Vocal Tract Detection using Yolo v7

Haroldo G.B. Filho[1], Luã C. Saunders[1]

[1]Universidade Federal do Maranhão (UFMA)
Av. dos Portugueses, 1966 - Vila Bacanga, 65080-805– São Luís – MA – Brazil

haroldo.gbf@ufma.br, saunders.lua@discente.ufma.br

Abstract. *Speech is a means by which an individual can interact with the society in which he or she is inserted, but its weakness can lead to social exclusion and pathological stigma, so it is necessary to understand the speech process in a systematized and specified way. In this paper, this understanding is proposed through the detection of vocal tract using the YOLO v7 framework in a vocal tract represented by a magnetic resonance image, to observe the accuracy and performance of the model in order to help the specialist recognize a pattern in the individual's speech and consequently the maturity and evolution of the applied framework..*

Resumo. *A fala é um meio pelo qual o indivíduo pode interagir com a sociedade em que está inserido, mas a sua fragilidade pode levar à exclusão social e estigmas, por isso é necessário compreender o processo da fala de forma sistematizada e especificada. Neste trabalho, propõe-se a detecção do trato vocal utilizando o framework YOLO v7 para analise da viabilidade do processo da fala representado por uma imagem de ressonância magnética, para observar a precisão e o desempenho do modelo, auxiliando o especialista a reconhecer um padrão na fala do indivíduo e consequentemente a maturidade e evolução do framework.*

1. Introduction

Speech is characterized as a primordial attribute in such development, to establish interrelationships between sender and receiver in a social environment through certain linguistic pattern (manipulation of the verbal meaning system), composed of an auditory and phonoarticulatory structure, responsible for perception acoustics, vocal production and phonemic articulation, that is, the act of producing speech can be defined as the modularization of oral structures and analysis of organic effects (acoustic-articulatory and tactile-kinetic) through skeletal muscles of the vocal tract – lungs, oropharynx andrigid structures; as well as the articulatory synthesis of the vocal tract (responsible for the vibration of the vocal cords and consequent verbal “intonation”), which (Bresch and Narayanan 2009) attests to being composed of eight anatomical phonating or speech articulating components, which are: larynx, epiglottis, tongue, lips, pharyngeal wall, glottis, soft palate and hard palate.

Knowledge regarding the position and movement of these articulators is essential for studies and investigation into speech production and articulatory process. The vocal tract is a dynamic system and physiological structure that integrates speech production and "real world" perception for neuromotor instructions, aiming for a systematic and contextualized flow for reasonably regular and accurate communication.

2. Dataset of Vocal Tract

The Dataset originally having 1210 frames in DICOM (Digital Imaging and Communications in Medicine) format extracted from MRI scans, after the pronunciation of 15 words in the Portuguese language for the observation of phonetic variability [Hamer 2011], being them: "palhaço", "bolsa", "cadeira", "galinha", "vassoura", "alho", "xícara", "mesa", "navio", "livro", "sapo", "tambor, "sapato", "balde, "faca" and "fogão". To train the model with this dataset, we used 50 epochs with a resolution of 640 px on the Google Colab GPU.

We partitioned the dataset to 515 images and 70% (training), 20% (validation) and 10% (test) exclusively on frames that had movements, cutting out the pauses or "rest" of the patient between one word and the next.

For implementation of object detection, each grid cell would predict bounding boxes along with the dimensions and confidence scores (see figure 1). The confidence score was indicative of the absence or presence of an object within the bounding box.

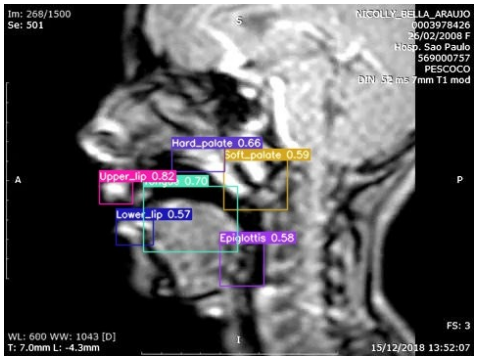

Figure 1. Detection objects in the vocal tract

Once the dataset is formalized, it is necessary to define the labeling. Labeling is nothing more than marking the bounding boxes in the images, that is, it is manually identifying the objects of interest within an image so that, in this way, it is possible to carry out the training. In the context of this work, the objects of interest that were "marked" are five class in vocal tract: lower lip, upper lip, hard palate, soft palate and epiglottis using annotation software (Dwyer et al. 2022).

3. Detection of Speech Articulators using Yolo v7

The proposed methodology (Figure 2), aims to evaluate the implementation of an architecture in its most recent version, YOLO v7, to observe its best performing backbones in the development of a model that detects with high precision the upper and lower lips of an individual, during speech, in the same way, segment the objects of

interest for effective pattern recognition, thus assisting in the pre-diagnosis of speech pathologies with high accuracy.

YOLOv7 (Wang et. al 2022) was published in ArXiv in July 2022 by the same authors of YOLOv4 and YOLOR. At the time, it surpassed all known object detectors in speed and accuracy in the range of 5 FPS to 160 FPS. Like YOLOv4, it was trained using only the MS COCO dataset without pre-trained backbones. YOLOv7 proposed a couple of architecture changes and a series of bag-of-freebies, which increased the accuracy without affecting the inference speed, only the training time. Figure 1 shows the detailed architecture of YOLOv7 for detecting objects in the vocal tract.

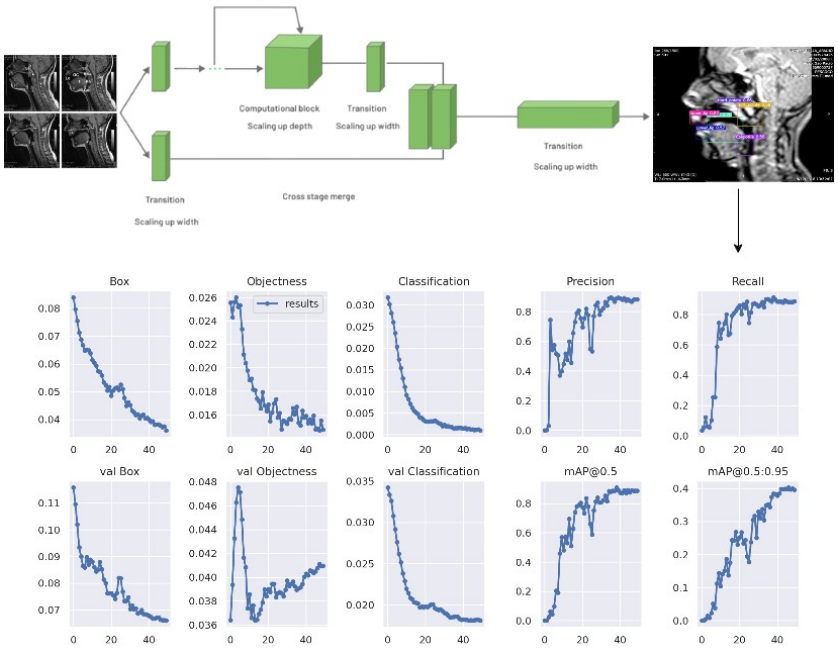

Figure 2. Yolo v7 architecture for detecting and evaluating a vocal tract

The architecture changes of YOLOv7 are:

• Extended efficient layer aggregation network (E-ELAN). ELAN is a strategy that allows a deep model to learn and converge more efficiently by controlling the shortest longest gradient path. YOLOv7 proposed E-ELAN that works for models with unlimited stacked computational blocks. E-ELAN combines the features of different groups by shuffling and merging cardinality to enhance the network's learning without destroying the original gradient path.

• Model scaling for concatenation-based models. Scaling generates models of different sizes by adjusting some model attributes. The architecture of YOLOv7 is a concatenation-based architecture in which standard scaling techniques, such as depth scaling, cause a ratio change between the input channel and the output channel of a transition layer which, in turn, leads to a decrease in the hardware usage of the model. YOLOv7 proposed a new strategy for scaling concatenation-based models in which the depth and width of the block are scaled with the same factor to maintain the optimal structure of the model.

The bag-of-freebies used in YOLOv7 include:

• Planned re-parameterized convolution. Like YOLOv6, the architecture of YOLOv7 is also inspired by re-parameterized convolutions (RepConv) (Ding et. al 2021). However, they found that the identity connection in RepConv destroys the residual in ResNet (He et. al 2016) and the concatenation in DenseNet (Huang et. al 2017). For this reason, they removed the identity connection and called it RepConvN.

4. Results and Conclusion

The proposed methodology for detecting objects in a vocal tract aimed to analyze the feasibility of understanding movement and variability through the implementation of a deep learning based architecture called Yolo V7, which showed an average accuracy of 83%, demonstrated through the accuracy and recall metrics.

Demonstrating the relevance of understanding the speech process through the use of MRI, without exhaustive tests requiring “audio files” from the patient, considered a non-intrusive and highly scalable technique for producing image samples, thus helping the Speech therapist or Otorhinolaryngologist, frame by frame, to locate a speech bottleneck or specific weakness.

References

Bresch E. and Narayanan S., "Region Segmentation in the Frequency Domain Applied to Upper Airway Real-Time Magnetic Resonance Images" in IEEE Transactions on Medical Imaging, vol. 28, no. 3, pp. 323-338, March 2009, doi: 10.1109/TMI.2008.928920.

C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” arXiv preprint arXiv:2207.02696, 2022

Dwyer, B., Nelson, J. (2022), Solawetz, J., et. al. Roboflow (Version 1.0) [Software]. Available from https://roboflow.com.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708, 2017

MARCELINO, F. C; HAMER, B. L. Intervenção fonoaudiológica nos atrasos de linguagem: uma visão integral. In: LOPES-HERRERA, S. A; MAXIMINO, L. P. Fonoaudiologia:Intervenções e alterações da linguagem oral infantil. 1a edição. RibeirãoPreto: Editora Novo Conceito, 2011

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13733–13742, 2021.