SPPAS : un outil « user-friendly » pour l'alignement texte/son

Brigitte Bigi
Laboratoire Parole et Langage, CNRS & Aix-Marseille Université,
5 avenue Pasteur, BP80975, 13604 Aix-en-Provence France
brigitte.bigi@lpl-aix.fr

RÉSUMÉ

Cet article présente *SPPAS*, le nouvel outil du LPL pour l'alignement texte/son. La segmentation s'opère en 4 étapes successives dans un processus entièrement automatique ou semi-automatique, à partir d'un fichier audio et d'une transcription. Le résultat comprend la segmentation en unités inter-pausales, en mots, en syllabes et en phonèmes. La version actuelle propose un ensemble de ressources qui permettent le traitement du français, de l'anglais, de l'italien et du chinois. L'ajout de nouvelles langues est facilitée par la simplicité de l'architecture de l'outil et le respect des formats de fichiers les plus usuels. L'outil bénéficie en outre d'une documentation en ligne et d'une interface graphique afin d'en faciliter l'accessibilité aux non-informaticiens. Enfin, SPPAS n'utilise et ne contient que des ressources et programmes sous licence libre GPL.

ABSTRACT

SPPAS : a tool to perform text/speech alignement

This paper presents SPPAS, a new tool dedicated to phonetic alignments, from the LPL laboratory. SPPAS produces automatically or semi-automatically annotations which include utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. SPPAS is currently implemented for French, English, Italian and Chinese. There is a very simple procedure to add other languages in SPPAS : it is just needed to add related resources in the appropriate directories. SPPAS can be used by a large community of users : accessibility and portability are important aspects in its development. The tools and resources will all be distributed with a GPL license.

MOTS-CLÉS : segmentation, phonétisation, alignement, syllabation.

KEYWORDS: segmentation, phonetization, alignement, syllabification.

1 Introduction

De nombreux développements de logiciels sont effectués dans les laboratoires de recherche comme support à la recherche ou aboutissement d'une recherche. Ces développements sont souvent innovants et intéressent rapidement d'autres entités que le laboratoire. Il se pose alors la question des choix pour permettre et pour accompagner leur valorisation, pour augmenter leur visibilité et leur capacité à susciter des collaborations. La portabilité et l'accessibilité du logiciel

JEP-TALN-RECITAL 2012, Atelier DEGELS 2012: Défi GEste Langue des Signes, pages 85–92,
Grenoble, 4 au 8 juin 2012.

en sont des points clés. La portabilité, car choisir une plate-forme pour un programme revient à en restreindre l'audience. L'accessibilité aux contenus et aux fonctions du logiciel s'avère également essentielle pour sa diffusion large à différentes communautés d'utilisateurs, car un logiciel est à la fois un objet scientifique mais aussi potentiellement un objet de transfert de technologie.

De nombreuses boîtes à outils pour réaliser différents niveaux de segmentations de la parole et l'apprentissage des modèles sous-jacents sont mis à disposition sur le web. Elles bénéficient parfois d'une large documentation, d'une communauté d'utilisateurs, de tutoriaux et de forums actifs. Des ressources (dictionnaires, modèles) sont également disponibles pour quelques langues. Pourtant, lorsqu'il s'agit d'effectuer des alignements texte/son, la plupart des phonéticiens choisissent de le faire manuellement même si plusieurs heures sont souvent nécessaire pour n'aligner qu'une seule minute de signal. Les raisons principalement évoquées concernent le fait qu'aucun outil n'est à la fois disponible librement, utilisable de façon simple et ergonomique, multiplateforme et, bien sûr, qui prend en charge la langue que veut traiter l'utilisateur. Ainsi, bien qu'elles soient très utilisées par les informaticiens, des boîtes à outils telles que, par exemple, HTK (Young, 1994), Sphinx (Carnegie Mellon University, 2011) ou Julius (Lee *et al.*, 2001), ne bénéficient toujours pas d'un développement qui permette une accessibilité à une communauté plus large d'utilisateurs, en particulier, à des utilisateurs non-informaticiens. HTK (Hidden Markov Toolkit), en effet, requiert un niveau de connaissances techniques très important à la fois pour son installation et pour son utilisation. Par ailleurs, HTK nécessite de s'enregistrer et il est proposé sous une licence qui limite les termes de sa diffusion (« *The Licensed Software either in whole or in part can not be distributed or sub-licensed to any third party in any form.* »). En outre, la dernière version (3.4.1) date de 2005. Malgré cela, HTK est largement utilisé et ses formats de données ont été largement repris par d'autres outils. Contrairement à HTK, Sphinx et Julius sont diffusés sous licence GPL. À ce titre, ils peuvent être re-distribués par des tiers, et ils sont régulièrement mis à jour. Par rapport à Sphinx, Julius offre toutefois l'avantage de pouvoir utiliser des modèles et dictionnaires au format HTK et de s'installer très facilement.

Développer un outil d'alignement automatique, s'appuyant uniquement sur des ressources libres (outils et données) et regroupant les critères nécessaire à son accessibilité à des non-informaticiens n'est pas uniquement un défi technique. On suppose en effet que si tel était le cas, cet outil existerait depuis longtemps ! Quelques outils sont toutefois déjà disponibles. P2FA (Yuan et Liberman, 2008) est un programme python multi-plate-forme qui permet de simplifier l'utilisation d'HTK pour l'alignement. De même, EasyAlign (Goldman, 2011) repose sur HTK, pour l'alignement automatique. Il se présente sous la forme d'un plugin pour le logiciel Praat (Boersma et Weenink, 2009), très utilisé pour l'annotation phonétique. EasyAlign (Goldman, 2011) offre l'avantage d'être simple à utiliser et propose une segmentation semi-automatique en Unités Inter-Pausales (IPUs), mots, syllabes et phonèmes pour 5 langues, mais il ne fonctionne que sous Windows. Dans (Cangemi *et al.*, 2010), les auteurs proposent les ressources pour l'italien et un logiciel d'alignement (licence GPL), également seulement pour Windows.

L'outil présenté dans cet article s'appelle SPPAS, acronyme de « SPeech Phonetization Alignment and Syllabification ». L'article en présente d'abord une vue d'ensemble puis décrit les 4 modules principaux : la segmentation en unités inter-pausales, la phonétisation, l'alignement et la syllabation. Enfin, une évaluation de la phonétisation est proposée.

2 SPPAS : vue d'ensemble

SPPAS peut être utilisé de diverses façons. La manière la plus simple d'utiliser SPPAS consiste à utiliser le programme *sppas.command* sous un système Unix, ou *sppas.py* sous Windows, qui lance l'interface graphique (voir figure 1). La division de SPPAS en différentes étapes (ou modules) permet une utilisation semi-automatique. Chacune des étapes de SPPAS peut être lancée puis le résultat corrigé manuellement avant de lancer l'étape suivante. Pour des utilisateurs avertis, SPPAS peut aussi être utilisé en ligne de commande : soit avec le programme général *sppas.py*, soit étape par étape (un ensemble d'outils est disponible dans le répertoite *tools*).

Un des points importants pour favoriser la diffusion destinée à une large communauté concerne la licence. SPPAS n'utilise que des ressources et des outils déposés sous licence GPL. SPPAS peut ainsi être distribué sous les termes de cette licence libre. Par ailleurs, pour des raisons de compatibilité, SPPAS manipule des fichiers TextGrid (format natif de Praat).

SPPAS permet de traiter différentes langues avec la même approche car la connaissance linguistique est placée dans les ressources et non dans les algorithmes. Actuellement, cet outil peut traiter des données en anglais, français, italien ou chinois. Ajouter une nouvelle langue L dans SPPAS consiste à ajouter les ressources nécessaires à chacun des modules, à savoir :

1. pour la phonétisation : un dictionnaire au format HTK, dans dict/L.dict,
2. pour l'alignement : un modèle acoustique (au format standard HTK-ASCII, appris à partir de fichiers audio échantillonés à 16000Hz), dans models/models-L,
3. pour la syllabation : un fichier de règles, dans syll/syllConfig-L.txt.

Si ces fichiers sont placés dans les répertoires appropriés et respectent la convention de nommage et le format requis, la nouvelle langue sera prise en compte automatiquement.

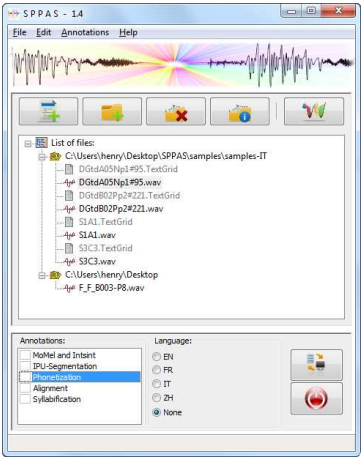

FIGURE 1 – *SPPAS* - Version 1.4

Afin d'assurer la portabilité, SPPAS est développé dans le langage python. En plus d'être multi-plateforme, le langage python offre l'avantage d'être orienté objet. Il permet ainsi un développement modulaire intéressant compte-tenu des objectifs du logiciel. En outre, python est interprété, il ne nécessite donc pas l'étape de compilation qui peut être difficile pour un non informaticien. L'interface graphique de SPPAS est développée à l'aide de la librairie wxPython, également très facile à installer.

Actuellement, SPPAS a notamment permis au LPL de participer à la campagne d'évaluation Evalita 2011, pour la tâche d'alignement forcé de dialogues en map-task, en italien (Bigi, 2012). SPPAS a également été choisi pour traiter les données du corpus AixOx, enregistrées dans le cadre du projet Amennpro, sur les phrases lues en français, par des locuteurs français ou des apprenants anglophones (Herment *et al.*, 2012). Il est par ailleurs régulièrement utilisé au LPL pour réaliser différentes tâches d'alignement.

Sur la figure 1, on voit une étape supplémentaire nommée « Momel and INTSINT ». Elle implémente la modélisation automatique de la mélodie proposée dans (Hirst et Espesser, 1993).

3 Les modules de SPPAS

3.1 Segmentation en IPUs

Cette segmentation consiste à aligner les macro-unités d'un texte (segments, phrases, etc) avec le son qui lui correspond. C'est un problème de recherche ouvert car, à notre connaissance, seul EasyAlign propose un tel algorithme. SPPAS utilise les pauses indiquées (manuellement) dans la transcription. L'algorithme s'appuie sur la recherche des pauses dans le signal et leur alignement avec les unités proposées dans la transcription (en supposant qu'une pause sépare chaque unité). Pour une durée fixée de pause et une durée fixée des segments de parole, une recherche dichotomique permet d'ajuster le volume pour trouver le bon nombre d'unités. Selon que le nombre d'unités trouvées est inférieur ou supérieur au nombre souhaité d'unités, la recherche est relancée avec des valeurs de durées de pauses et de durée des unités plus élevées ou moins élevées. La recherche s'arrête lorsque les 3 paramètres sont fixés correctement, c'est-à-dire qu'ils permettent de trouver le bon nombre d'unités. Cet algorithme a été appliqué à un corpus de lecture de mots et au corpus AixOx (Herment *et al.*, 2012) de lecture de petits paragraphes (3-6 phrases). La figure 2 montre une segmentation de ce dernier. SPPAS a ainsi permis ainsi un gain de temps substantiel.

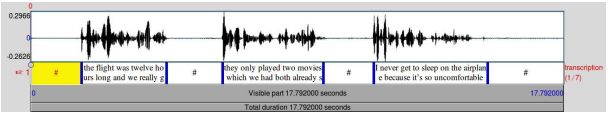

FIGURE 2 – Segmentation en IPUs

3.2 Phonétisation

La phonétisation, aussi appelée conversion graphème-phonème, consiste à représenter les unités (mots, syllabes) d'un texte par des symboles phonétiques. Il existe deux familles d'approches dans les méthodes de phonétisation : celles reposant sur des règles (proposées par des experts et/ou apprises sur corpus) et celles s'appuyant uniquement sur un dictionnaire. SPPAS implémente cette dernière approche. SPPAS propose aussi un plugin, nommé ESPPAS, qui permet d'utiliser l'approche à base de règles pour le français.

Approche à base de dictionnaire : Il n'y a pas d'algorithme spécifique dans cette approche. Le principe réside simplement à consulter le dictionnaire pour en extraire la prononciation de chaque entrée observée. Les deux situations suivantes peuvent survenir :

– une entrée peut se prononcer de différentes manières. C'est le cas notamment des homographes hétérophones, mais aussi des accents régionaux ou des phénomènes de réductions propres à l'oral. Dans ce cas, SPPAS ne choisit pas *a priori* la prononciation. Toutes les variantes présentes dans le dictionnaire sont cumulées dans la phonétisation.
– une entrée peut être absente du dictionnaire. SPPAS peut soit la remplacer par le symbole UNK, soit produire une phonétisation automatique. Dans ce cas, l'algorithme repose sur une recherche « longest matching », indépendante de la langue. Il cherche, de gauche à droite, les segments les plus longs dans le dictionnaire et recompose la phonétisation des segments pour créer la phonétisation du mot absent.

Par exemple, pour les mots « je » et « suis » le dictionnaire propose :

je [je] jj	suis [suis] ss yy ii
je(2) [je] jj eu	suis(2) [suis] ss yy ii zz
je(3) [je] ch	suis(3) [suis] ss uu ii
	suis(3) [suis] yy ii

Pour l'énoncé « je suis », SPPAS propose alors la phonétisation : « jj|jj.eu|ch ss.yy.ii|ss.yy.ii.zz|ss.uu.ii|yy.ii » dans laquelle les espaces séparent les mots, les points séparent les phonèmes et les barres verticales séparent les variantes. L'utilisateur peut laisser la phonétisation telle quelle (processus entièrement automatique) : c'est l'aligneur qui choisira la phonétisation. La phrase phonétisée sera l'une des combinaisons possibles des variantes proposées par SPPAS pour chaque mot. Dans un processus semi-automatique, l'utilisateur peut choisir la phonétisation appropriée (ou la modifier) manuellement. Pour des raisons de compatibilité, SPPAS utilise des dictionnaires au même format que ceux d'HTK. C'est un format ASCII éditable ; il peuvent donc être facilement modifiés avec un éditeur de texte.

Approche à base de règles : Dans le cadre de notre étude, notre choix s'est porté sur l'outil LIA_Phon (Bechet, 2001), pour deux raisons. La première parce qu'il est diffusé sous licence GPL, donc facilement accessible et par ailleurs, suffisamment bien documenté, facile d'utilisation et multi-plateformes. La seconde car il est connu pour produire une phonétisation de qualité.

En dehors de l'étape de transcription graphème-phonème, généralement traitée par une approche à base de règles, de nombreux traitements linguistiques sont nécessaires afin de lever les ambiguïtés d'oralisation du texte écrit (formatage du texte, homographes hétérophones, liaisons, phonétisation des noms propres, sigles ou emprunts à des langues étrangères, etc). Les outils inclus dans le LIA_Phon peuvent se décomposer en trois modules : les outils de formatage et d'étiquetage, les outils de phonétisation et les outils d'exploitation des textes phonétisés. Dans

la présente étude, nous faisons appel aux deux premiers modules. Le plugin « ESPPAS » (pour Enriched-SPPAS) encapsule le LIA_Phon pour l'utiliser facilement dans SPPAS.

3.3 Alignement en phonèmes et en mots

L'alignement en phonèmes consiste à déterminer la localisation temporelle de chacun des phonèmes d'une unité. SPPAS fait appel à Julius pour réaliser l'alignement. Julius est essentiellement dédié à la reconnaissance automatique de la parole. Il est distribué sous licence GPL, avec des versions exécutables simples à installer. Pour réaliser l'alignement, Julius a besoin d'une grammaire et d'un modèle acoustique. La grammaire contient la (ou les) prononciation(s) de chaque mot et l'indication des transitions entre les mots. L'alignement requiert aussi un modèle acoustique qui doit être au format HTK-ASCII, appris à partir de fichiers audio en 16000hz. Dans une première étape, Julius sélectionne la phonétisation et la segmentation en phonèmes est effectuée lors d'une seconde étape. La segmentation en mots est déduite de cette dernière.

La table 1 synthétise les informations relatives aux ressources incluses dans SPPAS. Il contient le nombre d'entrées du dictionnaire et la quantité de données utilisées pour l'apprentissage des modèles acoustiques. Les ressources de l'anglais proviennent du projet VoxForge (http ://www.voxforge.org), avec le dictionnaire du CMU.

Langue	Dictionnaire	Modèle Acoustique
Français	348k, 305k variantes	7h30 CID et 30min AixOx, triphones
Italien	390k, 5k variantes	3h30 CLIPS dialogues map-task, triphones
Chinois simplifié	353 syllabes	1h36 phrases lues, monophones

TABLE 1 – Ressources de SPPAS - Version 1.4

3.4 Syllabation

SPPAS encapsule le syllabeur du LPL (Bigi *et al.*, 2010). Il consiste à définir un ensemble de règles de segmentation entre phonèmes. Il repose sur les deux principes suivants : 1/ une syllabe contient une seule voyelle ; 2/ une pause est une frontière de syllabe. Ces deux principes résument le problème de syllabation en la recherche de frontières de syllabes entre deux voyelles. Les phonèmes sont alors regroupés en classes et des règles de segmentation entre ces classes sont établies, comme dans l'exemple suivant :

Transcription	et donc on mange sur la baignoire donc c'est c'est ça
Phonèmes	e d ɔ̃ k ɔ̃ m ɑ̃ ʒ s y ʁ l ɑ b e n w ɑ ʁ d ɔ̃ k s e s e s ɑ
Classes	V O V O V N V F F V L L V O V N G V L O V O F V F V F V
Syllabes	e . dɔ̃ . kɔ̃ . mɑ̃ʒ . syʁ . lɑ . be . nwɑʁ . d ɔ̃ k . se . se . sɑ

Le programme utilise un fichier de configuration qui décrit la liste des phonèmes et leur classe, ainsi que la liste de toutes les règles. Il peut être facilement modifié, ce qui rend l'outil applicable à d'autres langues. SPPAS inclut les fichiers de configuration pour la syllabation du français et de l'italien (ce dernier n'a pas été évalué).

4 Évaluations

Nous présentons ici des évaluations que nous avons réalisées sur la phonétisation du français. Les évaluations sur la phonétisation et l'alignement de l'italien sont présentées dans (Bigi, 2012). Nous avons d'abord construit un corpus, nommé « MARC-Fr - Manual Alignments Reference Corpus for French », entièrement phonétisé et aligné manuellement par un expert phonéticien. Il est déposé sous licence GPL sur la forge SLDR [1]. Il est composé de 3 corpus :

– 143 secondes d'extraits du CID, corpus conversationnel décrit dans (Bertrand *et al.*, 2008)
– 137 secondes d'extraits du corpus AixOx, corpus de lecture décrit dans (Herment *et al.*, 2012),
– 134 secondes d'un extrait d'Yves Cochet lors d'un débat à l'Assemblée nationale portant sur le « Grenelle II de l'environnement » décrit dans (Bigi *et al.*, 2011).

La transcription est en orthographe standard et contient les pauses pleines, les pauses perçues, les rires, les bruits, les amorces et les répétitions. D'autres résultats avec différents enrichissements de la transcription sont proposés dans (Bigi *et al.*, 2012)). Les évaluations sont effectuées avec l'outil Sclite (NIST, 2009). Il calcule le taux d'erreurs de la phonétisation (Err) qui somme les erreurs de substitution (Sub), de suppression (Del) et d'insertion (Ins). Les résultats sont présentés dans le tableau 2. La phonétisation du CID est meilleure en utilisant SPPAS, tandis que pour les deux autres corpus, la phonétisation est meilleure en utilisant le LIA_Phon. Cependant, puisque SPPAS utilise une approche à base de dictionnaire qui dépend énormément des ressources dont il dispose, il bénéficie d'une marge de progression assez importante. Le dictionnaire pourrait en effet être amélioré en vérifiant manuellement les entrées. Il faudrait aussi améliorer le modèle acoustique, en ajoutant des données d'apprentissage.

		Sub	Del	Ins	Err
CID	SPPAS-dico	3,6	2,1	7,6	**13,2**
	LIA_Phon	2,7	1,4	10,3	14,4
AixOx	SPPAS-dico	3,1	2,4	2,9	8,4
	LIA_Phon	1,4	2,3	2,9	**6,5**
Grenelle	SPPAS-dico	1,7	1,7	4,1	7,4
	LIA_Phon	1,0	1,2	4,1	**6,2**

TABLE 2 – Pourcentages d'erreurs de la phonétisation

5 Perspectives

SPPAS est un outil qui permet d'aligner automatiquement textes et sons. Sa particularité vient du fait qu'il s'adresse à une communauté très large d'utilisateurs. De nombreux efforts ont été réalisés en ce sens lors de son développement : portabilité, accessibilité, modularité, licence libre, etc. Les développements à venir suivent 3 directions : la première consiste valoriser la version actuelle (documentation, tutoriel, dépôt dans une forge, packaging, etc), le deuxième consiste en l'ajout de nouveaux modules (détection de pitch, tokenizer multilingue), la troisième est l'ajout de nouvelles ressources pour la prise en charge de nouvelles langues (et/ou la création d'un modèle multilingue).

1. Speech Language Data Repository, http ://www.sldr.fr

Références

BECHET, F. (2001). LIA_PHON - un système complet de phonétisation de textes. *Traitement Automatique des Langues*, 42(1).

BERTRAND, R., BLACHE, P., ESPESSER, R., FERRÉ, G., MEUNIER, C., PRIEGO-VALVERDE, B. et RAUZY, S. (2008). Le CID - Corpus of Interactional Data. *Traitement Automatique des Langues*, 49(3):105–134.

BIGI, B. (2012). The SPPAS participation to Evalita 2011. *In EVALITA 2011 : Workshop on Evaluation of NLP and Speech Tools for Italian*, Rome, Italie.

BIGI, B., MEUNIER, C., NESTERENKO, I. et BERTRAND, R. (2010). Automatic detection of syllable boundaries in spontaneous speech. *In Language Resource and Evaluation Conference*, pages 3285–3292, La Valetta, Malta.

BIGI, B., PORTES, C., STEUCKARDT, A. et TELLIER, M. (2011). Multimodal annotations and categorization for political debates. *In In ICMI Workshop on Multimodal Corpora for Machine learning (ICMI-MMC)*, Alicante, Espagne.

BIGI, B., PÉRI, P. et BERTRAND, R. (2012). Orthographic Transcription : Which Enrichment is required for Phonetization ? *In The eighth international conference on Language Resources and Evaluation*, Istanbul (Turkey).

BOERSMA, P. et WEENINK, D. (2009). Praat : doing phonetics by computer, http ://www.praat.org.

CANGEMI, F., CUTUGNO, F., LUDUSAN, B., SEPPI, D. et COMPERNOLLE, D.-V. (2010). Automatic speech segmentation for italian (assi) : tools, models, evaluation, and applications. *In 7th AISV Conference*, Lecce, Italie.

CARNEGIE MELLON UNIVERSITY (2011). CMUSphinx : Open Source Toolkit For Speech Recognition. http ://cmusphinx.sourceforge.net.

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

HERMENT, S., LOUKINA, A., TORTEL, A., HIRST, D. et BIGI, B. (2012). A multi-layered learners corpus : automatic annotation. *In 4th International Conference on Corpus Linguistics Language, corpora and applications : diversity and change*, Jaén (Espagne).

HIRST, D. J. et ESPESSER, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15:75–85.

LEE, A., KAWAHARA, T. et SHIKANO., K. (2001). Julius — an open source real-time large vocabulary recognition engine.". *In European Conference on Speech Communication and Technology*, pages 1691–1694.

NIST (2009). Speech recognition scoring toolkit, http ://www.itl.nist.gov/iad/mig/tools/.

YOUNG, S. (1994). The HTK Hidden Markov Model Toolkit : Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44.

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