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

Kadria Ezzine, Imen Ben Othmane, Joseph Di Martino, Mondher Frikha

Titre : 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]

[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

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. Par rapport à la voix laryngée, ES se caractérise par un bruit spécifique qui ressemble à des éructations avec une fréquence fondamentale chaotique, une faible intensité et un timbre généralement dur.

Récemment, plusieurs approches basées sur la conversion vocale ont été proposées [1], [2], [3], [4] qui visent à rapprocher les caractéristiques de la

voix œsophagienne (source) de celles de la voix laryngée (cible).

Dans ce travail, nous proposons une méthode de rehaussement de la parole œsophagienne basée sur une technique séquence à séquence (SEQ2SEQ) [5], [6] 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.

Premièrement, un réseau BiLSTM (Bidirectionnel Long Short Time Memory) [7] est utilisé comme encodeur qui traite chaque séquence d'entrée dans un espace de caractéristiques de grande dimension puis l'encode dans un vecteur de contexte de longueur fixe. Ensuite, le décodeur avec son mécanisme d'attention vise à améliorer la qualité et la précision des sorties de l'encodeur. Enfin, pour préserver l'identité du locuteur cible, 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 arbre de recherche binaire en interrogeant celui-ci par les coefficients du conduit vocal précédemment prédits par le modèle SEQ2SEQ. Au niveau de la resynthèse, nous avons appliqué la méthode addition-recouvrement à court terme OLA-FFT.

Dans nos expériences, nous avons adopté deux méthodes référence de comparaison qui sont DNN [8] et LSTM [9]. Trois corpus parallèle ont été utilisés pour évaluer notre système de rehaussement de la voix œsophagienne. Trois mesures, y compris l'évaluation perceptuelle de la

qualité de la parole (PESQ), l'intelligibilité objective à court terme (STOI) et la distorsion Mel-Cepstral (MCD) ont été utilisées pour mesurer objectivement la qualité de la parole. De plus un test MOS a été utilisé pour évaluer les résultats de manière subjective.

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[1].

Références bibliographiques

[1] TODA, T., NAKAMURA, K., SARUWATARI, H., & SHIKANO, K. (2014). Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. IEEE/ACM transactions on audio, speech, and language processing, 22(1-2), 172-183.

[2] LACHHAB, O., Di MARTINO, J., ELHAJ, E. I., & HAMMOUCH, A. (2015). A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion. SpringerPlus, 4(1), 644.

[3] 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), 99-110.

[4] BEN OTHMANE, I., Di MARTINO, J., & OUNI, K. (2017). Vers la transformation de la parole

1 http://techtech-solution.com/Kadria/seq2seq-ES-enhancement.html

oesophagienne en voix laryngée à l'aide de techniques de conversion vocale, *7ème Journées de Phonétique Clinique* – JPC 7, Paris.

[5] SUTSKEVER, I., VINYALS, O., & LE, Q. V. (2014). "Sequence to sequence learning with neural networks," Neural Information Processing Systems, pp. 3104–3112.

[6] CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., & BENGIO, Y. (2014). "Learning phrase representations using RNN encoder–decoder for statistical machine translation," Empirical Methods in Natural Language Processing, pp. 1724–1734.

[7] 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.

[8] WU, Z., WATTS, O., & KING, S. (2016, September). Merlin: An Open Source Neural Network Speech Synthesis System. In SSW (pp. 202-207).

[9] CHEN, J., & WANG, D. (2017). Long short-term memory for speaker generalization in supervised speech separation. The Journal of the Acoustical Society of America, 141(6), 4705-4714.