Project description

We thus only consider the temporal geometry of the vocal tract, the aero-acoustic phenomena, and the vocal-fold activity. The advantage is that there exist minimally invasive measuring devices that allow access to the shape of the vocal tract (Magnetic Resonance Imaging) and the activity of the vocal folds (ElectroPhotoGlottoGraphy).

The vocal tract shape, and especially its temporal evolution, has to be modeled so as to provide the numerical acoustic simulations with the relevant geometry at each time point of the synthesis. The shape changes according to the positions of the speech-articulators over time. The articulators move continuously, and the speaker must anticipate the positions to be reached in order to produce the desired sounds.

A speech sound is thus not produced independently of the surrounding sounds. Coarticulation covers the influence of the surrounding sounds on the current sound to be articulated. It should be noted that an articulator that is not critical for the production of a sound, i.e. has not acoustical impact, can anticipate its position for the coming sounds. For instance, during the production of /pa/ the tongue is not recruited by the production of /p/ and thus can anticipate the position required by /a/ well before the acoustic onset of the vowel.

The quantitative prediction of the coarticulation effects is a challenging task. One of the first numerical models was proposed by Ohman [8] and consists of superimposing the effect of the consonants onto the trajectories followed by the articulators between two consecutive vowels. Despite its simplicity, this model is still used for its ease of implementation and relatively good results.

The overlapping of coordinated gestures corresponding to critical articulatory variables (for example the glottal aperture, labial protrusion and aperture, the place and degree of constriction of the tongue tip or body.) is a key element of articulatory phonology. Attempts to calculate gestures from speech and articulatory data [2] are always based on simplifying assumptions so strong that they severely limit the scope of the results.

The approach proposed by Cohen and Massaro [3] relies on the idea of finding the influence domain and the coarticulatory effects of each phoneme. These two sets of parameters are trained from a single pdf (ElectroPhotoGlottoGraphy).

The vocal tract is therefore not taken into account in its entirety, and additionally those approaches are unable to involve a true aero-acoustic dimension.

The objective of this work is to train a coarticulation model that covers all the articulators and guarantees that the target sounds can be generated.

Principales activités

Context

Our long term objective is to achieve articulatory synthesis of speech, i.e. the generation of the acoustic signal by simulating the production of the speech signal by a human being.

In order to keep this problem affordable, we do not consider the bio-mechanical phenomena involved in the movement of speech articulators (jaw, tongue, lips, soft palate, larynx and epiglottis). Indeed, the number of muscles involved, their complex organization, the lack of maturity of numerical models applied to muscles and the lack of data make numerical simulations too far from real speech.

We thus only consider the temporal geometry of the vocal tract, the aero-acoustic phenomena, and the vocal-fold activity. The advantage is that there exist minimally invasive measuring devices that allow access to the shape of the vocal tract (Magnetic Resonance Imaging) and the activity of the vocal folds (ElectroPhotoGlottoGraphy).

The vocal tract shape, and especially its temporal evolution, has to be modeled so as to provide the numerical acoustic simulations with the relevant geometry at each time point of the synthesis. The shape changes according to the positions of the speech-articulators over time. The articulators move continuously, and the speaker must anticipate the positions to be reached in order to produce the desired sounds.

A speech sound is thus not produced independently of the surrounding sounds. Coarticulation covers the influence of the surrounding sounds on the current sound to be articulated. It should be noted that an articulator that is not critical for the production of a sound, i.e. has not acoustical impact, can anticipate its position for the coming sounds. For instance, during the production of /pa/ the tongue is not recruited by the production of /p/ and thus can anticipate the position required by /a/ well before the acoustic onset of the vowel.

The quantitative prediction of the coarticulation effects is a challenging task. One of the first numerical models was proposed by Ohman [8] and consists of superimposing the effect of the consonants onto the trajectories followed by the articulators between two consecutive vowels.

Despite its simplicity, this model is still used for its ease of implementation and relatively good results.

The overlapping of coordinated gestures corresponding to critical articulatory variables (for example the glottal aperture, labial protrusion and aperture, the place and degree of constriction of the tongue tip or body.) is a key element of articulatory phonology. Attempts to calculate gestures from speech and articulatory data [2] are always based on simplifying assumptions so strong that they severely limit the scope of the results.

The approach proposed by Cohen and Massaro [3] relies on the idea of finding the influence domain and the coarticulatory effects of each phoneme. These two sets of parameters are trained from a single pdf (ElectroPhotoGlottoGraphy).

The vocal tract is therefore not taken into account in its entirety, and additionally those approaches are unable to involve a true aero-acoustic dimension.

The objective of this work is to train a coarticulation model that covers all the articulators and guarantees that the target sounds can be generated.
Since this year the IADI laboratory with which we have been collaborating for many years has been equipped with a real-time MRI data acquisition system (at 50 Hz) that allows us to monitor the evolution of the midsagittal shape of the vocal tract during speech production. This represents a considerable asset in the perspective of studying and modeling coarticulation for several speakers.

The work proposed consists of exploiting these data and is organized in two stages.

The first will consist of tracking articulators in MRI data. Unlike several approaches which process the complete vocal tract as a single object we want to track each articulator independently because even if their movements are coordinated, they are not necessarily synchronized. The fact of connecting all the articulator contours in one general contour from the glottis to the lips thus prevents the coarticulation from being studied at the level of articulators. We already drawn articulatory contours in about a thousand images, and preliminary tests we carried out show that this enables fairly good results for the tongue. The objective is to implement a deep-learning auto-encoding approach which, in a first step learns the image and the associated contour for the images with outlined contours, and in a second step retrains the first hidden layers without the contours so as to enable the reconstruction of the contours, and thus tracking, without their prior knowledge [4,5,6].

The second step will be devoted to the modeling of coarticulation via deep learning techniques by identifying the role of each articulator in order to integrate the phenomena of acoustic compensation between articulators.

References

Compétences

Required qualifications
PhD in computer science or acoustics. Knowledge about speech processing and speech production is a decisive plus.

Language
French or English.

Avantages
- 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

Rémunération
Salary: 2653€ gross/month