"Train&Align: Un outil d'alignement phonétique automatique disponible en ligne"

Brognaux, Sandrine ; Roekhaut, Sophie ; Drugman, Thomas ; Beaufort, Richard

Abstract

Plusieurs outils d'alignement phonétique automatique de corpus oraux sont actuellement disponibles. Ils exploitent généralement des modèles indépendants du locuteur pour aligner de nouveaux corpus. Leur désavantage est qu'ils couvrent un nombre très limité de langues et fournissent parfois un alignement de qualité réduite quand ils sont appliqués sur différents styles de parole. Cet article présente Train&Align, un nouvel outil d'alignement phonétique automatique disponible en ligne (http ://cental.fltr.ucl.ac.be/train_and_align). Sa spécificité est qu'il entraîne les modèles directement sur le corpus à aligner, ce qui le rend applicable à toutes les langues et à tous les styles de parole. Des tests effectués sur six corpus dans plusieurs langues et styles de parole montrent qu'il produit un alignement de qualité comparable aux autres outils d'alignement. Train&Align permet également d'optimiser certains paramètres d'entraînement. Ainsi, une ...

Document type : *Communication à un colloque (Conference Paper)*

Train&Align : un outil d'alignement phonétique automatique disponible en ligne

Sandrine Brognaux[1, 2] Sophie Roekhaut[1] Thomas Drugman[2] Richard Beaufort[3]
(1) Cental - Université catholique de Louvain (UCL), Belgium
(2) TCTS - UMons, Belgium (3) Nuance Communications, Inc., Belgium *
sandrine.brognaux@uclouvain.be, sophie.roekhaut@uclouvain.be,
thomas.drugman@umons.ac.be, richard.beaufort@nuance.com

RÉSUMÉ
Plusieurs outils d'alignement phonétique automatique de corpus oraux sont actuellement disponibles. Ils exploitent généralement des modèles indépendants du locuteur pour aligner de nouveaux corpus. Leur désavantage est qu'ils couvrent un nombre très limité de langues et fournissent parfois un alignement de qualité réduite quand ils sont appliqués sur différents styles de parole. Cet article présente Train&Align, un nouvel outil d'alignement phonétique automatique disponible en ligne (http ://cental.fltr.ucl.ac.be/train_and_align). Sa spécificité est qu'il entraîne les modèles directement sur le corpus à aligner, ce qui le rend applicable à toutes les langues et à tous les styles de parole. Des tests effectués sur six corpus dans plusieurs langues et styles de parole montrent qu'il produit un alignement de qualité comparable aux autres outils d'alignement. Train&Align permet également d'optimiser certains paramètres d'entraînement. Ainsi, une partie manuellement alignée du corpus peut notamment être utilisée afin d'améliorer la qualité des modèles. Les tests montrent une amélioration du taux d'alignement dépassant les 15%, quand 30 secondes de corpus aligné manuellement sont utilisées.

ABSTRACT
Train&Align : An automatic phonetic alignment tool available online

Several automatic phonetic alignment tools have been proposed in the literature. They usually rely on pre-trained speaker-independent models to align new corpora. Their drawback is that they cover a very limited number of languages and might not perform properly for different speaking styles. This paper presents Train&Align, a new tool for automatic phonetic alignment available online. Its specificity is that it trains the models directly on the corpus to align, which makes it applicable to any language and speaking style. Experiments on six corpora in different languages and speaking styles show that it provides results comparable to other existing tools. Train&Align also allows the tuning of some training parameters. A manually-aligned part of the corpus can, for instance, be used as bootstrap to improve the model quality. Alignment rates were found to significantly increase, by more than 15%, using only 30 seconds of bootstrapping data.

MOTS-CLÉS : Phonétique, alignement, HMM, corpus oraux, annotation.

KEYWORDS: Phonetics, alignment, HMM, speech corpora, annotation.

*. Cette étude a été réalisée alors que Richard Beaufort travaillait au CENTAL (UCL, Belgium).

1 Introduction

Les corpus oraux de taille importante jouent un rôle crucial tant en recherche linguistique que dans les technologies de la parole. Une particularité de ces corpus est que le son est rarement étudié seul. Ses transcriptions orthographique et phonétique sont généralement nécessaires. Les phonèmes, en particulier, doivent être alignés avec le son. Des outils d'annotation tels que Praat (Boersma et Weenink, 2009), EXMARaLDA (Schmidt, 2012) ou ELAN (Wittenburg *et al.*, 2006) permettent de définir plusieurs couches d'annotation. Ils offrent également la possibilité d'aligner manuellement ces différentes annotations avec le son. Cependant, l'alignement manuel présente deux problèmes majeurs. D'une part, un tel traitement requiert un temps considérable : de 130 à 800 fois la durée du son (Kawai et Toda, 2004; Schiel et Draxler, 2003). D'autre part, ce travail nécessite des phonéticiens entraînés et une bonne concertation entre les différents annotateurs.

Pour tenter de résoudre ce problème, des outils d'alignement automatique ont vu le jour. Ils permettent de fournir un alignement à la fois cohérent et reproductible. Les méthodes fondées sur l'utilisation de modèles de Markov cachés (HMM) sont reconnues comme fournissant les meilleurs résultats d'alignement phonétique (Adell *et al.*, 2005; van Niekerk et Barnard, 2009). Comme en reconnaissance vocale, des modèles acoustiques de chaque (groupe de) phonème(s) sont appris sur un corpus de taille plus ou moins importante. Ces modèles servent ensuite à aligner d'autres corpus avec leur transcription phonétique.

Le toolkit HTK (Young *et al.*, 1995) offre les outils nécessaires à l'alignement phonétique automatique. Utilisable en ligne de commande, il nécessite cependant un certain niveau de programmation de la part de l'utilisateur. Plusieurs outils, avec ou sans interface graphique, tels que EasyAlign (Goldman, 2011), SPPAS (Bigi, 2012) ou P2FA (Yuan et Liberman, 2008) sont donc venus se greffer sur HTK ou sur des toolkits similaires. Ils fournissent des modèles indépendants du locuteur entraînés sur de grandes bases de données multi-locuteurs ainsi que les méthodes pour aligner d'autres corpus à l'aide de ces modèles. Ces outils présentent trois limites majeures. Premièrement, les modèles ne sont proposés que dans un nombre très limité de langues (p. ex. quatre langues pour EasyAlign). Il est alors impossible d'aligner un corpus dans une langue non couverte. Une seconde limitation concerne le fait que la phase d'entraînement n'est pas accessible. Ces outils proposant des modèles pré-entraînés, il n'est pas possible d'optimiser les paramètres d'entraînement. Enfin, les modèles fournis sont censés offrir une représentation globale de la langue, non spécifique à un locuteur ou un style de parole particulier, on parle ainsi de modèles génériques. Cependant, les modèles dépendent fortement du corpus d'entraînement, généralement un corpus de parole neutre et lue. Leur utilisation pour l'alignement d'autres styles de parole (spontanée, expressive, etc.) produit donc souvent un alignement de qualité réduite. Cette dernière limitation concerne également les phonèmes utilisés lors de l'annotation : seuls les phonèmes présents dans le modèle peuvent être alignés. Ainsi, EasyAlign ne propose pas de modèle pour le phonème /ŋ/, prononcé notamment en fin du mot "camping", tandis que SPPAS n'effectue pas la distinction entre les phonèmes /ɛ̃/ et /œ̃/ (de "brin" et "brun"). Enfin, si certains phonèmes étaient rares ou sous-représentés dans le corpus d'entraînement, ils seront plus enclins à être mal alignés, comme le soulignent Goldman et Schwab (2011).

Train&Align [1] est un nouvel outil d'alignement phonétique automatique qui offre une solution pour pallier ces problèmes. Il se distingue des outils existants par le fait qu'il entraîne directement les modèles acoustiques sur le corpus à aligner. Il ne nécessite donc pas de modèle préalablement

1. /http ://cental.fltr.ucl.ac.be/train_and_align

entraîné sur un corpus multi-locuteurs et peut ainsi être utilisé pour aligner n'importe quelle langue ou n'importe quel style de parole. De plus, Train&Align implémente différentes options d'entraînement qui permettent d'améliorer la qualité de l'alignement. Contrairement à EasyAlign qui ne propose qu'une version sous Windows, notre outil est directement accessible en ligne et peut ainsi être utilisé sur tous les systèmes d'exploitation. Enfin, le système offre la possibilité d'évaluer automatiquement la qualité de l'alignment produit si une partie manuellement alignée du corpus est fournie. Il faut cependant noter que notre outil présuppose actuellement l'utilisation d'une transcription phonétique correcte et ne permet donc pas l'introduction de variantes phonétiques lors de l'alignement comme proposé par d'autres études et outils (Goldman, 2011; Boula de Mareuil *et al.*, 2008; Bigi, 2012).

L'objectif de cet article est de présenter ce nouvel outil d'alignement ainsi que les techniques sur lesquelles il repose. Une description de l'outil et de son fonctionnement est donnée en section 2. La section 3 présente le protocole expérimental établi pour évaluer les performances de Train&Align, suivi de la comparaison des taux d'alignement obtenus avec ceux atteints par les modèles fournis par d'autres outils existants. Enfin, la section 5 conclut l'article.

2 Présentation de Train&Align

Train&Align est un nouvel outil d'alignement phonétique automatique. Sa particularité est qu'il entraîne directement les modèles acoustiques sur le corpus à aligner, ce qui le rend applicable à toutes les langues et styles de parole. L'intérêt de cette technique a déjà été mise en évidence par de nombreuses recherches en reconnaissance vocale (Leggetter et Woodland, 1995) qui montrent que les modèles dépendant du locuteur permettent d'atteindre de meilleurs résultats.

Dans un premier temps, l'entièreté du corpus (non aligné) et sa transcription phonétique sont utilisés afin d'entraîner les modèles acoustiques de chaque phonème (monophones à 5 états) [1]. Notons que le modèle de silence repose sur une configuration particulière qui permet une modélisation plus adéquate de sa durée. Des modèles de courtes pauses ('sp') sont également automatiquement insérés entre les mots. Ils correspondent à des modèles de silence spécifiques et permettent la détection de silences non annotés dans la transcription phonétique. Ces modèles sont également implémentés dans EasyAlign et P2FA. SPPAS, au contraire, ne propose pas de modèle pour les silences qui sont traités dans une phase indépendante de l'alignement.

Les modèles acoustiques sont initialisés à l'aide d'un alignement uniforme du son et de sa transcription phonétique. Chaque phonème est ainsi modélisé par des paramètres moyens, communs à tous les phonèmes. Les paramètres sont ensuite adaptés itérativement par l'algorithme de Baum-Welsh (Young *et al.*, 1995). Dans un second temps, ces modèles sont utilisés pour aligner le corpus d'entraînement lui-même. HTK est utilisé pour l'entraînement et l'alignement.

Deux options principales sont proposées par Train&Align et permettent de modifier certains paramètres de l'entraînement des modèles.

Tout d'abord, la configuration des modèles peut être modifiée. Au lieu d'entraîner un modèle par phonème (monophone), il est possible d'entraîner un modèle par triphone, c'est-à-dire par séquence de trois phonèmes. L'utilisation de tels modèles nécessite cependant un corpus de

1. Les paramètres des modèles sont 12 Mel Frequency Cepstral Coefficients (MFCC) et un coefficient d'énergie, ainsi que leur première et seconde dérivée.

taille importante afin de modéliser correctement chaque triphone. De plus, l'augmentation du nombre de modèles s'accompagne d'une augmentation du temps de traitement. Afin de pallier ce problème, il est également possible d'utiliser des tied-state triphones. Dans ce cas, le contexte gauche et droit des triphones est remplacé par des classes phonétiques (fricatives, voyelles, plosives, etc.) ce qui réduit considérablement le nombre de modèles. Notons que la liste des classes phonétiques est ouverte et peut être définie librement pour chaque nouvelle langue ou nouveau style de parole à aligner. L'utilisation de triphones ou de tied-triphones est une option que peuvent plus difficilement offrir les outils reposant sur des modèles pré-entraînés. Idéalement, l'outil devrait alors proposer des modèles pour tous les triphones possibles de la langue, ce qui demande l'utilisation d'un corpus d'entraînement de taille très importante. Dans le cas contraire, l'alignement échoue lorsqu'il rencontre des triphones inconnus. Naturellement, ce problème ne survient pas lorsque l'on entraîne les modèles sur le corpus à aligner.

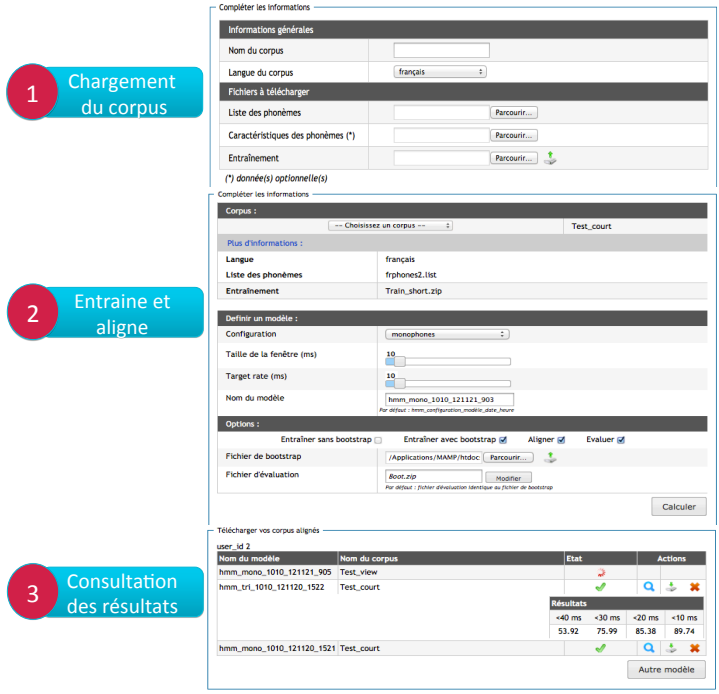

FIGURE 1 – Aperçu de Train&Align

Une seconde option est le bootstrap. Si une petite partie du corpus a été alignée manuellement, elle peut être exploitée afin d'améliorer l'initialisation des modèles et l'alignement qui en résulte. Dans ce cas, l'initialisation uniforme est remplacée par une procédure itérative qui détermine les paramètres d'après l'alignement manuel.

Train&Align permet également d'évaluer automatiquement la qualité de l'alignement. Si une partie du corpus alignée manuellement est disponible, elle peut être comparée à l'alignement automatique. On calcule ainsi le pourcentage de frontières de phonèmes qui sont similaires dans les deux alignements, avec un certain seuil de tolérance. Nous considérons donc la proportion de

frontières pour lesquelles l'erreur temporelle est inférieure à un seuil qui varie de 10 à 40 ms, par pas de 10 ms. Le bootstrap est associé à une évaluation particulière. Si le même sous-corpus est fourni pour l'évaluation et le bootstrap, une 'évaluation croisée à 5 plis' est réalisée : 4/5 du sous-corpus sont, tour à tour, exploités pour le bootstrap, le cinquième restant servant à l'évaluation. Les taux d'alignement correct obtenus à chaque itération sont ensuite moyennés. Il est également possible d'utiliser deux sous-corpus différents pour l'évaluation et le bootstrap.

Notons enfin que seul le son et sa transcription phonétique sont nécessaires pour entraîner un nouveau modèle et ainsi produire l'alignement du corpus. Cette transcription peut être fournie sous la forme d'un simple fichier .txt. Train&Align est également compatible avec le format Praat, acceptant des TextGrids en entrée et produisant des TextGrids contenant une couche phonétique alignée. Il permet aussi d'exclure de l'alignement des parties du TextGrid qui ne doivent pas être alignées ou n'ont pas été transcrites. L'interface de l'outil peut être observée en Figure 1.

3 Évaluation de Train&Align

3.1 Protocole expérimental

Nous proposons ici une évaluation de Train&Align sur différents corpus, de langues et styles de parole différents. Dans le cadre de cette étude, six corpus, manuellement alignés, sont utilisés :
– Pour le français, un corpus neutre lu de 110 minutes utilisé en synthèse vocale (*Marie* (Colotte et Beaufort, 2005)) et un corpus expressif spontané de commentaires sportifs de 15 minutes (*Sportic* (Brognaux *et al.*, 2013)) ;
– Pour l'anglais, un corpus neutre lu de 12 minutes utilisé en synthèse vocale (*Will* [3]) et un corpus expressif lu de 51 minutes composé de 5 émotions et 5 locuteurs (*Woggle* (Dellaert *et al.*, 1996)) ;
– Deux corpus neutres de langues peu dotées : le *féroïen* [3] (21 min) et le *gaélique* [4] (8 min).
Les corpus expressifs sont caractérisés par un degré important de variabilité.

En Section 3.2, l'alignement par Train&Align est comparé à celui obtenu en utilisant les modèles fournis par des outils d'alignement existants : P2FA (Yuan et Liberman, 2008) pour l'anglais, SPPAS (Bigi, 2012) pour le français et l'anglais et EasyAlign (Goldman, 2011) pour le français. Les outils existants ne fournissant pas de modèles pour le féroïen et le gaélique, ces corpus sont donc uniquement alignés par Train&Align. Les résultats obtenus en exploitant le contexte phonétique sont ensuite analysés en Section 3.3. Enfin, la Section 3.4 étudie l'amélioration obtenue lors de l'utilisation d'une partie manuellement alignée du corpus comme bootstrap.

Pour tous les corpus, la transcription phonétique correcte est fournie pour l'alignement.

3.2 Version de base

L'alignement des corpus par Train&Align (T&A) est ici comparé à celui obtenu par les modèles fournis par les outils existants. Les taux d'alignement sont calculés sur un corpus de chaque

3. gracieusement fourni par Acapela Group SA
4. gracieusement fourni par le Phonetics & Speech Laboratory, Trinity College, Dublin

langue (neutre pour le français et expressif pour l'anglais). Deux précisions méthodologiques doivent être faites. Tout d'abord, SPPAS fournit des modèles triphones pour l'anglais. Cela signifie que l'alignement à l'aide de ce modèle échoue lorsqu'il présente des triphones inexistants. Alors que le programme SPPAS a été développé afin de permettre de résoudre ce problème, nous avons décidé ici d'exclure toutes les phrases contenant des triphones inconnus et ce afin de permettre une comparaison des modèles entre eux. Seule une petite moitié des fichiers en anglais a donc pu être évaluée par les modèles de SPPAS. Il faut ensuite noter que P2FA fournit des modèles différents selon l'accent que reçoit la voyelle (non accentuée, accents primaire et secondaire). Afin de mesurer la pleine potentialité de ce modèle, notre corpus a donc également été annoté accentuellement pour l'alignement avec ce modèle seulement. Notons cependant que les résultats obtenus sans distinguer les différents niveaux d'accentuation avec le modèle de P2FA sont nettement inférieurs (diminution de plus de 6 % à 20 ms près).

TABLE 1 – Taux d'alignement (%) avec Train&Align et les modèles fournis par d'autres outils

	Modèle	Correct <10 ms	Correct <20 ms	Correct <30 ms	Correct <40 ms
Français (*Marie*)	SPPAS	43.04	68.91	81.5	88.63
	EasyAlign	52.54	77.54	87.72	92.11
	T&A	60.67	84.56	92.88	96.69
Anglais (*Woggle*)	T&A	43.44	63.9	78.34	87.18
	SPPAS	37.3	65.2	82.07	88.69
	P2FA	46.68	69.74	81.2	87.46
Féroïen	T&A	47.1	74.09	87.24	93.61
Gaélique	T&A	75	88.78	93.59	95.98

Les résultats sont présentés en Table 1. Pour le corpus en français, nous observons que Train&Align dépasse largement les performances des modèles des autres aligneurs. Le taux d'alignement obtenu est d'ailleurs relativement élevé et comparable aux taux d'accord interannotateurs rapportés par Goldman (2011) qui étaient d'environ 80 % à 20 ms près. En ce qui concerne le corpus en anglais, les résultats de Train&Align sont sensiblement comparables à ceux obtenus par le modèle de SPPAS avec une précision supérieure de 6 % à 10 ms près mais des taux légèrement inférieurs pour les seuils plus élevés. Seul P2FA permet d'atteindre de meilleurs résultats. Ceci s'explique notamment par son utilisation d'un modèle différent pour chaque type d'accent, ce qui permet une meilleure modélisation de la langue. De plus, les modèles utilisés par P2FA sont entraînés sur plus de 25 heures de parole alignées manuellement au mot. Afin de fournir de tels résultats pour chaque langue, un travail manuel important devrait donc être effectué, ce qui est économiquement peu réalisable. La qualité globalement plus faible de l'alignement, comparé au corpus français, est probablement dûe à la plus grande variabilité de ce corpus qui contient plusieurs émotions et locuteurs. Il nous faut tout de même noter que Train&Align offre des résultats relativement comparables à P2FA, surtout pour des seuils de tolérance élevés, bien qu'il ne nécessite pas de modèle pré-entraîné, ni d'annotation de l'accentuation. Enfin, les taux d'alignement correct obtenus pour le féroïen et le gaélique sont assez proches de ceux obtenus sur le corpus français, ce qui indique une bonne adaptation de notre outil à des langues peu dotées. Notons enfin que l'application de Train&Align sur un corpus en néerlandais lors d'une précédente étude nous a également permis d'obtenir de bons taux d'alignement (Michaux *et al.*, 2014).

Train&Align entraînant les modèles acoustiques directement sur le corpus à aligner, la taille de ce corpus joue un rôle évident dans la qualité des modèles. Les taux d'alignement pour le corpus en français sont ici assez élevés mais les tests sont effectués sur un corpus de plus de 100 minutes de parole. On peut donc s'interroger sur la qualité de l'alignement lorsque la taille du corpus est plus réduite. L'influence de la taille du corpus sur la qualité de l'alignement a été étudiée. Nous avons ainsi pu observer que les taux d'alignement restent relativement stables jusqu'à l'utilisation d'un corpus de 5 minutes. Pour des corpus neutres avec peu de variabilité (p. ex. *Marie*), un taux correct d'alignement peut être conservé jusqu'à 2 minutes. En deçà de ce seuil, la qualité de l'alignement décroit rapidement.

3.3 Considération du contexte phonétique

Comme mentionné précédemment, Train&Align offre différentes options d'entraînement. Une première modification possible concerne la configuration des modèles. Au lieu d'un modèle par phonème (mono), un contexte phonétique plus large peut être considéré. Les résultats obtenus lors de l'utilisation de triphones (tri) et de tied-state triphones (tied) sont présentés en Table 3. Ces options ne sont ici testées que sur les corpus anglais et français, langues pour lesquelles nous disposons d'une table des caractéristiques acoustiques de chaque phonème utilisé. On observe que l'utilisation de tels modèles permet d'atteindre, globalement, un meilleur alignement des corpus. Ainsi, pour *Sportic*, l'utilisation de tied-state triphones est recommandée et permet d'obtenir une augmentation de près de 3 % à 20 ms près. L'utilisation des tied-state triphones sur le corpus expressif anglais montre cependant une amélioration pour les seuils de 30 ms ou plus seulement. Il est donc important de sélectionner ou non cette option en fonction du corpus à aligner. Si une petite partie manuellement alignée du corpus est disponible, une évaluation avec les différentes options d'entraînement permettra de choisir la meilleure configuration.

TABLE 2 – Taux d'alignement (%) avec Train&Align-mono, -tri et -tied

	Français				Anglais			
	<10 ms	<20 ms	<30 ms	<40 ms	<10 ms	<20 ms	<30 ms	<40 ms
NEUTRE			*Marie*				*Will*	
mono	60.67	**84.56**	92.8	96.69	49.86	**76.85**	87.14	92.68
tri	62.7	85.44	92.56	96.53	49.05	75.95	86.81	91.99
tied	63.64	**86.05**	92.91	96.63	50.79	**78.27**	87.68	93.06
EXPRESSIF			*Sportic*				*Woggle*	
mono	51.75	**72.6**	83.63	89.23	43.44	**63.9**	78.34	**87.18**
tri	51.66	72.71	84.15	89.89	42.85	63.38	78.78	88.17
tied	54.88	**75.57**	85.97	91.26	42.43	**62.63**	78.37	**88.24**

3.4 Bootstrap

Une seconde option proposée par Train&Align est le bootstrap. Une petite partie du corpus alignée manuellement peut être exploitée afin de permettre une meilleure initialisation des modèles et ainsi une meilleure qualité de l'alignement. Nos tests montrent qu'un sous-corpus manuellement aligné de 30 secondes (hors silences initiaux et finaux) permet d'atteindre une nette amélioration du taux d'alignement (voir Table 3). L'alignement obtenu est ici évalué sur le même sous-corpus de 2 minutes avec et sans bootstrap afin d'éviter tout biais introduit par le

sous-corpus d'évaluation. Au delà de 30 secondes de bootstrap, la qualité de l'alignement reste majoritairement stable, certains corpus témoignant encore d'une augmentation avec un bootstrap de 1 minute à 2 minutes. Comme mentionné précédemment, il a été montré que l'alignement manuel nécessite environ 130 fois la durée du son (Kawai et Toda, 2004). Pour 30 secondes, une heure de travail manuel serait donc nécessaire, ce qui est économiquement faisable. Pour *Woggle*, le corpus en anglais expressif qui obtenait des taux d'alignement assez bas avec la version de base, on observe ainsi une augmentation de près de 20 %, ce qui lui permet d'atteindre une qualité d'alignement proche de celle du corpus neutre. Ceci semble montrer que l'utilisation de bootstrap est particulièrement indiquée pour l'alignement de corpus caractérisés par un haut degré de variabilité.

TABLE 3 – Taux d'alignement des corpus sans et avec 30 secondes de bootstrap

		<10 ms	<20 ms	<30 ms	<40 ms
Marie	mono	58.84	**81.69**	91.43	95.13
	boot (30 sec.)	62.65	**84.76**	93.97	96.08
Woggle	mono	44.58	**65.34**	78.33	87.18
	boot (30 sec.)	64.27	**85.11**	92.31	95.53
Féroïen	mono	45.95	**74.83**	87.86	94.51
	boot (30 sec.)	52.66	**79.10**	90.52	95.68
Gaélique	mono	75.31	**89.10**	94.29	96.39
	boot (30 sec.)	78.33	**91.60**	95.67	97.11

4 Conclusion

Cet article a présenté un nouvel outil d'alignement phonétique automatique disponible en ligne : Train&Align . Contrairement aux outils existants, il entraîne les modèles acoustiques directement sur le corpus à aligner, ce qui lui permet d'aligner n'importe quelle langue ou n'importe quel style de parole. Les tests montrent que Train&Align fournit des résultats comparables aux outils existants et permet également d'obtenir de bons taux d'alignement pour des langues peu dotées. De plus, il offre la possibilité de modifier certains paramètres d'alignement. Ainsi, il a été montré que l'utilisation de 30 secondes de bootstrap, permet d'atteindre une amélioration allant jusqu'à 20 % pour un corpus caractérisé par un haut degré de variabilité. Enfin, l'outil fournit une évaluation automatique de l'alignement si une partie manuellement alignée est disponible.

Remerciements

S. Brognaux et T. Drugman sont soutenus par le « FNRS ». S. Roekhaut est financée par la convention ARC 12/17-044. Les auteurs tiennent également à remercier J.-P. Goldman et B. Bigi pour leur aide lors de l'utilisation de leur outil et pour leur enthousiasme concernant cette étude.

Références

ADELL, J., BONAFONTE, A., GOMEZ, J. A. et CASTRO, M. J. (2005). Comparative study of automatic phone segmentation methods for TTS. *In ICASSP*.

BIGI, B. (2012). SPPAS : a tool for the phonetic segmentation of speech. *In LREC*.

BOERSMA, P. et WEENINK, D. (2009). Praat : doing phonetics by computer (version 5.1.05) [computer program].

Boula de MAREUIL, P., VIERU-DIMULESCU, B., WOEHRLING, C. et ADDA-DECKER, M. (2008). Accents trangers et rgionaux en franç. *Traitement Automatique des Langues*, 49 (3):135–163.

BROGNAUX, S., PICART, B. et DRUGMAN, T. (2013). A new prosody annotation protocol for live sports commentaries. *In Interspeech*.

COLOTTE, V. et BEAUFORT, R. (2005). Linguistic features weighting for a text-to-speech system without prosody model. *In Interspeech*.

DELLAERT, F., POLZIN, T. et WAIBEL, A. (1996). Recognizing emotion in speech. *In ICSLP*.

GOLDMAN, J.-P. (2011). EasyAlign : an automatic phonetic alignment tool under Praat. *In Interspeech*.

GOLDMAN, J.-P. et SCHWAB, S. (2011). Easyalign spanish : An (semi-) automatic segmentation tool under Praat. *In 5th CFE*.

KAWAI, H. et TODA, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. *In ICASSP*.

LEGGETTER, C. et WOODLAND, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185.

MICHAUX, M.-C., BROGNAUX, S. et CHRISTODOULIDES, G. (2014). The production and perception of L1 and L2 Dutch stress. *In Speech Prosody*.

SCHIEL, F. et DRAXLER, C. (2003). The production of speech corpora. Rapport technique, Bavarian Archive for Speech Signals.

SCHMIDT, T. (2012). Exmaralda and the folk tools. *In LREC*.

van NIEKERK, D. et BARNARD, E. (2009). Phonetic alignment for speech synthesis in under-resourced languages. *In Interspeech*.

WITTENBURG, P., BRUGMAN, H., RUSSEL, A., KLASSMANN, A. et SLOETJES, H. (2006). Elan : a professional framework for multimodality research. *In LREC*.

YOUNG, S., KERSHAW, D., ODELL, J., OLLASON, D., VALTCHEV, V. et WOODLAND, P. (1995). *The HTK Book (for HTK Version 3)*. Cambridge University.

YUAN, J. et LIBERMAN, M. (2008). Speaker identification on the SCOTUS corpus. *In Acoustics '08*.