Projet SASHIMI : clonage de voix et intelligence artificielle

Sujet proposé par Pascal Vallet
pascal.vallet@bordeaux-inp.fr

Thématiques. Traitement du signal, machine learning, intelligence articifielle

Mots-clés. Analyse spectrale, filtrage, conversion de voix, GMM, réseaux de neurones

Contexte. Dans le domaine de la sécurité informatique, et plus particulièrement de l'authentification, les enjeux actuels évoluent en faveur de procédures simplifiées, visant à éviter aux utilisateurs la mémorisation de mots de passe multiples, et leurs changements fréquents. En particulier, les techniques issues de la biométrie, c'est-à-dire la reconnaissance d'individus à partir de caractéristiques physiques/biologiques, sont un axe de développement envisagé depuis une quinzaine d'années pour la sécurisation de l'habitation et des dispositifs personnels. L'identification peut alors se faire par reconnaissance vocale, digitale, faciale, etc., et il convient dès lors de connaître la robustesse de ses systèmes face aux attaques par usurpation. En particulier, la reconnaissance vocale peut être induite en erreur par des voix présentant des caractéristiques spectrales proches.

Clonage vocal. La *conversion de voix* (ou clonage vocal) regroupe un ensemble de techniques de traitement de signal permettant de modifier les caractéristiques de voix d'un individu « source » de telle manière à se rapprocher des caractéristiques de voix d'un individu « cible », tout en préservant le contenu linguistique du son. Les applications potentielles de la conversion de voix concernent également d'autres domaines comme la synthèse de voix pour des personnes en situation de handicap, les systèmes « text to speech », les traducteurs vocaux instantanés, etc. Une approche possible pour le clonage de voix consiste (1) à extraire des paramètres spectraux d'un échantillon de voix source et cible prononçant les mêmes mots, puis (2) à « apprendre » une fonction permettant de convertir les paramètres source en paramètres cible. [3, 2]

Projet. L'objectif du projet SASHIMI (Spoofing AttackS of Human voIce using Machine learnIng) est de réaliser un logiciel fonctionnel de conversion de voix, ne nécessitant qu'un court extrait de voix cible (inférieur à 5 secondes), utilisant l'approche décrite ci-dessus. Si le temps le permet, des tests d'usurpation pourront être effectués sur des logiciels de reconnaissance vocale (par exemple [1]). Le langage utilisé sera de préférence Python.

1. Etat de l'art et simulation d'algorithmes d'extraction de caractéristiques spectrales de sons vocaux (coefficients MFCC, LP, etc.) ;
2. Etat de l'art et simulation de méthodes d'apprentissage pour la régression non-linéaire (Gaussian Mixture Models, réseaux de neurones, etc.) ;
3. Développement d'une IHM (capture d'un signal vocal + conversion) ;

Références

[1] `https://azure.microsoft.com/fr-fr/services/cognitive-services/speaker-recognition/`.

[2] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A.W Black, and K. Prahallad. Voice conversion using artificial neural networks. In *ICASSP*, pages 3893–3896. IEEE, 2009.

[3] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2) :131–142, 1998.