Proposition de travaux de thèse, IRISA, équipe Expression

Titre : *Optimisation de scripts d'enregistrement pour la lecture expressive de livres audio*

Mots-clefs : synthèse de la parole expressive ; optimisation et apprentissage.

Contexte : Le projet vise à étudier la réalisation automatique de livres audio à l'aide d'une voix de synthèse. La durée d'écoute de l'intégralité d'une œuvre nécessite une voix de haute qualité à l'expressivité adaptée.

Un système de synthèse vocale à partir du texte (TTS) produit un signal de parole correspondant à une vocalisation d'un texte donné. Ces dernières années, la TTS a fait de nombreux progrès en termes de qualité acoustique et d'intelligibilité, mais la production d'une voix expressive de très bonne qualité reste un verrou scientifique (voir [1] et ses références). Cette qualité vocale dépend fortement du système TTS (paramétrique, ou par sélection et concaténation d'unités sonores) et du corpus de parole utilisé.

Fréquemment, la création d'un tel corpus de parole nécessite l'enregistrement de la lecture d'un script spécifique avec des types d'expressivité donnés. Ce processus d'enregistrement étant complexe et coûteux, de nombreux travaux portent sur la création d'un script permettant de couvrir un maximun d'événements souhaités tout en minimisant sa durée (voir [2, 3, 4] et leurs références).

Proposition : La problématique étudiée dans ce projet de thèse est la création de livres audio sous une forme hybride : il s'agit d'enregistrer une partie minimale des livres visés pour produire une voix de synthèse la mieux adaptée au texte restant à vocaliser. Plus généralement, le sujet porte sur l'étude des méthodes de construction et d'enrichissement automatique de scripts d'enregistrement afin de produire une voix de synthèse de haute qualité pour des textes pré-définis d'expressivité variée. Cette approche se formalise en un problème d'optimisation d'un compromis entre qualité des messages acoustiques finaux et quantité de texte à enregistrer.

Un premier axe de travail concerne la problématique de l'évaluation subjective et objective. Dans le cadre général de la synthèse de la parole, l'évaluation de la qualité des signaux produits est un problème qui fait l'objet de nombreuses études (voir par exemple [5, 6, 7]) mais qui reste difficile. En quoi le fait de connaître à l'avance le texte à vocaliser ou de disposer de signaux de parole naturelle réalisés dans le même contexte permet de simplifier ce problème ? D'un autre coté, le livre audio produit sera un mélange de signaux naturels et de signaux de synthèse. Il sera donc nécessaire d'étudier et de proposer des approches spécifiques pour évaluer de tels objets et, en particulier, dépasser l'évaluation subjective à l'échelle de la phrase.

Un deuxième axe de travail porte sur la construction automatique du script d'enregistrement et la définition d'un compromis entre la qualité des signaux et la taille de l'enregistrement associé. Plusieurs verrous sont déjà identifiés. Comment les descripteurs textuels influencent-ils la qualité finale ? En particulier, quelles méthodes d'apprentissage, guidées par des mesures objectives de qualité, conduisent aux jeux de descripteurs optimaux ?

Un dernier axe de travail porte sur l'étude de la prise en compte des altérations entre le résultat théorique attendu lié au script d'enregistrement et le signal acoustique réel issu de la phase d'enregistrement. Comment détecter ces variations et adapter dynamiquement le script afin de conserver la qualité acoustique finale initialement attendue ?

Environnement de travail : le projet sera réalisé au sein de l'équipe Expression de l'IRISA, dans sa composante lannionnaise spécialisée sur les problématiques de synthèse de la parole et de traitement automatique des langues. Il sera encadré conjointement par Damien Lolive

et Jonathan Chevelu (IRISA-ENSSAT Lannion, Université de Rennes1) et bénéficiera d'un financement sur trois ans (financement des conseils départemental et régional). L'équipe dispose d'un moteur de synthèse de la parole par corpus, d'un moteur statistique (HTS), d'un studio d'enregistrement, d'une plate-forme de tests d'écoute [8] et d'une collection de livres audio annotés [9] qu'elle enrichit dans le cadre d'un projet ANR.

Profil du candidat : Le candidat sera diplômé d'un master informatique ou de toute autre formation équivalente. Compte-tenu du sujet, des compétences avancées en algorithmique et programmation seront requises. Le candidat disposera de la motivation et des facultés nécessaires pour aborder les domaines de recherche de la synthèse de la parole, de l'apprentissage artificiel et du traitement automatique des langues.

Contacts :
Damien LOLIVE (damien.lolive@irisa.fr) et Jonathan CHEVELU (jonathan.chevelu@irisa.fr)

Bibliographie

[1] D. Govind, S. R. Mahadeva Prasanna, Expressive speech synthesis : a review, Int. J. of Speech Tech., p. 1-24, 2013.

[2] H. François, Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue, thèse de l'Univ. de Rennes 1, 2002

[3] D. Cadic, Optimisation du procédé de création de voix en synthèse par sélection, thèse de l'Univ. de Paris 11, 2011

[4] N. Barbot, O. Boëffard, J. Chevelu, A. Delhay, Large linguistic corpus reduction with SCP algorithms, Computational Linguistics 41(3) : 355-383, 2015

[5] N. Campbell, Evaluation of speech synthesis : from reading machines to talking machines, Evaluation of Text and Speech Synthesis, (L. Dybjoer at al. Eds.) , Chapitre 2, 2007

[6] J. Chevelu, D. Lolive, S. Le Maguer, D. Guennec, How to compare TTS systems : a new subjective evaluation methodology focused on differences, Interspeech, 2015

[7] C.-T. Do, M. Evrard, A. Leman, C. d'Alessandro, A. Rilliard, J.-L. Crebouw, Objective evaluation of HMM-based Speech synthesis system using Kullback-Liebler divergence, Interspeech, 2015

[8] L. Blin, O. Boëffard, V. Barreaud, WEB-based listening test system for speech synthesis and speech conversion evaluation, LREC, 2008

[9] O. Boëffard, L. Charonnat, S. Le Maguer, D. Lolive, Towards fully automatic annotation of audio books for TTS, LREC, 2012