Vers la transformation de la parole oesophagienne en voix laryngée à l'aide de techniques de conversion vocale

Imen Ben Othmane, Joseph Di Martino, Kaïs Ouni

Vers la transformation de la parole œsophagienne en voix laryngée à l'aide de techniques de conversion vocale

Imen Ben Othmane [1], Joseph Di Martino [2], Kaïs Ouni [1]
imen.benothmen@hotmail.fr , joseph.di-martino@loria.fr , kais.ouni@enicarthage.rnu.tn
(1) Unité de Recherche Systèmes Mécatroniques et Signaux, SMS, Ecole Nationale d'Ingénieurs de Carthage, Tunisie
(2) Laboratoire Lorrain de Recherche en Informatique et ses Applications, LORIA, Vandœuvre-lès-Nancy, France

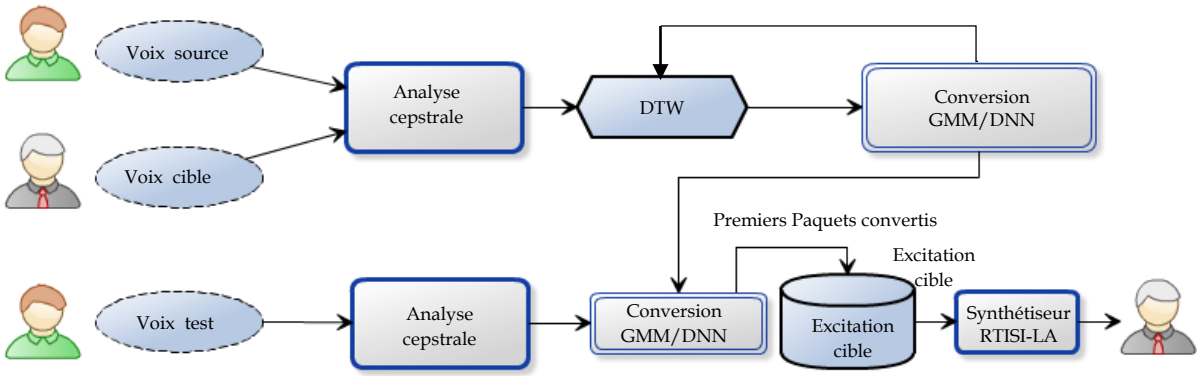
Ce travail concerne le développement d'un système de conversion de voix œsophagienne dans le but est de rendre plus intelligible celle-ci. La conversion de voix est une technique de transformation d'un signal de parole d'un locuteur source, de manière à ce qu'il semble, à l'écoute, être prononcé par un locuteur cible.

Etant donnée la spécificité de la voix œsophagienne, nous proposons dans cette étude d'appliquer une nouvelle technique de conversion vocale en tenant compte de la particularité de l'appareil vocal des patients qui ont subi une ablation de larynx. En effet, l'ablation des cordes vocales perturbe profondément le signal glottique et par conséquent la voix œsophagienne acquise par le patient laryngectomisé est difficile à comprendre, rauque et faible en intensité.

Dans la littérature, plusieurs techniques de conversion des voix ont été proposées, parmi lesquelles, la technique du codage linéaire prédictif pour la conversion vocale [1] et la régression linéaire multi-variée [2] qui vise à réduire la discontinuité et la distorsion spectrale.

D'autres chercheurs ont proposé de transformer les caractéristiques acoustiques de la voix œsophagienne afin de l'améliorer, par exemple, à l'aide de lissage [3] ou par filtrage en peigne [4].

Dans ce travail, nous proposons un système de conversion basée sur les étapes suivantes :



Deux corpus parallèles ont été utilisés : un concernant la voix œsophagienne comme source et l'autre concernant la voix normale comme cible préalablement alignées à l'aide d'un algorithme de programmation dynamique DTW [7]. Par la suite, un module d'apprentissage a été appliqué sur les premiers paquets cepstraux des voix source et cible afin de déterminer une fonction de transformation.

Cette fonction a été établie par deux approches différentes : la première, par réseaux de neurones profonds [5] et la deuxième, à l'aide de modèle de mélange gaussien [6]. Ces deux approches ont permis d'obtenir des résultats équivalents.

La fonction établie permet de convertir les premiers paquets cepstraux.

Les paquets convertis ont été utilisés ensuite pour estimer les coefficients cepstraux relatifs au signal d'excitation glottique avec une recherche dans l'espace d'apprentissage cible préalablement codé sous la forme d'un arbre binaire. Pour préserver les caractéristiques du conduit vocal du locuteur source les premiers paquets cepstraux n'ont pas été modifiés au niveau de la resynthèse.

Après resynthèse par le synthétiseur RTISI-LA [8], une voix « laryngée » plus naturelle que l'originale a été obtenue, avec une reconstruction effective des informations prosodiques. Et ce, tout en conservant, et c'est le point fort de notre étude, les caractéristiques du conduit vocal inhérentes au locuteur source.

Mots clés :
Voix œsophagienne, Conversion de Voix, Coefficients Cepstraux, Modèle de Mélange Gaussien, DNN, Excitation Glottique, Arbre de Recherche.

Bibliographie

[1] Y. Qi, Replacing tracheoesophageal voicing sources using LPC synthesis. Journal of Acoustical Society of America, 88, pp. 1228–1235, 1990.

[2] N. Bi and Y. Qi, Application of speech conversion to alaryngeal speech enhancement. IEEE Transactions on Acoustics, Speech, Signal Processing, vol. 5, no. 2, pp. 97–105, 1997.

[3] K. Matui, N. Hara, N. Kobayashi, and H. Hirose, Enhancement of esophageal speech using formant synthesis. Proc. ICASSP, pp. 1831–1834, Phoenix, Arizona, 1999.

[4] A. Hisada, and H. Sawada, Real-time clarification of oesophageal speech using a comb filter. International Conference on Disability, Virtual Reality and Associated Technologies, pp. 39–46, 2002.

[5] D. Yu, L. Deng, Deep learning and its applications to signal and information processing. IEEE Signal Processing Magazine, pp. 145–154, 2011.

[6] Y. Stylianou, O. Cappé and Eric Moulines, Continous probabilistic transform for voice conversion. IEEE transactions on Speech and Audio Processing, Vol. 6, No. 2, pp. 131-142, March 1998.

[7] H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, No. 1, pp. 43- 49, 1978.

[8] X. Zhu, G. T. Beauregard and L. L. Wyse, Real-time signal estimation from modified short-time Fourier transform magnitude spectra. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 5, pp. 1645-1653, July 2007.