De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée

Dominique Fohr, Odile Mella, Denis Jouvet

DE L'IMPORTANCE DE L'HOMOGÉNÉISATION DES CONVENTIONS DE TRANSCRIPTION POUR L'ALIGNEMENT AUTOMATIQUE DE CORPUS ORAUX DE PAROLE SPONTANÉE

Dominique Fohr, Odile Mella, Denis Jouvet
Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France

1 Introduction

Afin de pouvoir offrir à la communauté scientifique un Corpus d'Étude pour le Français Contemporain écrit et oral (CEFC), le projet ANR ORFEO (Outils et Recherches sur le Français Ecrit et Oral) [6] a décidé de rassembler sur une plate-forme plusieurs corpus oraux existants en associant à chacun un ensemble de couches d'annotation. La couche d'annotation la plus proche du signal audio est le résultat de l'alignement automatique en phonèmes et en mots d'un fichier audio à partir de la transcription orthographique associée à ce fichier audio. Cette annotation temporelle en phonèmes et en mots des corpus pourra être utilisée dans des études linguistiques pour extraire des segments de parole, pour les écouter ou pour effectuer des traitements acoustiques comme par exemple une analyse prosodique.
Les corpus rassemblés dans le projet ont été orthographiquement transcrits par différents laboratoires en utilisant des conventions propres à chaque laboratoire et donc hétérogènes. Plus généralement, dans le cadre des recherches sur l'analyse, la synthèse et la reconnaissance automatique de la parole, il a été indispensable de développer des outils d'alignement automatique texte-parole [2][4]. Au LORIA, nous avons développé le logiciel ASTALI (Automatic Speech-Text ALIgnment) pour réaliser automatiquement l'alignement en phonèmes et en mots de corpus oraux. L'objet de cet article est de présenter les difficultés rencontrées lors de l'adaptation de notre outil pour l'alignement des différents corpus ORFEO du fait de l'hétérogénéité des conventions de transcription. Cette présentation est précédée d'une brève description du logiciel ASTALI.

2 Fonctionnement d'ASTALI

Comme le montre la figure 1, ASTALI est un logiciel qui, à partir d'un signal audio en langue française et de sa transcription orthographique, fournit un alignement au niveau du mot et du phonème. La transcription peut être fournie soit sous la forme d'un fichier texte simple (uniquement la suite de mots prononcés) pour des fichiers audio courts, soit sous la forme d'un fichier Transcriber (trs) contenant des tours de parole [1]. Après une étape de prétraitement de la transcription, le logiciel génère les différentes prononciations possibles pour chacun des mots du texte. Il construit ensuite le graphe phonétique des prononciations possibles du texte en tenant compte des liaisons potentielles et en ajoutant un silence optionnel entre chaque mot [7]. Puis, il détermine la prononciation la plus proche de celle réalisée par le locuteur en recherchant le meilleur chemin dans ce graphe à l'aide de modèles acoustiques statistiques (modèles de phonèmes et de bruits) fondés sur les modèles de Markov cachés (HMM, Hidden Markov Models). L'alignement obtenu peut être fourni au format HTK (mlf) [8] ou au format Praat (textgrid) [3]. La figure 2 montre un fichier résultat affiché par le logiciel Praat.
Dans le cas d'une transcription Transcriber, le logiciel ASTALI possède les fonctionnalités suivantes : indication des locuteurs dans le fichier résultat, restitution des informations signalées par les balises Transcriber « Event » (événements) et « Comment » (commentaires) dans le fichier résultat, traitement de la parole superposée, et traitement des fichiers anonymisés.
De plus, l'utilisateur peut fournir un dictionnaire phonétique contenant les mots dont il souhaite donner lui-même la phonétisation, par exemple des noms propres.
Dans le cadre de l'Equipex ORTOLANG [5], une version Web du logiciel ASTALI est en cours de développement (astali.loria.fr).

3 Les difficultés liées à l'hétérogénéité des conventions de transcription

3.1 Divergences de notation orthographique

La première source de divergence entre les transcriptions orthographiques concerne la présence de ponctuation, l'utilisation de majuscules et la manière d'écrire les sigles et les noms propres. Afin de respecter les notations choisies par l'annotateur, ASTALI conserve la casse et la ponctuation du texte original. Plus précisément, pour aligner, il supprime la ponctuation mais la réintègre dans le fichier résultat final. Les phonétisations d'un mot sont d'abord recherchées, éventuellement en modifiant la casse, dans deux dictionnaires phonétiques, celui de l'utilisateur ou celui du logiciel. En cas d'échec, elles sont générées automatiquement à partir de la graphie.
Une seconde source de divergence concerne les conventions d'écriture des unités lexicales contenant des chiffres comme, par exemple, les dates, les heures, les numéros de téléphones, les nombres, etc. ASTALI considère un

certain nombre de conventions et réécrit en interne l'unité lexicale en toutes lettres avant de le phonétiser. Toutefois, il ne peut pas être exhaustif au niveau de ces conventions. Quelques exemples de traitement sont donnés dans le tableau 1.

Texte dans la transcription orthographique	Réécriture interne pour phonétisation
2h15	deux heures quinze ; deux heures et quart
A31	a trente et un
02/01/2015	deux janvier deux mille quinze ; zéro deux zéro un deux mille quinze

Tableau 1 : Exemples de traitement d'unités lexicales comportant des chiffres.

3.2 Parole spontanée

Les corpus de français contemporain oral rassemblés dans le projet ORFEO sont des enregistrements d'interviews, de dialogues ou de réunions. Ils sont donc essentiellement constitués de parole spontanée comportant des hésitations, des pauses, des reprises, des apocopes, néologismes, des prononciations déviantes, des suites de syllabes incompréhensibles. Les enregistrements incluent également de nombreux bruits : spécifiques à l'enregistrement, bruits d'ambiance (froissement de papier, chaises, sonneries, …), respirations, rires, exclamations, applaudissements. Tous ces phénomènes sont décrits dans les transcriptions sous différentes formes et souvent en utilisant un texte libre en langage naturel comme le montre l'exemple suivant présentant deux codages différents de parole incompréhensible, soit « *** » soit une balise XML :

*Il y a un rapport avec le *** Ombres et Lumières*
Il m'a dit hein <Event desc= « suite de syllabes incompréhensibles »/> je te jure

Le logiciel ASTALI prend en compte certains de ces événements comme les reprises, les pauses et certains bruits. L'existence de conventions de transcription communes permettrait d'améliorer les outils d'alignement. Ainsi pour les bruits cela permettrait de mettre dans le graphe phonétique le modèle acoustique correspondant le mieux à l'évènement décrit (rires, applaudissements, sonnerie, etc). Pour les prononciations déviantes, cela permettrait de les ajouter automatiquement aux dictionnaires phonétiques et donc d'améliorer la qualité de l'alignement.

3.3 Parole superposée

Concernant la parole superposée, les corpus à aligner avaient été transcrits de deux manières différentes : soit avec des balises XML « Who » soit avec des symboles situés directement dans le texte de la transcription orthographique et donc sans marques temporelles associées au début et à la fin de la parole superposée. Un exemple est donné Figure 3. Ceci a imposé l'écriture d'une version spécifique d'ASTALI pour tenir compte de ce codage non standard.

3.4 Anonymisation

Les corpus réels oraux sont souvent anonymisés pour pouvoir être diffusés mais les méthodes d'anonymisation ne sont pas homogènes. Nous avons été confrontés aux cas suivants :

- le nom propre anonymisé a été remplacé dans le texte par une chaîne de caractères identifiable (exemple : *P*) et le segment de parole a été remplacé par un bip ou un silence. Un exemple est donné à la figure 2. Nous avons appris un modèle de bip spécifique et ASTALI a été adapté pour compléter le dictionnaire phonétique.
- le segment de parole a été remplacé par un bip et un silence mais dans le texte le nom propre a été remplacé par un autre nom propre afin de préserver l'anonymat (par exemple « Jean Dupont »). Malheureusement, il n'est pas possible d'aligner correctement le fichier audio puisque les noms propres remplacés ne sont pas identifiables.

Une convention concernant d'une part le signal acoustique utilisé pour anonymiser le signal audio et d'autre part le codage dans la transcription textuelle permettrait de rendre plus automatique et plus fiable le processus d'alignement.

4 Conclusion

L'alignement en phonèmes et en mots de corpus oraux de parole spontanée est intrinsèquement difficile à cause de la variabilité de la parole (reprises, hésitations, contractions, …, d'un enregistrement dans un environnement non contrôlé (présence de nombreux bruits), d'une grande diversité de locuteurs (genre, âge, accents) et de la présence de superpositions de parole. A ces difficultés s'ajoute le fait que les corpus ont été transcrits au sein d'entités différentes ayant leurs propres conventions. Ce manque d'homogénéité oblige à apporter des modifications à l'outil d'alignement pour s'adapter à chaque corpus. De plus, certaines indications qui pourraient être utiles pour améliorer l'alignement ne peuvent pas être prises en compte car leur traitement n'est pas automatisable. Il serait souhaitable pour une meilleure réutilisabilité des corpus oraux que les concepteurs de futurs corpus respectent un format de transcription commun ou à défaut emploient des annotations dont le traitement automatique soit réalisable.

5 Remerciements

Le développement de l'outil ASTALI a été en partie financé par le projet ANR ORFEO et celui de l'application Web par l'Equipex ORTOLANG.

6 Bibliographie

[1] Barras C., Geoffrois E.,Wu Z., Liberman M., “Transcriber: development and use of a tool for assisting speech corpora production”, Speech Communication special issue on Speech Annotation and Corpus Tools, Vol 33, No 1-2, January 2000.

[2] Bigi B. “SPPAS: a tool for the phonetic segmentations of Speech”, The eight international conference on Language Resources and Evaluation, Istanbul (Turkey), pages 1748-1755, 2012.

[3] Boersma P., Weenink D. “Praat: doing phonetics by computer” [Computer program],Version 5.4.08, retrieved 24 March 2015 from http://www.praat.org/, 2015

[4] Goldman J-Ph “EasyAlign: an automatic phonetic alignment tool under Praat” InterSpeech, Firenze (Italy), 2011.

[5] http://www.ortolang.fr/

[6] http://www.projet-orfeo.fr/

[7] Mella O., Fohr D."Two tools for Semi-automatic Phonetic Labelling of Large Corpora", First International Conference on Language Resources and Evaluation, Grenade, Espagne,1998

[8] Young S. J., Evermann G., Gales M. J. F., Hain T. , Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P. C. “The HTK Book”, version 3.4, 2006.

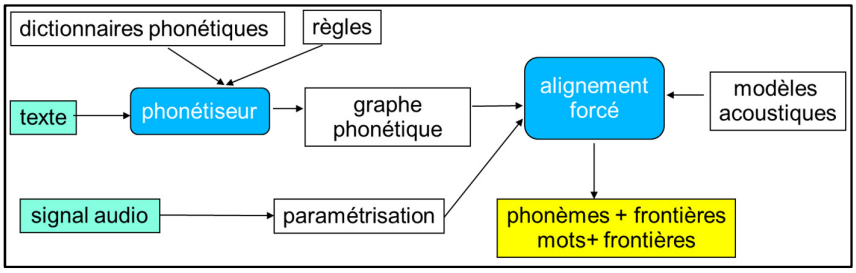

Figure 1 : Fonctionnement de la partie alignement parole-texte du logiciel ASTALI.

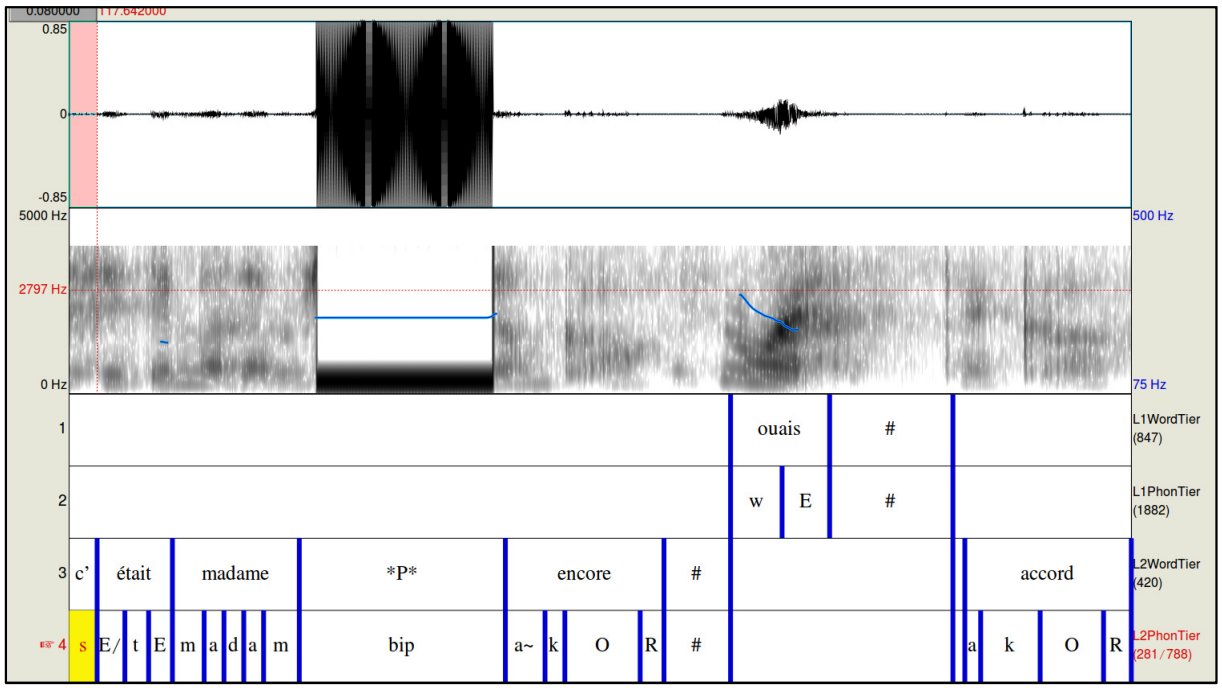

Figure 2 : Exemple de fichier résultat affiché par le logiciel Praat avec un segment de parole anonymisé (audio +texte). Les deux premiers « tiers » correspondent à l’alignement en mots et en phonèmes pour le locuteur L1 ; les deux derniers correspondent au locuteur L2.

```
<Turn speaker="spk2" startTime="181.866" endTime="182.279">
<Sync time="181.866"/>
par contre < tu écoutes un
</Turn>
<Turn speaker="spk3" startTime="182.279" endTime="183.208">
<Sync time="182.279"/>
ça me choque > aussi
</Turn>
```

```
<Turn speaker="spk1 spk3" startTime="737.741" endTime="742.324">
<Sync time="737.741" />
<Who nb="1" />
   les habitudes alimentaires changent
 <Who nb="2" />
   il y a moins de petits commerces
 </Turn>
```

Figure 3 : Exemple de deux variantes de transcription de parole superposée : à gauche, utilisation des symboles « > » et « < » directement dans le texte pour indiquer la superposition ; à droite utilisation de balises xml « Who ».