Une nouvelle méthodologie prédictive fondée sur un modèle séquence à séquence utilisé pour la transformation de la parole œsophagienne en voix laryngée

Kadria Ezzine[1], Imen BEN Othmane[2], Joseph Di Martino[3], Mondher Frikha[1]

kadria.ezzine@gmail.com, imen.benothmen01@gmail.com, joseph.di-martino@loria.fr, mondher.frikha@enetcom.usf.tn

(1) Unité de Recherche en Advanced Technologies For Image And Signal Processing, ATISP, ENET'COM, Université de Sfax, Tunisie

(2) Laboratoire de Recherche Electricité Intelligente & TIC, Ecole Nationale d'Ingénieurs de Carthage, ENICarthage, Université de Carthage, Tunisie

(3) Laboratoire Lorrain de Recherche en Informatique et ses Applications, LORIA, Vandœuvre-lès-Nancy, France

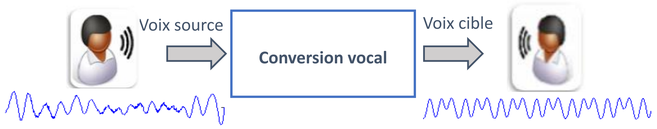
Résumé

Dans ce travail, nous proposons une méthode de rehaussement de la parole œsophagienne basée sur une technique séquence à séquence (SEQ2SEQ) combinée à un mécanisme d'attention auditive. Le point fort de la méthode proposée est qu'elle ne nécessite pas d'alignement temporel durant la phase d'apprentissage ce qui permet de réduire considérablement le temps de calcul de celle-ci.

Problématique et Objectifs

❖ Définition

La conversion de la parole œsophagienne (ES) en voix plus naturelle est un moyen efficace pour améliorer la qualité auditive et l'intelligibilité de cette parole pathologique.



❖ Caractéristiques de la parole œsophagienne:

- Des bruits spécifiques qui ressemblent à des éructations
- Une fréquence fondamentale chaotique
- Une faible intensité
- Un timbre généralement dur

⍰ Ces variations instables d'intensité et de F0 sont responsables de la mauvaise qualité audio de l'ES

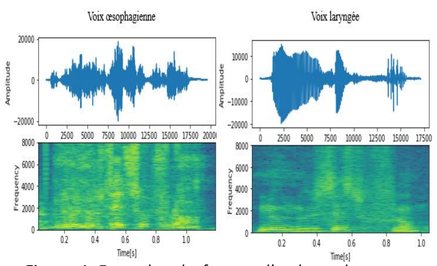


Figure. 1: Exemples de formes d'onde, et de spectrogrammes de la voix laryngée et œsophagienne.

❖ Objectifs

- ✓ Améliorer la qualité auditive de la parole œsophagienne en termes de naturalité et d'intelligibilité
- ✓ Rendre la communication des personnes laryngectomisées plus aisée

Domaine d'application

- Text-To-Speech (TTS) synthèse
- Personnalisation des appareils parlants
- Doublage et traduction de films
- Pathologie vocale (conception d'aides à la parole)

Contributions

Notre contribution consiste en une méthode de rehaussement de la parole œsophagienne basée sur une technique prédictive séquence à séquence (SEQ2SEQ) combinée à un mécanisme d'attention auditive.
La méthodologie SEQ2SEQ proposée consiste en une architecture codeur-décodeur composée d'un modèle d'apprentissage et d'une technique de conversion vocale.

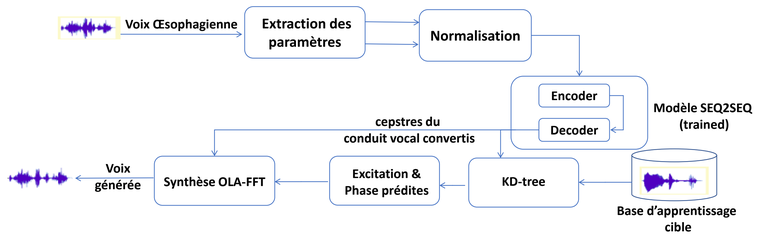


Figure. 2. Diagramme de la phase de conversion de la méthode proposée

1) Tout d'abord, un réseau Bidirectionnel LSTM (Long Short Time Memory) est utilisé comme encodeur qui traite chaque séquence d'entrée.
2) Ensuite, le décodeur avec son mécanisme d'attention vise à améliorer la qualité et la précision des sorties de l'encodeur.
3) Les coefficients de l'excitation et de la phase sont estimés à partir de l'espace d'apprentissage cible structuré sous la forme d'un KD-tree.
4) Enfin, la méthode addition-recouvrement OLA-FFT a été appliquée au niveau de la resynthèse.

Dans nos expériences, nous avons adopté deux méthodes référence de comparaison qui sont:

- ➢ Les réseaux de neurones profonds DNN
- ➢ Les réseaux LSTM sans mécanisme d'attention.

Trois corpus parallèle ont été utilisés pour évaluer notre système de rehaussement de la voix œsophagienne:

- Locuteurs sources (voix œsophagienne): "PC" et "MH"
- Locuteur cible (voix laryngée): "AL"

Résultats

Trois mesures objectives ont été utilisés pour évaluer la qualité de la parole;

- L'évaluation perceptuelle de la qualité de la parole (PESQ),
- L'intelligibilité objective à court terme (STOI)
- La distorsion Mel-Cepstral (MCD)

Un test perceptif de type MOS a été utilisé pour évaluer les résultats de manière subjective.

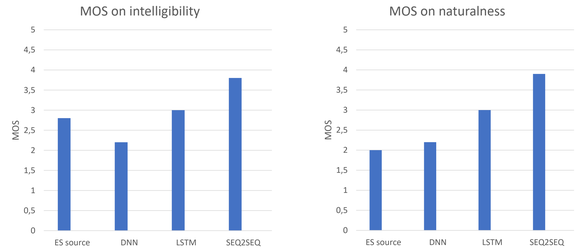


Figure. 3 : Résultats expérimentaux des tests subjectifs MOS

Conclusion

Nous avons décrit une nouvelle méthodologie prédictive fondée sur un modèle SEQ2SEQ combiné à un mécanisme d'attention auditive utilisé pour la transformation de la parole œsophagienne en voix laryngée. Les résultats expérimentaux démontrent que notre méthode se comporte mieux et atteint de meilleures performances même dans certains cas difficiles. En effet elle surpasse les méthodes conventionnelles en termes de rendu naturel et d'intelligibilité.

Références

BEN OTHMANE, I., Di MARTINO, J., & OUNI, K. (2019). Enhancement of esophageal speech obtained by a voice conversion technique using time dilated Fourier cepstra. International Journal of Speech Technology, 22(1), (pp. 99-110)

BEN OTHMANE, I., Di MARTINO, J., & OUNI, K. (2017). Vers la transformation de la parole œsophagienne en voix laryngée à l'aide de techniques de conversion vocale, 7ème Journées de Phonétique Clinique – JPC 7, Paris.

TODA, T., NAKAMURA, K., SARUWATARI, H., & SHIKANO, K. (2014). A laryngeal speech enhancement based on one-to-many eigen voice conversion. IEEE/ACM transactions on audio, speech, and language processing, 22(1-2), (pp.172-183)

SUN, L., KANG, S., Li, K., & MENG, H. (2015, April) .Voice conversion using deep bidirectional long short term memory based recurrent neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), (pp. 4869-4873). IEEE.