Vers une annotation automatique de corpus audio pour la synthèse de parole

Olivier Boëffard Laure Charonnat Sébastien Le Maguer
Damien Lolive Gaëlle Vidal
Université de Rennes 1, Enssat, Lannion, France
olivier.boeffard@irisa.fr, laure.charonnat@univ-rennes1.fr,
sebastien.le_maguer@irisa.fr, damien.lolive@irisa.fr,
gaelle.vidal@univ-rennes1.fr

RÉSUMÉ

La construction de corpus de parole est une étape cruciale pour tout système de synthèse de la parole à partir du texte. L'usage de modèles statistiques nécessite aujourd'hui l'utilisation de corpus de très grande taille qui doivent être enregistrés, transcrits, annotés et segmentés afin d'être exploitables. La variété des corpus nécessaire aux applications actuelles (contenu, style, etc.) rend l'utilisation de ressources audio disponibles, comme les livres audio, très attrayante. C'est dans ce cadre que s'inscrit notre proposition de chaîne d'acquisition, de segmentation, et d'annotation de livres audio. Cette proposition tend vers la mise en place d'un processus automatique. Le processus proposé s'appuie sur une structure de données, ROOTS, qui établit des relations entre différents niveaux d'annotation. Cette méthodologie a été appliquée avec succès sur 11 heures de parole extraites d'un livre audio. Une vérification manuelle sur une partie du corpus annoté a montré l'efficacité du procédé.

ABSTRACT

Towards Fully Automatic Annotation of Audio Books for Text-To-Speech (TTS) Synthesis

Building speech corpora is a crucial step for every text-to-speech synthesis system. Nowadays, statistical models require enormous corpora that need to be recorded, transcribed, annotated and segmented to be usable. The variety of corpora necessary for recent applications (content, style, etc.) makes the use of existing audio resources very attractive. Taking the above considerations into account, a complete acquisition, segmentation and annotation chain for audio books, which tends to be fully automatic, is proposed. This process relies on a data structure, ROOTS, which establishes the relations between the different annotation levels. This methodology has been applied successfully on 11 hours of speech extracted from an audio book. A manual check, on a part of the corpus, has shown the efficiency of the process.

MOTS-CLÉS : Livres audio, annotation, segmentation, synthèse de la parole.

KEYWORDS: Audio books, annotation, segmentation, text-to-speech synthesis.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 731–738,
Grenoble, 4 au 8 juin 2012.

1 Usage de grands corpus pour la synthèse

L'usage de modèles statistiques, issus du domaine de la reconnaissance de la parole, intervient dans toutes les disciplines du traitement automatique des langues et de la parole. L'apprentissage de tels modèles, sur des unités de parole, nécessite un grand nombre d'observations, ce qui implique la mise en place de corpus de parole de grande taille. Une conséquence directe de la taille de ces corpus est que les méthodes jusqu'alors utilisées pour les constituer, les segmenter et les annoter montrent leurs limites. En synthèse de la parole, l'unité est généralement un phonème pris dans un contexte linguistique et acoustique précis. La représentation dans une base de données de l'ensemble des unités générées par les combinaisons de toutes les caractéristiques linguistiques et acoustiques (une vingtaine de descripteurs est utilisée pour HTS (Tokuda *et al.*, 2002)) est impossible. Cependant, l'apprentissage des modèles des unités présentes dans la langue visée peut être envisagé par l'analyse de grands corpus de parole naturelle.

La généralisation des ressources numérisées favorise la disponibilité de données de grand corpus de parole naturelle. Leurs natures sont très variées, ils peuvent être accompagnés ou non du texte, monolocuteurs ou multilocuteurs, amateurs ou professionnels, représentant une parole spontanée ou lue, etc. Dans le cadre d'une utilisation en synthèse de la parole, nous considérons ici le cas du livre audio qui permet de disposer d'un enregistrement accompagné du texte lu par un locuteur professionnel et de bonne qualité acoustique. Complétant voire remplaçant des corpus ad hoc, ils présentent de nombreux avantages. On peut trouver des livres audio différents mais enregistrés par un même locuteur permettant de multiplier les registres littéraires, comme on peut trouver une même œuvre lue par des locuteurs différents permettant des travaux sur des corpus parallèles. En outre, une particularité des données textuelles associées à un livre audio est qu'elles ne varient pas. On notera toutefois la possibilité de variations pour les cas ayant fait l'objet d'éditions différentes (traductions, œuvres inachevées).

Une étape importante pour la création d'une voix est celle de l'annotation du texte associé au signal de parole. Cette opération peut paraître simple si on se limite à la synchronisation des mots sur le signal prononcé, mais fait appel à des relations temporelles complexes si on s'intéresse à différents niveaux d'analyse comme l'annotation sémantique, lexicale, grammaticale, syntaxique, phonétique, prosodique, etc. Pour chacun de ces niveaux, des travaux ont permis de mettre au point des systèmes d'annotation automatique définissant l'ensemble des étiquettes à apposer sur le texte en lien avec le signal. Ces systèmes sont le plus souvent indépendants les uns des autres, ne travaillent pas toujours à la même échelle et peuvent recourir à des formats de description différents. Leur mise en œuvre sur un corpus demande souvent beaucoup de manipulations et génère un ensemble de fichiers hétérogènes et dispersés. Afin de limiter la désynchronisation des informations de description, nous avons récemment proposé une solution fondée sur la mise en relation de séquences, ROOTS (Barbot *et al.*, 2011). Cette approche permet de définir un ensemble de relations minimales qui existent entre différents niveaux de description. Des relations primitives sont définies et un mécanisme de composition de relations permet par des règles algébriques de décliner toutes les relations souhaitées entre deux séquences d'annotation.

L'objet de notre étude consiste à décrire une chaîne d'acquisition de corpus de parole à partir de livres audio. Le processus d'annotation automatique a été mis en œuvre pour traiter plusieurs dizaines d'heures en continu. L'étude se limite à des enregistrements monolocuteurs. La seule contrainte est de disposer du contenu sonore et du texte associé à ce contenu. En sortie du système de traitement, et selon les niveaux d'annotation souhaités, un ensemble d'énoncés ROOTS

stockés au format XML, structure les diverses relations allant du texte au signal.

Dans la partie 2, la chaîne d'annotation proposée est décrite. La structure permettant la représentation des informations liées à l'énoncé est ensuite présentée dans la partie 3. Cette structure est utilisée pour le processus de découpage et d'alignement du texte, détaillé en section 4, ainsi que pour conserver toutes les informations obtenues suite à la phase d'annotation, section 5. La partie 6 illustre tout ce processus en proposant une application de cette méthodologie.

2 Procédé d'annotation

La mise en place du procédé d'annotation doit respecter un certain nombre de contraintes dictées par l'usage du corpus annoté et la maîtrise des performances. La première contrainte concerne le texte qui devra être conservé sous sa forme originale, les écarts de lecture devront être signalés par des balises. La seconde concerne le découpage du corpus. Notre objectif étant d'annoter un corpus sur différents niveaux, nous sommes amenés à traiter des fragments de parole ou de texte de taille variable. En effet il est souhaitable pour une analyse syntaxique de disposer d'une phrase complète alors qu'une segmentation en phones est plus efficace sur des extraits courts. Le texte et les plages d'enregistrement seront donc découpés avec une granularité suffisamment fine pour pouvoir travailler sur des fragments courts qui pourront, le cas échéant, être réunis et former une phrase ou un fragment plus long à condition de toujours conserver la cohérence du texte. Enfin, pour garantir une annotation de qualité, nous veillerons à ce qu'une intervention manuelle soit possible, elle sera guidée par des indicateurs de confiance fournis par les différents outils intervenant dans le processus.

La chaîne d'annotation, présentée sur la figure 1, est constituée de deux étapes : la première consiste a fractionner l'enregistrement de plusieurs heures de parole et d'y associer le texte correspondant. Cette étape, d'autant plus coûteuse en temps que le découpage du signal est fin, nécessite le recours à un système de reconnaissance de la parole pour retrouver dans le texte complet l'ancrage de la transcription associée au signal. Les extraits sont ensuite regroupés pour reconstituer les phrases du texte original. La deuxième étape concerne l'annotation des phrases du texte et du signal mis en correspondance. La représentation des données par ROOTS est réalisée dès l'étape de découpage du livre-audio en phrases. Elle est enrichie au fur et à mesure de l'annotation des données par l'ajout de nouvelles séquences de description et de relations entre ces séquences.

3 Représentation du corpus

ROOTS est une librairie conçue pour manipuler un ensemble de structures de données comme un ensemble cohérent permettant la description et l'annotation de la parole. Chaque type d'annotation correspond à une séquence d'items. Un item correspond à des objets de nature très variée comme par exemple une transcription sous forme de texte, un label, un segment acoustique (i.e. défini par un début et une fin), etc. La seule contrainte imposée est qu'au sein d'une séquence les items doivent être homogènes. Cela assure la cohérence des séquences et une bonne séparation des types d'éléments. Des relations permettent de connecter les items des séquences entre eux, et de créer des relations de type n-vers-m. ROOTS produit des fichiers

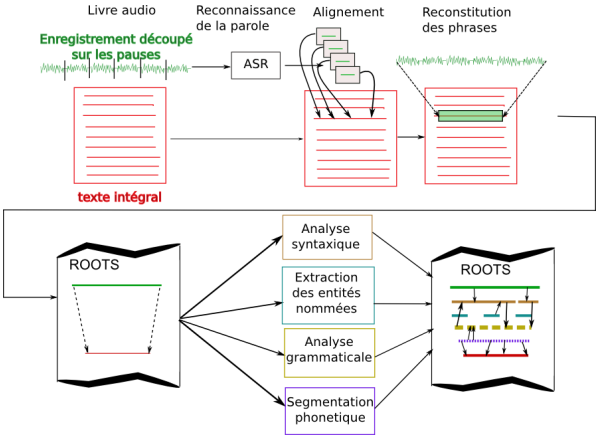

FIGURE 1 – Processus d'annotation d'un livre-audio

XML qui donnent la description complète des énoncés du corpus mais permet aussi d'exporter le contenu décrit vers des formats d'usage courant afin de garantir une interopérabilité avec les outils existants tels que Wavesurfer ou Transcriber (Barras *et al.*, 2001).

Un autre point positif apporté par ROOTS est que sa structure modulaire permet de conserver plusieurs séquences du même type en parallèle. Il est par exemple important de pouvoir conserver le texte d'origine, le texte prononcé (texte avec éventuellement des corrections prenant en compte des erreurs sur les mots ou bien des prononciations particulières), et également le texte en sortie de l'analyse syntaxique (par exemple, « j'avais » correspond réellement à deux mots). Ces trois types de texte correspondent à trois séquences d'items reliées les unes aux autres par des relations précisant la correspondance entre les éléments. Ce point de vue permet donc de ne perdre aucune information et de conserver dans une structure unique (pouvant correspondre physiquement à plusieurs fichiers) les différentes annotations d'un énoncé.

4 Découpage du signal de parole et alignement avec le texte

Plusieurs travaux portent sur l'alignement de longs textes et des plages audio correspondantes. (Braunschweiler *et al.*, 2010) propose un système automatique alignant des zones de textes d'un livre-audio pour des applications en synthèse de la parole à partir du texte. L'objectif pour d'autres était de faire face à des transcriptions approximatives (Tao *et al.*, 2010) ou d'effectuer un alignement sans découper le texte (Moreno et Alberti, 2009)(Prahallad *et al.*, 2007). Dans notre cas, le texte devra être découpé pour effectuer les différentes analyses et pour faciliter sa représentation, en particulier dans l'hypothèse d'une vérification manuelle.

Nous avons choisi d'effectuer l'alignement du texte et du son en 3 étapes : (1) découpage de l'enregistrement sur des pauses, (2) reconnaissance du texte associé à chaque fragment sonore

par un système de reconnaissance automatique de la parole (ASR), (3) alignement entre le texte reconnu et le texte original.

Le découpage de l'enregistrement repose sur l'observation des niveaux d'énergie et sur la longueur des silences. Les seuils sont fixés selon le débit du locuteur et les niveaux d'enregistrement. L'idéal est d'obtenir un fragment sonore en dessous de la phrase, permettant ainsi de reconstituer les phrases tout en gardant des points d'ancrage à l'intérieur de chaque phrase dans le cas des phrases longues.

La reconnaissance du texte correspondant à chaque extrait sonore est réalisé par le système de reconnaissance de Nuance (Nuance, 2010). Les modèles sont indépendants du locuteur, le modèle de langage est appris sur le texte intégral du livre. En comparant le texte original et le texte issu de la reconnaissance, il est possible de mesurer le taux d'erreur de reconnaissance des mots. Ces erreurs peuvent être dues à une défaillance du système de reconnaissance ou à une lecture erronée du texte (mauvaise prononciation ou modification du texte). Les écarts entre texte reconnu et texte original pourront être signalés afin de permettre à un opérateur de contrôler les passages concernés.

Le texte obtenu par le système de reconnaissance est ensuite aligné sur le texte original par le calcul d'une distance de Levenshtein définie au niveau du mot. Les extraits reconnus sont traités dans l'ordre du texte, ce qui permet de traiter des occurrences d'un même fragment de texte situé en différents endroits de l'énoncé. Lorsque la position du premier extrait dans le texte original est déterminée, il est associé au fichier sonore puis supprimé du texte original. L'opération est reproduite pour les extraits suivants jusqu'à la fin du texte.

Enfin pour conserver au mieux la structure du texte, les extraits sont regroupés en phrases en respectant les ponctuations majeures. Lorsqu'un extrait n'est pas terminé par une ponctuation forte il est simplement regroupé avec l'extrait suivant. Cependant l'information sur la frontière entre les deux extraits est conservée. Le texte et le signal découpés sont mis dans un format compatible avec le logiciel Transcriber (Barras *et al.*, 2001). Un tour de parole est constitué d'une phrase ou d'un ensemble de phrases contiguës de manière à ne pas forcer une découpe trop tôt dans la chaîne des traitements. Des points de synchronisation matérialisent dans le texte la frontière entre deux fragments sonores consécutifs. L'usage de transcriber permet aussi de placer des balises signalant les désaccords entre le texte original et le texte reconnu en particulier lorsqu'ils surviennent à la frontière entre deux extraits ce qui peut mettre en cause la pertinence du découpage du texte par rapport au signal de parole. La mise en relation texte/signal est ensuite automatiquement convertie au format ROOTS.

5 Annotation des données

Au cours de la deuxième étape, les descriptions textuelles et sonores sont fournies par l'objet ROOTS aux différents systèmes d'annotation qui donneront en retour leur propre analyse. Actuellement, les niveaux d'annotations utilisés sont les suivants : une extraction d'entités nommées, une analyse syntaxique, une analyse en POS (Part-Of-Speech), une segmentation phonétique et une extraction des proéminences prosodiques. Les analyses syntaxiques et grammaticales sont réalisées par des logiciels fournis par Synapse Développement, les analyses acoustiques sont réalisées à l'aide de nos propres outils. Les informations obtenues à chaque analyse sont intégrées

au fichier ROOTS affinant ainsi la description du corpus et permettant d'établir de nouvelles relations entre les éléments des différentes annotations. Par exemple, après une analyse complète, il est possible de retrouver aisément pour un phonème donné le mot auquel il appartient ainsi que son étiquette grammaticale.

6 Application

Le processus complet d'annotation a été expérimenté sur une œuvre de Marcel Proust "Albertine disparue". Ce livre contient environ 120 000 mots, son enregistrement dure 11 heures et 43 minutes. Proust ayant réécrit certains passages, le texte a tout d'abord été contrôlé pour en garantir la version.

Les plages audio ont été découpées sur les silences d'une durée supérieure à une seconde. Le découpage a produit 11693 fichiers qui correspondent à des phrases ou groupes de souffle d'une durée moyenne de 3 secondes. Les fichiers ont ensuite été transmis au système de reconnaissance automatique de la parole. Les écarts entre le texte en sortie du système de reconnaissance et le texte d'origine portent sur 5,2% des mots, ce qui inclut les différences sur l'orthographe des mots et sur leurs accords grammaticaux. Ce taux pourrait être réduit en effectuant une adaptation des modèles au locuteur à partir de quelques phrases du livre. Les textes issus du système de reconnaissance ont ensuite été utilisés pour découper le texte original conformément au découpage du signal. Lorsqu'un désaccord porte sur un mot en début ou en fin de segment, en particulier lorsqu'il s'agit d'une insertion ou d'une élision, l'alignement ne permet pas de garantir un bon découpage du texte autour de ce mot. Le cas est alors signalé pour permettre une intervention manuelle. Sur l'ensemble des 11693 fichiers, le cas s'est produit 969 fois (8,3% des segments). Une vérification manuelle a pu établir qu'une erreur de découpage avait réellement eu lieu dans 8% des alertes soit sur 78 fichiers. Le nombre élevé de fausses alertes provient du fait que tous les désaccords en frontière de segment ont été signalés (insertion, élision et substitution), or de nombreuses substitutions ne donnent pas lieu à un mauvais alignement.

Les fichiers ont ensuite été regroupés au sein de fichiers de type Transcriber (un fichier par plage de CD soit une quinzaine de minutes de parole en moyenne). Les textes ont été regroupés en tours selon leur ponctuation. Nous obtenons un total de 3340 tours dont les durées varient entre 660 ms et 1 min 22 s. Un tour regroupe en moyenne 3,5 segments définis lors de la précédente étape. Comme nous l'avons précisé les positions d'ancrage de ces segments sont conservées.

Lorsque les fichiers Transcriber sont créés toutes les informations concernant les désaccords entre le texte d'origine et les sorties du système de reconnaissance sont signalés par une balise afin de faire l'objet d'une vérification ciblée. L'opérateur peut alors ajouter de nouvelles balises signalant les erreurs de prononciation ou les erreurs de lecture. Dans notre cas, 86 balises lexicales (remplacement d'un mot ou d'une expression par une autre) et 213 balises de prononciation ont été ajoutées, et 78 corrections ont été effectuées sur le découpage du texte.

Le corpus ainsi découpé et aligné est alors mis dans une structure de type ROOTS que viennent interroger les différents systèmes d'annotation. À l'heure actuelle, les analyses linguistiques n'ont pas encore été validées mais la segmentation en phonèmes a été l'objet d'une vérification manuelle sur une partie du corpus d'une durée de 2 heures et 16 minutes.

La segmentation automatique est obtenue par l'application de modèles de Markov (un modèle

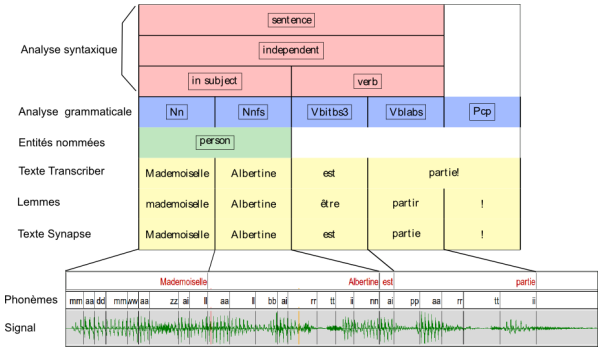

FIGURE 2 – Représentation d'une phrase annotée du corpus

par phonème, un modèle de pauses courte et longue, un modèle d'inspiration, un modèle de début et fin de phrase, soit un total de 40 modèles différents) sur un graphe de phonétisation. Les modèles sont indépendants du contexte, ils sont appris sur des vecteurs d'observation constitués de 39 coefficients (12 coefficients MFCC, leurs dérivées au premier et second ordre et l'energie) calculés sur des trames de 30 ms décalées de 10 ms. Le graphe de phonétisation est obtenu par la phonétisation du texte par Liaphon (Bechet, 2001) enrichie de variantes concernant certains phonèmes : les phonèmes /ø/ et /ə/ sont optionnels pour une majorité des mots, les liaisons sont optionnelles et peuvent être précédées ou suivies de courts silences, et enfin les pauses peuvent être remplacées par des inspirations. Les segments de paroles utilisés pour la segmentation sont les plus courts possibles afin de pouvoir augmenter la complexité des graphes par l'ajout de variantes sans remettre en cause l'efficacité des algorithmes d'alignement. Le texte fourni au système de segmentation ne tient pas compte des labels de lexique et de prononciation insérés dans les fichiers Transcriber mais tient compte des corrections concernant le découpage du texte.

Le résultat de la segmentation automatique a conduit à l'annotation de 419 742 segments (phones et silences). Sur 82 936 unités phonétiques validées manuellement, 94% ont été correctement étiquetées, 2,5% sont absentes et 3,3% ont été remplacées par une autre étiquette. À cela s'ajoute 2,8% d'insertion de phonèmes. Une grande partie des élisions (46%) concerne le phonème /ø/ qui est la plupart du temps optionnel et dont 24% des occurrences n'ont pas été détectées. Les substitutions sont dominées par le remplacement de /e/ par /ɛ/ du à une prononciation particulière du locuteur (41% des substitutions) suivi des confusions pause/inspiration (19%). Les insertions concernent les ajouts de pauses ou d'inspiration (72% des insertions). Ces résultats sont corrects mais ils peuvent être améliorés par l'ajout de variantes adaptées à la prononciation du locuteur et par un post-traitement sur les pauses et inspirations. L'alignement des phonèmes sur le signal est également correct, nous avons 86% des frontières de phonèmes qui sont placées à moins de 20 ms de la position définie par l'opérateur humain, l'écart moyen étant de 8,7ms ce qui est inférieur au décalage entre deux trames d'observation. Ces mesures ont également été effectuées avant l'intervention manuelle corrigeant les erreurs de découpage du texte, le nombre d'étiquettes correctes était de 91% et la proportion de frontière placées à moins de 20ms était de 86%.

La structure ROOTS est ensuite enrichie des informations fournies par les systèmes d'annotation. La mise en relation de ces informations permet de connaître pour chaque fragment de texte ou de son l'ensemble des étiquettes auquel il est rattaché. La figure 2 présente l'exemple d'une phrase où l'on constate que le texte est représenté dans plusieurs séquences : le texte Transcriber donne un découpage selon les espaces, alors que le texte Synapse sépare les éléments selon leurs natures.

7 Conclusion

L'automatisation de l'ensemble des traitements d'un livre-audio pour obtenir un corpus annoté sur différents niveaux permet de constituer rapidement de nouveaux corpus pour des études en laboratoire. Le gain de temps réalisé pour la création d'une nouvelle voix de synthèse est considérable si on ajoute le temps épargné par la suppression de l'étape d'enregistrement à celui gagné lors de l'annotation du corpus. La représentation du corpus sous forme de fichiers XML obtenus grâce à ROOTS supprime les difficultés liées à l'hétérogénéité des fichiers fournis par chaque système d'annotation et simplifie sa manipulation. Cependant, la chaîne d'annotation doit être renforcée par des indices de confiance sur chaque étape de l'annotation, permettant, soit d'éliminer les zones douteuses, soit de fonctionner en mode supervisé en fournissant à un opérateur les informations nécessaires à une intervention rapide ainsi que les données sous un format adapté aux outils de vérification.

Références

BARBOT, N., BARREAUD, V., BOËFFARD, O., CHARONNAT, L., DELHAY, A., LE MAGUER, S. et LOLIVE, D. (2011). Towards a versatile multi-layered description of speech corpora using algebraic relations. *In Proc. of Interspeech*, pages 1501–1504.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (2001). Transcriber : Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

BECHET, F. (2001). Liaphon - un système complet de phonétisation de textes. *Traitement Automatique des Langues (T.A.L.), édition Hermes*, 42(1).

BRAUNSCHWEILER, N., GALES, M. et BUCHHOLZ, S. (2010). Lightly supervised recognition for automatic alignment of large coherent speech recordings. *In Proc. of Interspeech*, pages 2222–2225.

MORENO, P. et ALBERTI, C. (2009). A factor automaton approach for the forced alignment of long speech recordings. *In Proc. of IEEE ICASSP*, pages 4869–4872.

NUANCE (2010). Dragon Naturally Speaking - SDK Server Edition - version 10.

PRAHALLAD, K., TOTH, A. et BLACK, A. (2007). Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases. *In Proc. of Interspeech*, pages 2901–2904.

TAO, Y., XUEQING, L. et BIAN, W. (2010). A dynamic alignment algorithm for imperfect speech and transcript. *Computer Science and Information Systems*, 7(1):75–84.

TOKUDA, K., ZEN, H. et BLACK, A. W. (2002). An hmm-based speech synthesis system applied to english. *In Proceedings of the IEEE Workshop on Speech Synthesis*, pages 227–230.