement

Parmi les 7 causes d'erreurs repérées dans les données lors de la vérification auditive, on observe les proportions par corpus dans le Tableau 1.

Cause	ESTER	ETAPE	NCCFr	Taux d'erreur
aucune identifiable	65	115	17	24,56
prosodie inhabituelle	20	57	67	17,96
disfluences	20	43	47	13,72
bande passante	97	11	0	13,47
chevauchement de parole	6	95	6	13,34
bruit de fond	40	47	0	10,85
hypoarticulation	9	19	3	3,87
audio inaudible	3	11	4	2,24

TABLE 1 : Nombre d'alignements erronés par cause et par corpus et taux d'erreurs par cause tous corpus confondus.

On y constate qu'une grande partie des occurrences n'a pas de cause identifiable à l'oreille (197 occurrences, soit 24,56% des alignements erronés). Des études futures tenteront de déterminer si des paramètres acoustiques fins dans ces occurrences peuvent expliquer les erreurs d'alignement.

Outre cette catégorie inexplicable, la deuxième source d'erreur la plus répandue est la présence de disfluences dans la séquence Mot1-Mot2, qui représente 13,72% des erreurs, ce qui montre que les disfluences font partie intégrante de l'oralité et sont traitées sans problème par les humains (en l'occurrence la seconde autrice) mais restent problématiques pour les machines.

On constate néanmoins que les disfluences ont des effets différents selon le corpus, et corpus et sources d'erreurs d'alignement sont corrélés de façon significative ($\chi^2 = 397.13$, df = 14, $p < 0.0001$). Il est peu surprenant de constater que les disfluences impactent surtout le traitement de NCCFr (style informel mais amical), dont 32,64% des erreurs proviennent des disfluences, juste derrière les prosodies inhabituelles (46,53%). Les erreurs dans ETAPE n'ont, dans la majeure partie des cas, pas de source identifiée (28,89%) alors qu'il s'agit d'enregistrements professionnels. Parmi les sources identifiées, la proportion d'erreurs dues aux disfluences (10,80%) est moindre que celles liées aux prosodies inhabituelles (14,32%) et surtout aux chevauchements de parole (23,87%), ce qui est sans doute dû à la présence de débats parfois animés dans le corpus. Enfin, le taux d'erreurs liées aux disfluences dans ESTER est le plus bas des trois corpus, puisqu'elles y représentent 7,69% des sources d'erreurs. Elles semblent loin derrière les 15,38% d'erreurs dues au bruit de fond et surtout les 37,31% d'erreurs liées aux problèmes de bande passante, mais n'en demeurent pas moins la première cause d'ordre dialogique et non technique. Les disfluences sont donc une source d'erreur notable dans nos données, avec évidemment un taux plus important dans la parole dialogique informelle, mais avec un taux assez notable aussi dans la parole plus formelle, et même non nul dans la parole très préparée/lue.

3.2 Typologie des disfluences dans les erreurs d'alignement

Voyons à présent, sur les 110 erreurs d'alignement dues à des disfluences, quelles disfluences exactement provoquent les erreurs d'alignement, et si l'on trouve des taux différents selon les corpus ou le sexe des locuteurs.

Lors de la vérification auditive, la seconde autrice a repéré six types de disfluences dans nos données :

- les autocorrections, c'est-à-dire lorsque le locuteur se reprend (ex. *c'est un une*) ;
- les hésitations, ou pauses remplies, par exemple avec *euh* (ex. *c'est euh un*) ;
- les éclats de rire ;
- les pauses, et plus précisément les pauses silencieuses ;
- les répétitions, lorsque le locuteur se répète (ex. *c'est un un*) ;
- les bégaiements, quand le locuteur bafouille ou répète une partie du mot seulement.

Comme on peut le voir dans le Tableau 2, la disfluence la plus liée à des erreurs d'alignement dans nos données est la répétition, qui représente 25,00% de nos erreurs, loin devant les pauses en seconde place (12,82%). Ce résultat confirme la littérature (Adda-Decker et al., 2004 ; Goldwater et al., 2009).

	auto-correction	hésitation	rire	pause	répétition	bégaiement
ESTER	3	2	0	4	5	6
ETAPE	3	6	3	11	14	6
NCCFr	2	6	12	5	20	2
Total	8	14	15	20	39	14

TABLE 2 : Nombre d'alignements erronés par type de disfluence et par corpus.

Le type de disfluences est là encore dépendant du corpus ($\chi^2 = 22.985$, df = 10, p = 0.0108). Les répétions sont surtout sources d'erreur dans NCCFr, dont elles représentent 42,55%. NCCFr est aussi le corpus dans lequel on trouve le plus d'erreurs liées à des éclats de rire (7,69% des erreurs tous corpus confondus). Les répétitions sont aussi très représentées dans ETAPE, dont elles représentent 32,56% des sources d'erreurs, suivi de la présence de pauses, qui y représentent 25,58% des erreurs. Cependant, ces résultats sont à prendre avec prudence. En effet, nous avons sélectionné pour cette étude les séquences où Mot1 était un verbe *être* conjugué, et Mot2 l'article *un* ou *une*. Néanmoins, certaines pauses remplies, notamment par des *euh* allongés, ont été transcrites dans les corpus, ce qui fait que des séquences du type *c'est euh un*, n'ont pas été prises en compte. Il est donc probable que de nombreuses pauses remplies soient en fait absentes des données.

Par ailleurs, tous corpus confondus, les femmes représentent 30,91% des erreurs d'alignements dues à des disfluences (cf. Tableau 3), alors qu'elles représentent 27,68% des 802 alignements erronés et 36,81% des 1695 occurrences vérifiées auditivement. Il y a donc moins d'erreurs d'alignement pour la parole féminine, mais les disfluences n'y jouent pas un rôle beaucoup plus important.

	autocorrection	hésitation	rire	pause	répétition	bégaiement	Total
femme	3	3	10	4	13	1	34
homme	5	11	5	16	26	13	76

TABLE 3 : Nombre d'alignements erronés par type de disfluence et par sexe du locuteur.

Parmi les 34 occurrences erronées dans la parole féminine, la disfluence la plus représentée est la répétition (38,23%), suivi par les éclats de rire (29,41%). Chez les hommes en revanche, la disfluence la plus représentée est aussi la répétition (34,21%) mais suivie de la pause silencieuse (21,05%). Type de disfluence et genre du locuteur sont corrélés significativement ($\chi^2 = 14.658$, df = 5, p = 0.01193).

3.3 L'effet des disfluences dans les erreurs d'alignement

À présent, intéressons-nous aux effets des disfluences sur l'alignement, et en particulier à quelle consonne, parmi les quatre autorisées ([p, t, n, z]), est favorisée par quel type de disfluence.

	0	n	p	t	z	total
autocorrection	1	2	4	0	1	8
hésitation	0	6	6	1	1	14
rire	0	5	2	2	6	15
pause	2	3	12	2	1	20
répétition	0	9	6	13	11	39
bégaiement	0	5	4	4	1	14
total	3	30	34	22	21	110

TABLE 4 : Nombre d'alignements erronés par type de disfluence et par consonne.

Comme on peut le voir dans le Tableau 4, les disfluences ont peu tendance à aligner une absence de consonne. En revanche, les autocorrections (50,00%), les hésitations (42,86%) et surtout les pauses (60,00%) semble favoriser l'alignement avec [p]. La disfluence la plus représentée, néanmoins, c'est-à-dire la répétition, favorise l'alignement avec [t] (33,33%) et [z] (28,20%). Le type de disfluence et la consonne alignée sont corrélés de façon statistiquement significative ($\chi^2 = 41.293$, df= 20, p<0.005).

4 Les disfluences à distance

Dans cette partie, nous proposons une analyse complémentaire de l'effet des disfluences sur un empan de quatre mots. Nous observons ici les Mot1 et Mot2 du contexte de liaison, mais aussi le mot précédant Mot1 (Mot0) et le mot suivant Mot2 (Mot3). Les disfluences ont été établies ainsi :

- si une disfluence avait déjà été repérée lors de la vérification auditive entre Mot1 et Mot2,
- si une vérification manuelle de la transcription des données a permis de repérer une disfluence juste avant Mot1 (Mot0) ou juste après Mot2 (Mot3). Les disfluences en question peuvent être :
  - des pauses silencieuses (silences ou soupirs, qui sont présents dans la transcription),
  - des pauses remplies (ex. *ben, hem*),
  - des marqueurs discursifs (ex. *donc*, *enfin*),
  - des répétitions (ex. *c'est un un*),
  - des corrections, c'est-à-dire quand le mot est répété mais corrigé (ex. *c'est un une*),
  - des reprises, par exemple s'il y a rupture syntaxique entre Mot2 et Mot3 (ex. *c'est un il*),
  - des mots entièrement ou partiellement inintelligibles (ex. *c'est une af* pour *c'est une affaire*).

Notons d'abord que, parmi nos 1695 fichiers audios de quatre mots, 260 comprennent au moins une disfluence, soit 15,34%, ce qui est cohérent avec le taux d'Adda-Decker et al. (2004). De plus, comme

on peut le voir dans le Tableau 5, le taux de disfluences est toujours le plus élevé dans NCCFr avec 20,63% de disfluences, suivi d'ETAPE avec 14,87% et finalement ESTER avec 12,22%.

En comparant les 802 alignements erronés avec les 893 liaisons bien alignées, nous constatons que 21,82% des alignements erronés comprennent au moins une disfluence, contre 9,52% des alignements corrects. Ainsi, disfluences et alignements erronés sont corrélés ($\chi^2 = 48.297$, df = 1, $p < 0.0001$).

	avec disfluences	sans disfluences
ESTER	65	467
ETAPE	116	664
NCCFr	79	304
Total	261	1434

TABLE 5 : Nombre de fichiers audios avec et sans disfluences en fonction du corpus.

En revanche, si l'on compare les 893 liaisons bien alignées avec les 692 liaisons mal alignées mais dont l'erreur n'est pas due à une disfluence entre Mot1 et Mot2, c'est-à-dire les cas où la disfluence est à distance du site de liaison, seulement 9,39% des fichiers avec erreurs d'alignement comprennent une disfluence, et la corrélation entre disfluences distantes et alignements erronés n'est pas significative (χ^2=0, df=1, p=1). Ceci est vrai lorsqu'on considère uniquement le contexte gauche (Mot0), avec 2,60% de disfluences dans les alignements erronés et 2,58% dans les alignements corrects (χ^2=1.2515e-28, df=1, p=1), et lorsqu'on considère uniquement le contexte droit (Mot3), avec 6,79% de disfluences dans les alignements erronés et 7,28% dans les alignements corrects (χ^2=0.076381, df=1, p=0.7823). Ce résultat laisse supposer que les disfluences ne représentent pas un obstacle à l'alignement de la consonne de liaison lorsqu'elles ne se situent pas immédiatement sur le site de liaison.

5 Conclusion et discussion

Nous avons proposé ici une étude fine de l'effet des disfluences sur les erreurs d'alignement au niveau du phonème. Pour ce faire, nous avons analysé les erreurs d'alignement d'une consonne particulière, la consonne de liaison (dans le contexte « verbe *être* + *un*/*une* »), pour établir si les disfluences impactent le traitement des phonèmes isolés. Nous avons montré que les disfluences sont la deuxième source d'erreur, juste après les prosodies inhabituelles, et que, conformément à la littérature, la disfluence la plus présente dans les alignements erronés est la répétition, qui favorise l'alignement erroné des consonnes [t] et [z], suivie d'assez loin par la pause silencieuse, qui favorise l'alignement de [p]. De plus, la présence de disfluences dans l'environnement immédiat du site de liaison est bien corrélée à plus d'erreurs d'alignement, mais l'effet s'annule si la disfluence est à plus d'un mot de distance du site. Ces résultats montrent que l'alignement repose intimement sur la transcription orthographique du discours, qui n'intègre pas les disfluences et empêche le système d'identifier le matériel vocal s'il n'est pas transcrit. On peut donc se poser la question de savoir si dans le cas de la parole spontanée, avec disfluences, il ne faudrait pas intégrer les répétitions ou les pauses dans la transcription orthographique, ou bien s'il ne serait finalement pas plus efficace de se baser sur un système de reconnaissance « non supervisé », qui ne repose sur aucune transcription orthographique.

Nous espérons pouvoir affiner cette étude en explorant davantage de phonèmes, soit dans plus de contextes de liaison facultative, soit en permettant davantage de consonnes possibles. Néanmoins, notre étude préliminaire montre que les disfluences ont un effet immédiat sur la reconnaissance vocale, c'est-à-dire à l'endroit de leur réalisation, mais peu d'effet à distance. Ce résultat peut ouvrir

une porte intéressante pour tenter de limiter les erreurs au niveau du mot, surtout quand ces erreurs touchent les frontières de mots, c'est-à-dire ses premiers et derniers phones.

Remerciements

Cette recherche a été partiellement financée par le prix Excellence de L'Institut DATAIA et la MSH Paris-Saclay, par le projet SON-DISCOURS dans le cadre de l'appel Émergence de l'IdEX U. de Paris (ANR-18-IDEX-0001) et par le Labex EFL (ANR-10-LABX-0083).

Références

ADDA-DECKER M., HABERT B., BARRAS C., BOULA DE MAREÜIL Ph., PAROUBEK P. (2004). Une étude des disfluences pour la transcription automatique de la parole et l'amélioration des modèles de langage. *Actes des 25èmes JEP*.

ARNOLD J.E., FAGNANO M., TANENHAUS M.K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1), 25-36. Doi: 10.1023/a:1021980931292

CHRISTODOULIDES G., AVANZI M. (2014). Phonetic and Prosodic Characteristics of Disfluencies in French Spontaneous Speech. *LabPhon 2014*.

COLEMAN J., RENWICK M., TEMPLE R. (2016). Probabilistic underspecification in nasal place assimilation. *Phonology*, 33(3), 425-458.

DELATTRE P. (1947). La liaison en français, tendances et classification. *The French Review* 21 (2), 148-157

DURAND J., LYCHE C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, 18 (1), 33-66. https://doi.org/10.1017/S0959269507003158

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F., GRAVIER J. (2005). ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast Newshase II Evaluation Campaign for the Rich Transcription of French Broadcast News. *Proceedings of Interspeech 2005*, 2453–2456

GAUVAIN J.-L., LAMEL L., ADDA G. (2002). The LIMSI broadcast news transcription system. *Speech communication* 37 (1-2), 89–108

GOLDWATER S., JURAFSKY D., MANNING C.D. (2009). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52 (2010) 181–200

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A., GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. *Proceedings of LREC Eighth international conference on Language Resources and Evaluation*

HALLÉ P., ADDA-DECKER M. (2011). Voice assimilation in French obstruents: A gradient or a categorical process? *Tones and features: A festschrift for Nick Clements*, De Gruyter, 149–175

LEE L., JOUVET D., BARTKOVA K., KEROMNES Y., DARGNAT M. (2020). Correlation Between Prosody and Pragmatics: Case Study of Discourse Markers in French and English. *Proceedings of Interspeech 2020*, 1878-1882, doi: 10.21437/Interspeech.2020-2204

LEVELT W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press

MONIZ H., BATISTA F., TRANCOSO I., MATA DA SILVA A.I. (2012). Prosodic context-based analysis of disfluencies. *Proceedings of Interspeech*, Portland, Oregon

MONIZ H., TRANCOSO I., MATA DA SILVA A.I. (2009). Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. *Proceedings of Interspeech 2009*, Brighton, UK, 1719-1722

SHRIBERG E. (1994). *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, University of California, Berkeley

SHRIBERG E. (1999). Phonetic Consequences of Speech Disfluency. *Proceedings of the 14th ICPhS*. San Francisco, 619-622

SHRIBERG E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31(1), 153-169

TORREIRA F., ADDA-DECKER M., ERNESTUS M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, Elsevier: North-Holland, 2010, 52 (3)

VASILESCU I., CANDEA M., ADDA-DECKER M. (2004). Hésitations autonomes dans 8 langues : une étude acoustique et perceptive. *Actes de MIDL* 2004, 25-30

VASILESCU I., ROSSET S., ADDA-DECKER M. (2010). On the functions of the vocalic hesitation *euh* in interactive man-machine question answering dialogs in French. *Proceedings of DiSS-LPSS-2010*, 111-114

WEBB E. (2009). Minimalism and French /R/: Phonological representations in phonetically based phonology. *Journal of French Language Studies*, 19(1), 87-115. doi:10.1017/S095926950800358X