Vizart3D : Retour Articulatoire Visuel pour l'Aide à la Prononciation

Thomas Hueber[1] *Atef Ben-Youssef*[1]
Pierre Badin[1] *Gérard Bailly*[1] *Frédéric Eliséi*[1]
(1) GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France
(prénom.nom)@gipsa-lab.grenoble-inp.fr

RESUME

L'objectif du système Vizart3D est de fournir à un locuteur, en temps réel, et de façon automatique, un retour visuel sur ses propres mouvements articulatoires. Les applications principales de ce système sont l'aide à l'apprentissage des langues étrangères et la rééducation orthophonique (correction phonétique). Le système Vizart3D est basé sur la tête parlante 3D développée au GIPSA-lab, qui laisse apparaître, en plus des lèvres, les articulateurs de la parole normalement cachés (comme la langue). Cette tête parlante est animée automatiquement à partir du signal audio de parole, à l'aide de techniques de conversion de voix et de régression acoustico-articulatoire par GMM.

ABSTRACT

Vizart3D: Visual Articulatory Feedack for Computer-Assisted Pronunciation Training

We describe a system of visual articulatory feedback, which aims to provide any speaker with a real feedback on his/her own articulation. Application areas are computer-assisted pronunciation training (phonetic correction) for second-language learning and speech rehabilitation. This system, named Vizartd3D, is based on the 3D augmented talking head developed at GIPSA-lab, which is able to display all speech articulators including usually hidden ones like the tongue. In our approach, the talking head is animated automatically from the audio speech signal, using GMM-based voice conversion and acoustic-to-articulatory regression.

MOTS-CLES : retour visuel, aide à la prononciation, GMM, temps réel, tête parlante
KEYWORDS : visual feedback, pronunciation training, GMM, real-time, talking head

Plusieurs études semblent montrer que fournir à un locuteur un retour visuel sur ses propres mouvements articulatoires pouvait s'avérer utile pour la rééducation orthophonique et l'apprentissage des langues (Badin, 2010). Ce retour visuel peut notamment s'effectuer via une tête parlante *augmentée*, c'est-à-dire un clone orofacial virtuel qui laisse apparaître l'ensemble des articulateurs, externes (lèvres, mâchoire) comme internes (langue, voile du palais). Dans (Engwall, 2008), Engwall propose un paradigme expérimental du type « magicien d'Oz » pour montrer l'efficacité d'une telle approche (système ARTUR): un phonéticien expert évalue la nature du défaut de prononciation du sujet, et lui fait visualiser le geste articulatoire cible en sélectionnant l'animation adéquate parmi un ensemble d'animations pré calculées. Dans (Ben Youssef, 2011), nous avons proposé un système de retour articulatoire visuel également basé sur l'utilisation d'une tête parlante augmentée. Dans notre approche, la tête parlante augmentée est animée *automatiquement* à partir du

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 17–18,
Grenoble, 4 au 8 juin 2012.

signal audio, par inversion acoustico-articulatoire. Cependant, dans ce système, l'animation de la tête parlante ne peut débuter qu'une fois la phrase entièrement produite (approche par HMM, décodage acoustico-phonétique basé sur l'algorithme de Viterbi). C'est cette limitation que le système Vizart3D tente de lever, en proposant une version temps-réel de notre système de retour articulatoire visuel. Un schéma général du système Vizart3D est présenté à la Figure 1.

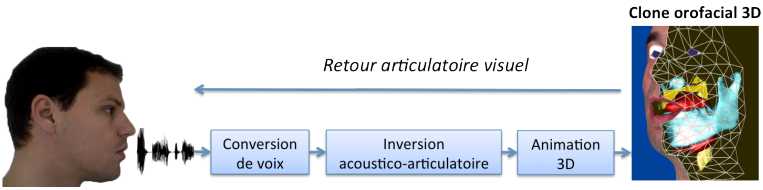

Figure 1 : Schéma général du système Vizart3D

Le système Vizart3D est basé sur la tête parlante augmentée, développée au GIPSA-lab à partir de données IRM, CT et vidéo, acquises sur un locuteur de référence. L'animation de cette tête parlante à partir de la voix d'un locuteur λ s'effectue en 3 étapes (exécutées toutes les 10 ms) : (1) Conversion de voix : l'enveloppe spectrale du locuteur λ, extraite par analyse mel-cepstrale, est transformée en une enveloppe spectrale dite « cible », qui peut être vue comme l'enveloppe qui aurait été obtenue si la même phrase avait été prononcée par le locuteur de référence ; dans notre implémentation, nous utilisons une approche basée sur une régression par GMM (*Gaussian Mixture Model*) – (2) Inversion acoustico-articulatoire : une cible articulatoire est estimée à partir de l'enveloppe spectrale cible (position de la langue (3 points), des lèvres (2 points), et de la mâchoire (1 point)); cette étape d'inversion est également basée sur une modélisation par GMM, à partir d'un corpus de données audio et articulatoires, acquises sur le locuteur de référence par articulographie électromagnétique 2D) – (3), les paramètres de contrôle de la tête parlante sont inférés par régression linéaire, à partir de la cible articulatoire estimée à l'étape 2.

Références

BADIN, P., BEN YOUSSEF, A., BAILLY, G., ELISEI, F., HUEBER, T. (2010) Visual articulatory feedback for phonetic correction in second language learning, Actes de SLATE, P1-10.

ENGWALL, O. (2008) Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes, Actes d'Interspeech, Brisbane, Australie, pp. 2631-2634.

BEN YOUSSEF A., HUEBER T., BADIN P., BAILLY G. (2011) Toward a multi-speaker visual articulatory feedback system, Actes d'Interspeech, Florence, Italie, pp. 489-492.