Offer #2024-08361

Internship - Dynamic three-dimensional reconstruction of the vocal tract during speech production

Contract type : Internship agreement

Level of qualifications required : Graduate degree or equivalent

Fonction : Internship Research

Context

Preamble:

This subject is part of the PhD tracks programme run by the Inria centre at the University of Lorraine and the Strasbourg site. The aim of this programme is to attract and support promising and motivated students currently enrolled in a Master 2 course towards a PhD by offering four years of combined funding covering a Master 2 internship + thesis. The Master 2 internship, lasting 5 to 6 months, will be paid at €4.35 /hour (plus or minus €670 per month). Candidates admitted to the programme will present the progress of their work to a jury in May 2025, which will validate their entry into the PhD programme (the PhD track should be discontinued in exceptional cases).
The programme, how to apply and the timetable are described in the PhD track section of the https://www.inria.fr/fr/centre-inria-universite-lorraine website.

Although very good text-to-speech systems now exist, understanding the human process of speech production remains an area of research in which there are several unresolved challenges. The first is the control of the temporal evolution of the three-dimensional geometric shape of the vocal tract, which defines the resonance cavities, and consequently the acoustic properties of speech sounds.

We have recently proposed a deep learning approach to synthesize the two-dimensional shape of the vocal tract from a sequence of phonemes to be articulated[1]. The system operates in the medio-sagittal plane and is trained on a large database of real-time MRI data acquired at the IADI laboratory in Nancy for a French-speaking speaker. \To the best of our knowledge, this is the most advanced system, as it provides the contour of all the articulators of speech, i.e. tongue, lips, larynx, etc. Its main limitation is that it only provides two-dimensional information, as the database only contains images in the medio-sagittal plane, due to technological constraints that cannot be easily overcome.

The aim of this project is to add the third dimension using several series of two-dimensional image recordings acquired orthogonally to the medio-sagittal plane (in the axial plane for the pharynx, then in the coronal plane for the oral cavity).

Inria MultiSpeech team and IADI laboratory have developed a sustained cooperation in the field of MRI data exploitation, and in particular real-time MRI to model articulatory gestures in speech. We have studied the direct problem of synthesizing the shape of the vocal tract from a sequence of phonemes, and the inverse problem of retrieving articulatory gestures from the speech signal. We have a real-time MRI recording system, unique in France, which enables us to acquire data at a frequency of 50 Hz.

This Master's subject can be pursued as a thesis in several directions. The first involves adapting a static 3D MRI to a new speaker without using dynamic data.

The second is to dynamically reconstruct the articulators of speech (tongue, lips...) in 3D, and not just a restricted set of cross-sections.

Assignment

The work will be based on a database acquired for a speaker for whom a short 45-second story was read once for 35 slices transverse to the medio-sagittal plane. The data comprise the debruited speech signal and real-time MRI images (50 frames per second) for the 35 slices.

A millimeter-precision 3D MRI acquisition is also available, enabling all acquisitions to be aligned to a single geometric reference frame.

The first step will be to temporally re-align all audio acquisitions to ensure that the images of all cross-sections correspond to the same sound. This work will be based on phonetic segmentation of the 35 recorded acquisitions and forced alignment tools using automatic speech recognition.

The second step will be to determine the cross-sectional area corresponding to air, and it will be possible to use the medio-sagittal shape to improve detection accuracy. For this task we can either adapt segmentation tools that we have developed[2] from R-CNN [3], or use available automatic tools such as SegmentAnyting https://segment-anything.com/.

The third step will be to design and train a prediction model that takes the medio-sagittal shape as input and generates the vocal tract area for the 35 cross-sections. This will then be used to determine the area of the vocal tract perpendicular to the central line corresponding to the propagation of the sound wave in the vocal tract.

It should be emphasized that this work will represent a major step forward for articulatory speech synthesis, as no satisfactory solution currently exists. Earlier work only partially addressed the problem[4] and very recent approaches to reconstructing the vocal tract in 3D only involve small real-time MRI movies[5].

[1] Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, Yves Laprie Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated Speech Communication, 141:1–13, 2022.

[2] Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Jacques Felblinger, Pierre-André Vuissoz, Yves Laprie. Automatic segmentation of vocal tract articulators in real-time magnetic resonance imaging Computer Methods and Programs in Biomedicine, 243, 2024.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. Mask R-CNN. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[4] Richard S. McGowan, Michel T-T. Jackson, Michael A. Berger. Analyses of vocal tract cross-distance to area mapping: An investigation of a set of vowel images. J. Acoust. Soc. Am. 1 January 2012, 131(1): 42–434.

[5] Isaieva K, Odille F, Laprie Y, Drouot G, Felblinger J, Vuissoz P.-A. Super-Resolved Dynamic 3D Reconstruction of the Vocal Tract during Natural Speech. J Imaging., 9(10):233, 2023

Skills

Technical skills and level required : Master 1 in computer science or applied mathematics

Languages : English or French

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

Internship bonus: €4.35/hour (plus or minus €670/month)

Remuneration for thesis: €2100 gross/month the 1st year

General Information

- **Theme/Domain :** Language, Speech and Audio
  Scientific computing (BAP E)
- **Town/city :** Villers lès Nancy
- **Inria Center :** Centre Inria de l'Université de Lorraine
- **Starting date :** 2025-02-01
- **Duration of contract :** 6 months
- **Deadline to apply :** 2024-12-01

Contacts

- **Inria Team :** MULTISPEECH
- **Recruiter :**
  Laprie Yves / yves.laprie@loria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.