Titre: Rehaussement de la parole œsophagienne par une technique de conversion de voix fondée sur l'estimation de multiples réseaux de neurones profonds et de fonctions de conversion spectrale

Mouad LARIBIA[1]

Imen BEN OTHMANE[2]

Joseph DI MARTINO[3]

Elhassane IBN ELHAJ[1]

[1]Laboratoire de Recherche Systèmes de Télécommunications Réseaux et Services,
Institut National des Postes et Télécommunications, INPT,
Rabat, Maroc
[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

Pour faire face aux dégradations de qualité et d'intelligibilité de la voix œsophagienne VO, nous proposons dans ce travail un système de rehaussement en utilisant la conversion de voix CV [1], combinant l'approche statistique de classification avec un modèle de mélange gaussien GMM et les réseaux de neurones profonds DNN pour transformer les vecteurs cepstraux de conduit vocal source vers ceux cible en tenant compte de la particularité de l'appareil vocal pathologique.

Dans la littérature, plusieurs techniques de rehaussement de la VO ont été proposées, parmi lesquelles, la technique de codage prédictif linéaire [2], la synthèse par formants [3], le filtrage en peigne [4], le filtrage de Kalman avec stabilisation de pôles [5], la CV basée sur l'analyse statistique [6] et la CV basée sur les postériorgrammes phonétiques (PPG) sans utilisation de corpus

parallèles [7]. Récemment, la CV basée sur l'utilisation des DNN ont permis d'obtenir de bonnes performances [8] [9] [10].

Les étapes de notre système de conversion sont les suivantes :

- Deux corpus parallèles ont été créés : un concernant la VO comme source et l'autre concernant la VN comme cible. Après extraction et séparation des coefficients cepstraux d'excitation et ceux relatifs au conduit vocal, une classification par GMM est utilisée pour regrouper conjointement les premiers paquets cepstraux source et cible, préalablement alignés par l'algorithme de programmation dynamique DTW [11].
- Ensuite, un apprentissage par DNN est réalisé afin de déterminer une fonction de transformation non-linéaire pour chaque classe.
- Les vecteurs cepstraux convertis sont utilisés d'abord pour estimer l'excitation et la phase à partir de l'espace d'apprentissage cible structuré sous la forme d'un arbre binaire à l'aide de l'algorithme KD-tree [12]. Ensuite ceux-ci sont utilisés pour déterminer des fonctions de conversion spectrale pour chaque classe en calculant la moyenne pondérée des rapports entre les spectres d'amplitude cible et ceux source appariés afin de réduire les irrégularités observées au niveau formantique. Quant à l'énergie, celle-ci est prédite séparément à l'aide d'un DNN.

L'évaluation objective de notre méthodologie en utilisant le rapport signal sur erreur (SER) indique une amélioration de qualité des signaux transformés. Et, de plus, il apparaît clairement, suite à des tests subjectifs informels, que l'approche proposée produit un son de meilleure qualité audio, par rapport au son original, avec une reconstruction effective des informations prosodiques mais avec une légère perte d'intelligibilité[1].

Références bibliographiques

[1] MOHAMMADI, S. H., and KAIN, A., An overview of voice conversion systems, *Speech Communication*, Vol. *88*, 2017, 65-82.

[2] SIRICHOKSWAD, R., BOONPRAMUK, P., KASEMKOSIN, N., CHANYAGORN, P., CHAROENSUK, W., SZU, H. H., Improvement

[1] http://esophageal-speech-enhancement.laribia.42web.io

of esophageal speech using LPC and LF model, in INTERNATIONAL CONFERENCE ON BIOMEDICAL AND PHARMACEUTICAL ENGINEERING, **405-408**, 2006.

[3] KENJI, M., and NORIYO, H., Enhancement of esophageal speech using formant synthesis, in INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP), **81-85**, 1999.

[4] HISADA, A., and SAWADA, H., Real-time clarification of esophageal speech using a comb filter, in INTERNATIONAL CONFERENCE ON DISABILITY, VIRTUAL REALITY AND ASSOCIATED TECHNOLOGIES, **39-46**, 2002.

[5] GARCIA, B., and MENDEZ, A., Esophageal speech enhancement using poles stabilization and Kalman filtering, in INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP), **1597-1600**, 2008.

[6] DOI, H., NAKAMURA, K., TODA, T., SARUWATARI, H., and SHIKANO, K., Statistical approach to enhancing esophageal speech based on Gaussian mixture models, in INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP), **4250-4253**, 2010.

[7] SERRANO, L., RAMAN, S., TAVAREZ, D., NAVAS, E., HERNAEZ, I. (2019) Parallel vs non-parallel voice conversion for esophageal speech, PROC. INTERSPEECH, **4549-4553**, 2019.

[8] BEN OTHMANE, I., DI MARTINO, J., and OUNI, K., Vers la transformation de la parole oesophagienne en voix laryngée à l'aide de techniques de conversion vocale, *7ème Journées de Phonétique Clinique - JPC 7*, 2017.

[9] BEN OTHMANE, I., DI MARTINO, J., and OUNI, K, Enhancement of esophageal speech using voice conversion techniques, in INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE, SIGNAL AND SPEECH PROCESSING (ICNLSSP).

[10] BEN OTHMANE, I., DI MARTINO, J., and OUNI, K., Enhancement of esophageal speech using statistical and neuromimetic voice conversion techniques, *Journal of International Science And General Applications ISGA*, Vol. **1/1**, 2018, 10.

[11] SAKOE, H., AND CHIBA, S., Dynamic programming algorithm optimization for spoken word recognition, in *IEEE Transactions On Acoustics, Speech And Signal Processing*, Vol. **ASSP-26/1**, 1978, 43-49.

[12] ARYA, S., *Nearest Neighbor Searching And Applications*, Thèse de doctorat, University of Maryland at College Park (United States), 1996.