Conversion expressive de l'identité vocale

Lieu : Laboratoire STMS, IRCAM - Equipe Analyse et Synthèse des Sons, Paris, France
Directeur de thèse : Axel Roebel (Analyse et Synthèses des Sons)
Encadrant : Nicolas Obin (Analyse et Synthèses des Sons)

Projet

La thèse se déroulera dans le cadre du projet ANR TheVoice (2017-2021). Le projet TheVoice s'attaque au design de voix pour l'industrie créative (films, séries, documentaires), un secteur très important pour son potentiel industriel mais extrêmement exigeant en termes de qualité. Le projet se fonde sur le constat qu'au coeur d'une industrie créative massivement numérique, la production des voix demeure artisanale et nécessite l'intervention d'opérateurs humains. La principale originalité de TheVoice sera de se confronter à l'étude de la voix d'acteurs professionnels en situation de jeu réelle, donc naturellement expressive, et de modéliser la palette vocale d'un acteur pour réaliser des outils d'assistance à la création de voix. Le projet créera une rupture des usages par la réalisation et l'industrialisation de nouvelles technologies pour la création de contenus vocaux naturels et expressifs. Le consortium, porté par un acteur majeur du secteur de l'industrie de la création de contenus numérique (Dubbing Brothers), et constitué de laboratoires de recherches reconnus (LIA et Ircam), ambitionne de consolider une position d'excellence de la recherche et des technologies numériques «Made-in-France» et la promotion de la culture française à travers le monde.

Contexte

La conversion d'identité de la voix consiste à convertir les caractéristiques d'une voix «source» pour reproduire celles d'une voix «cible», comprenant la source glottique (prosodie et qualité vocale) et le filtre du conduit vocal (timbre). Le projet TheVoice ambitionne la conversion de l'identité « expressive » de la voix à partir d'enregistrements d'acteurs professionnels en situation de jeu. Les possibilités d'exploitation en production sont nombreuses : permettre le réenregistrement de quelques séquences d'un acteur principal en son absence (en raison de son coût ou de son indisponibilité), pérenniser l'identité vocale d'une voix de doublage devenue indisponible (retraite ou autre), recréer la voix de personnalités historiques disparues, voir à terme de préserver l'identité vocale d'un acteur lors de son doublage.

La conversion est généralement réalisée par conversion spectrale pour le filtre [Stylianou, 1998; Toda, 2007; Sun, 2015] et conversion de la F0 pour la prosodie [Wu 2010]. L'Ircam a

une grande expérience en conversion de l'identité de la voix [Villavicencio 2009; Lanchantin 2010] et a récemment développé un algorithme de conversion [Huber, 2015a; Roebel, 2017], basé sur la sélection et la concaténation d'unités à partir d'exemples [Dutoit, 2007; Wu, 2013], qui a permis une conversion de haute qualité et rendu possible son exploitation en contexte professionnel. L'exploitation d'information phonétique, qui permet de guider la sélection des filtres à appliquer pour la conversion spectrale en fonction des phonèmes prononcés a grandement amélioré la conversion, et rendu possible l'utilisation de bases de voix du tout venant, non-parallèles.

Objectifs

Les travaux de recherches visent à faire passer un nouveau pallier à la conversion de l'identité de la voix pour le traitement de voix expressives d'acteurs en situation de jeu réelle, et s'appuieront sur le système de conversion de voix par concaténation existant à l'Ircam et sur l'expérience de l'Ircam en analyse/modélisation/transformation de la voix.

1) Conversion à partir de balises expressives

La description de la palette vocale sous la forme d'un partitionnement de l'espace acoustique (en fonction de : phonème, durée, hauteur, intensité, voisement etc...) permettra de mieux guider la sélection d'unités de contextes similaires et d'améliorer la conversion. En particulier, la description manuelle ou automatique des intentions de jeu de l'acteur devrait permettre de sélectionner des unités partageant une situation de jeu similaire, et donc d'augmenter la cohérence acoustique entre les voix source et cible pour mieux préserver l'intention de jeu au cours de la conversion.

2) Conversion de la prosodie et de la qualité vocale

Les travaux viseront également à exploiter la description de la prosodie sous la forme de contours mélodiques [Obin 2011, 2014a; Veaux 2011] pour guider la sélection des unités à partir de critères spectraux et prosodiques, corriger la prosodie d'un acteur, et améliorer la reproduction du jeu de l'acteur cible. Enfin, les travaux s'intéresseront à l'utilisation de nouveaux algorithmes des transformation de la source glottique synchrones aux impulsions glottiques [Degottex 2012; Huber 2015a, 2015b] pour la conversion de la qualité vocale (voix soufflée, craquée, etc...), essentielle pour une conversion d'identité expressive et de haute qualité.

Profil du candidat

Le profil recherché doit avoir un Master (ou équivalent) dans un ou plusieurs des domaines suivants : traitement du signal audio, apprentissage machine, informatique et des compétences en langage Python. Une expérience préliminaire en traitement automatique de la parole ou en synthèse / transformation du son sera un avantage pour la candidature.

Candidature

Les candidatures (lettre de motivation et CV) doivent être envoyées avant le **15 septembre 2017** à Nicolas.Obin@ircam.fr et Axel.Roebel@ircam.fr

Références

[Degottex, 2012] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, *Mixed Source Model and its Adapted Vocal Tract Filter Estimate for Voice Transformation and Synthesis*, Speech Communication, vol. 55, no. 2, pp. 278–294, 2012.

[Dutoit, 2007] T. Dutoit, et al. *Towards a voice conversion system based on frame selection*, IEEE International Conference on Audio, Speech, and Language Processing (ICASSP), 2007

[Huber, 2015a] S. Huber, *Voice Conversion by modelling and transformation of extended voice characteristics*, These Université Pierre et Marie Curie (Paris VI), 2015.

[Huber, 2015b] S. Huber, A. Roebel. *On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system*, Interspeech, 2015.

[Lanchantin, 2010] P. Lanchantin and X. Rodet. *Dynamic Model Selection for Spectral Voice Conversion*, Interspeech, 2010.

[Obin, 2011] N. Obin, *MeLos: Analysis and Modelling of Speech Prosody and Speaking Style*, PhD. Thesis, Ircam-Upmc, 2011.

[Sun, 2015] L. Sun, S. Kang, K. Li, and H. Meng. *Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks*, IEEE International Conference on Audio, Speech, and Language Processing (ICASSP), 2015.

[Veaux, 2011] C. Veaux, X. Rodet, *Intonation Conversion from Neutral to Expressive Speech*, Interspeech , Florence, Italy, p. 2765-2768, 2011.

[Villavicencio, 2009] F. Villavicencio, A. Robel, and X. Rodet. *Applying improved spectral modeling for High Quality voice conversion*, IEEE International Conference on Audio, Speech, and Language Processing (ICASSP), 2009.

[Wu, 2010] Wu, Z., Kinnunen, T., Chung, E., and Li, H. (2010). *Text-independent F0 transformations with non-parallel data for voice conversion*. In 11th Annual Conference of the International Speech Communication Association (Interspeech ISCA), pp. 1732–1735, 2010.

[Wu, 2013] Z. Wu, T. Virtanen, T. Kinnunen, Eng Siong Chng. *Exemplar-based Voice Conversion using Non negative Spectrogram deconvolution*. Speech Synthesis Workshop, pp. 201 -206, 2013.