2022-04654 - PhD Position F/M Acoustic to Articulatory Inversion by using dynamic MRI images

Level of qualifications required: Graduate degree or equivalent
Fonction : PhD Position

Context

This project is a LUE (Lorraine Université d'Excellence) project; it will be conducted jointly by the Loria laboratory (MultiSpeech team) and the IADI laboratory (INSERM U1254), which have been working together for several years on speech production and vocal tract imaging.

In particular, this will allow us to use the two-dimensional MRI acquisition system in real time (at 50 images per second) of the IADI laboratory. This system, unique in France, enables imaging the vocal tract at a frequency of 50 Hz in any direction, which is interesting for the recovery of the area function.

Yves Laprie (Yves.Laprie@loria.fr) and Pierre-André Vuissaz (pa.vuissaz@chu-nancy.fr) will co-supervise this PhD work.

Assignment

Scientific challenge

Articulatory synthesis mimics the speech production process by first generating the shape of the vocal tract from the sequence of phonemes to be pronounced, then the acoustic signal by solving the aeroacoustic equations [1, 2]. Compared to other approaches to speech synthesis which offer a very high level of quality, the main interest is to control the whole production process, beyond the acoustic signal alone.

The objective of this PhD is to succeed in the inverse transformation, called acoustic to articulatory inversion, in order to recover the geometric shape of the vocal tract from the acoustic signal. A simple voice recording will allow the dynamics of the different articulators to be followed during the production of the sentence.

Beyond its interest in terms of scientific challenge, articulatory acoustic inversion has many potential applications. Alone, it can be used as a diagnostic tool to evaluate articulatory gestures in an educational or medical context.

This project is a LUE (Lorraine Université d'Excellence) project, it will be conducted jointly by the Loria laboratory (MultiSpeech team) and the IADI laboratory (INSERM U1254), which have been working together for several years on speech production and vocal tract imaging.

In particular, this will allow us to use the two-dimensional MRI acquisition system in real time (at 50 images per second) of the IADI laboratory. This system, unique in France, enables imaging the vocal tract at a frequency of 50 Hz in any direction, which is interesting for the recovery of the area function.

State of the art and innovative character

Almost all current inversion works are based on the use of data from EMA (ElectroMagnetic Articulography) which gives the position of a few sensors glued on the tongue and other easily accessible articulators. From the point of view of the inversion techniques themselves, deep learning is currently very promising because it allows EMA data corpus to be efficiently exploited. Currently, the LSTM (LongShort-Term Memory) approach and its bidirectional variant gives the best results [3]. However, despite their very good geometric accuracy, and because EMA data can only cover the part of the vocal tract closest to the mouth, the current approaches do not allow the complete geometry of the vocal tract to be retrieved, while it is known, for example, that the larynx plays a determining role on the acoustics of the vocal tract. In practice, this considerably limits the interest of existing inversion techniques since the results cannot be used to reconstruct the speech signal.

The objective of this project is to remove this limitation and the originality is to recover the complete geometry of the vocal tract using dynamic MRI data that we can acquire in Nancy at the IADI laboratory.

This approach will open a real operational bridge between articulatory gestures and acoustics in both directions (physical numerical simulations for the direct transformation and inversion). Another innovative aspect of the inversion that we propose is the identification of the role of each articulator in order to take into account a possible perturbation concerning a specific articulator.

Description of work

The first objective is the inversion of the acoustic signal to recover the temporal evolution of the medio-sagittal slice. Indeed, dynamic MRI provides two-dimensional images in the medio-sagittal plane at 50Hz of very good quality and the speech signal acquired with an optical microphone can be very efficiently deconstructed with the algorithms developed in the MultiSpeech team (examples available on https://artspeech.loria.fr/resources/). We plan to use corpora already acquired in the process of being acquired. These corpora represent a very large volume of data (several hundreds of thousands of images) and it is therefore necessary to preprocess them in order to identify the contours of the articulators involved in speech production (mandible, tongue, lips, velum, larynx, epiglottis). Last year we developed an approach for tracking the contours of articulators in MRI images that gives very good results [8]. Each articulator is tracked independently of the others in order to keep the possibility to analyze the individual behavior of an articulator, e.g. in case one of them fails. The automatically tracked contours can therefore be used to train the inversion.
Initially, the goal is to perform the inversion using the LSTM approach on data from a small number of speakers for which sufficient data exists. This approach will have to be adapted to the nature of the data and to be able to identify the contribution of each articulator.

In itself, successful inversion to recover the shape of the vocal tract in the medio-sagittal plane will be a remarkable success since the current results only cover a very small part of the vocal tract (a few percent on the front part of the vocal tract). However, it is important to be able to transpose this result to any subject, which raises the question of speaker adaptation, which is the second objective.

The most recent speaker adaptation techniques are based on the construction of embeddings, which are widely used in speaker recognition or identification, with the idea of "embedding" an individual in a continuous space in order to adapt the system to a new speaker (6, 7). Here, both acoustic and anatomical data are available. In the context of this thesis, the objective is to construct anatomical plunger models, because we wish to be able to study each articulator independently of the others, which requires a fairly precise knowledge of its position and its immediate anatomical environment. This adaptation to the speaker on the basis of a few static MRIs only, answers a double constraint: the rarity and the cost of dynamic MRI on the one hand, and the impossibility of using MRI on the other hand, for example after the insertion of a cochlear implant whose compatibility with MRI is not guaranteed.

We have already addressed the issue of anatomical adaptation through the construction of dynamic atlases of consonant articulation (8), which is based on the use of a fairly classical transformation in medical image processing (5). It has the drawback of not identifying the remarkable anatomical landmarks as such, and the path we intend to follow will be inspired by anatomical plungering recently proposed for the processing of radiological images (4). In spirit, the idea of these plungers is quite close to LSTM (Long Short Term Memory) networks since they combine a global plunge and a local adaptation to the speaker on the basis of a few static MRIs only, answers a double constraint: the rarity and the cost of dynamic MRI on the one hand, and the impossibility of using MRI on the other hand, for example after the insertion of a cochlear implant whose compatibility with MRI is not guaranteed.

Main activities

Environment

The PhD student will be able to use the databases already acquired in the framework of the ANR ArtSpeech (about 10 minutes of speech for 10 speakers) and the much larger databases being acquired in the framework of the ANR Full3DTalkinghead (about 2 hours of speech for 2 speakers). The PhD student will of course also be able to acquire complementary data using the MRI system available in the IADI laboratory (40% of the MRI time reserved for research).

The scientific environment of the two teams is very complementary with a very strong competence in all fields of MRI and anatomy in the IADI laboratory and in deep learning in the MultiSpeech team of Loria. The two teams are geographically close (1.5 km). The PhD student will have an office in both laboratories and the technical means (computer access to the computing cluster) allowing him to work in very good conditions. A follow-up meeting will take place every week and each of the two teams will organize a weekly scientific seminar. The PhD student will also have the opportunity to participate in one or two summer schools and conferences in MRI and automatic speech processing. He/she will also be assisted in writing conference or journal papers.

References

2. Benjamin Elie, and Yves Laprie. Copy-synthesis of phrase-level utterances. EUSIPCO, Budapest 2016 https://hal.archives-ouvertes.fr/hal-01278462
12. Vincente Ribeire, Karyna Isaieva, Justine Leclère, Pierre-André Vuissoz, Yves Laprie. Towards the prediction of the vocal tract shape from the sequence of phonemes to be articulated. IntERSPEECH 2021, Aug 2021, Brno, Czech Republic. (hal-03360113)

Skills

depth learning, computer science, speech processing, applied mathematics

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


Monthly salary after taxes: around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year.