URGENT : Phd position in Computer science at IRISA, Team Expression

Title : *Script optimisation for the expressive synthesis of audio books*

Keywords : Expressive speech synthesis ; Optimisation ; Machine learning.

Context : The project targets the automatic generation of audio books using a high quality synthetic voice. The duration needed to listen to a full audio book implies the need for a high quality voice with adapted expressivity. A Text-To-Speech synthesis system produces a speech signal corresponding to an input text. Recently, TTS systems have made great progress in terms of acoustic quality and intelligibility. Despite of that, producing a high quality expressive voice remains a research problem (see [1] and references). The speech quality strongly depends on the system used (parametric or unit selection) and on the speech corpus used. Creating such a speech corpus requires the recording of a specific script with given expressivity types. This recording process being complex and costly, numerous works concern the creation of a script covering the largest possible number of events while minimizing the script duration (see [2, 3, 4] and references).

Work proposal : This problem studied in this Phd is the hybrid generation of audio books. The main idea is to record a minimal part of the audiobook to generate and use the recordings as the source to build a synthetic voice usable to synthesize the rest of the audiobook. More generally speaking, the goal of the Phd work is to propose methods to build and enrich recording scripts automatically in order to build a synthetic voice for which we are able to control the quality. This approach can be formalized as an optimisation problem trying to find the best trade-off between the quality of the final generated speech and the quantity of speech to record.

A first axis of this work concerns the problem of objective and subjective evaluations. In the general context of speech synthesis, evaluating the quality of the generated speech has been the subject of numerous studies (see for instance [5, 6, 7]). What will be the gain brought by the fact of knowing, in advance, the text to synthesize ? What is the impact of having at disposal natural speech realized in the same context ? Moreover, the generated speech will be a mix between natural speech and synthesized speech and its evaluation will require specific methods beyond the scale of the sentence.

A second work axis deals with the automatic generation of recording scripts and the association constraints as well as the definition of a trade-off between the quality of the speech signal and the associated script size. Several questions have already been identified. How textual features influence the final quality ? In particular, which learning methods, guided by objective quality measure, result in the best feature set ?

The last axis focuses on the study how to take into consideration the differences between the expected theoretical result related to the recording script and the real acoustic signal resulting from the recording script. How to detect automatically the variations and adapt dynamically the recording script to keep the final intended acoustic quality ?

Environment : This Phd will be directed jointly by Damien Lolive and Jonathan Chevelu (IRISA-ENSSAT Lannion, University of Rennes1). The financial support is running for 3 years. The student will be located in Lannion and will be member of the Expression team, Media and Interaction Department at IRISA, which studies the expressivity in gesture, speech and text. One of the main research axes of the team is Expressive Speech Synthesis. In this context, the team has a strong experience in the field and also collaborations in complementary fields. The team has a state of the art text-to-speech system and a recording booth to facilitate the generation of useful data to characterise expressivity.

Candidate profile : Anyone with a Master's degree Computer Science can apply. Excellent skills in Computer Science and in Speech and Language Processing are required. Personal Interest in Artificial Intelligence, Machine Learning and Deep Learning is a clear plus.

Phd position starting ASAP, duration : 36 months.

Contacts :
Damien LOLIVE (damien.lolive@irisa.fr) and 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