68. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)

17.09. - 21.09.23, Heilbronn

Meeting Abstract

Uncovering time and frequency characteristics of dysarthric speech by means of deep learning

Mayra Elwes - Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany

Antoine Serrurier - Clinic for Phoniatrics, Pedaudiology & Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Aachen, Germany

Ekaterina Kutafina - Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany

Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 68. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS). Heilbronn, 17.-21.09.2023. Düsseldorf: German Medical Science GMS Publishing House; 2023. DocAbstr. 42

doi: 10.3205/23gmds097 , urn:nbn:de:0183-23gmds0973

Published: September 15, 2023

Text

Introduction: Recent studies have been focused on the extraction and interpretation of dysarthria-related features derived from audio recordings using deep learning methods [1], [2]. In [1], features were extracted from the latent space of an encoder-decoder-network. In [2], a filterbank applied directly to the speech waveform is learned jointly with a dysarthria classifier and later interpreted to identify dysarthric characteristics. However, the time and frequency characteristics of dysarthric speech still remain to be identified and interpreted.

In this exploratory study, a Convolutional Neural Network (CNN) is trained to classify speech samples as dysarthric or healthy. In the second step, test samples are provided as input to the network, and the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm [3] is applied to it. This provides heatmaps that reveal the relevant time and frequency areas of the spectrogram for the classification of the samples as healthy or dysarthric.

Methods: Our method was tested on the TORGO Database [4] containing English speech samples, split by task, of 8 dysarthric and 7 control subjects. All audio samples were padded to 2.5 s and converted to mel-spectrogram. Then the samples were fed to a ResNet50 architecture taken from [5] that classified them as either dysarthric or control. For evaluation, a leave-one-subject-out scheme was considered. The Grad-CAM algorithm was applied to the last convolutional layer for all test samples. The heatmaps were then sorted by subject and correctness of classification and assessed manually. Averaged over all subjects, a classification accuracy of 73,1% was achieved.

Results: The heatmaps of the correctly classified control samples tend to highlight frequencies between 0 and 1024 Hz on the pre-voicing phase. In contrast, most of the correctly classified dysarthric samples tend to highlight frequencies between 1024 Hz and 4096 Hz on the voicing phase. The model seems therefore to detect dysarthria on the voicing phase in the medium frequency range. A lack of dysarthria seems, on the contrary, to be associated with a lack of dysarthria related characteristics in the middle frequencies or a presence of healthy characteristics in the pre-voicing phase of the lower frequencies.

Discussion: Our results generally agree with previous findings [2], which emphasize the importance of the frequencies around 2000 and 6500Hz. Additional information gained by using the Grad-CAM algorithm was the timing information, which highlighted the role played by the voicing and pre-voicing phases.

This exploratory study calls for further investigations. The current study considers a wide range of input speech samples. Considering similar speech tasks for all subjects may ensure comparability and possibilities to uncover dysarthria characteristics more robustly. Finding a way to improve, analyzing, and summarizing the visual explanations in an automated way is crucial. Finally, since the Grad-CAM algorithm results are quite coarse, other methods such as Shapley values or local surrogate models could provide interesting complementary outcomes.

Conclusion: We demonstrated a proof of concept to uncover audio time and frequency features of dysarthria, by applying the Grad-CAM algorithm to a dysarthria classifying CNN.

The authors declare that they have no competing interests.

The authors declare that an ethics committee vote is not required.

References

1. Korzekwa D, Barra-Chicote R, Kostek B, Drugman T, Lajszczak M. Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech. In: Proc. Interspeech 2019; Graz, Austria; 2019 Sep 15-19. 2019. p. 3890-3894. DOI: 10.21437/Interspeech.2019-1206
2. Millet J, Zeghidour N. Learning to detect dysarthria from raw speech. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 12-17 May; Brighton, United Kingdom. IEEE; 2019. p. 5831-5835. DOI: 10.1109/ICASSP.2019.8682324
3. Selvaraju RR, Cogswell M, Das A, Vendantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: International Conference on Computer Vision (ICCV); 2017 Oct 22-29 Oct; Venice, Italy. IEEE; 2017. p. 618–626. DOI: 10.1109/iccv.2017.74
4. Rudzicz F, Namasivayam AK, Wolff T. The Torgo database of acoustic and articulatory speech from speakers with dysarthria. Lang Resour Eval. 2012;46(4):523-541. DOI: 10.1007/s10579-011-9145-0
5. Sekhar SM, Kashyap G, Bhansali A, Abishek A, Singh K. Dysarthric-speech detection using transfer learning with convolutional neural networks. ICT Express. 2022;8(1):61-64. DOI: 10.1016/j.icte.2021.07.004