Some aspects to be investigated include the combined use of only the neutral speech data of the spectral characteristics to produce expressive speech through deep learning-based approaches. The thesis will essentially focus on the two last points, i.e., a correct prediction of prosody and correlates and modification of the spectral characteristics.

Speech synthesis: alteration of pronunciations and presence of pauses, modification of prosody parameters, which leads to the synthetic speech signal. All the levels are involved in expressive parameters (durations of sounds, pitch values, etc.); and finally the generation of acoustic sequences of basic units (phonemes, pauses, etc.) from the source text; the generation of prosody leads to achieve high intelligibility. The speech style was then typically a “reading style”, which resulted from the style of the speech data used to develop TTS systems (reading of a large set of sentences). Although a reading style is acceptable for occasional interactions, TTS systems should benefit from more variability and expressivity in the generated speech signals, for example, for lengthy interactions between machines and humans, or for entertainment applications. This is the goal of recent or emerging research on expressive speech synthesis. Contrary to neutral speech, which is typically read speech without conveying any particular emotion, expressive speech can be defined as speech carrying an emotion, or spoken as in spontaneous speech, or also as speech with emphasis set on some words.

Over the last decades, text-to-speech synthesis (TTS) has reached good quality and intelligibility, and is now commonly used in information delivery services, as for example in call center automation, in navigation systems, and in voice assistants. In the past, the main goal when developing TTS systems was to achieve high intelligibility. The speech style was then typically a “reading style”, which resulted from the style of the speech data used to develop TTS systems (reading of a large set of sentences). Although a reading style is acceptable for occasional interactions, TTS systems should benefit from more variability and expressivity in the generated speech signals, for example, for lengthy interactions between machines and humans, or for entertainment applications. This is the goal of recent or emerging research on expressive speech synthesis. Contrary to neutral speech, which is typically read speech without conveying any particular emotion, expressive speech can be defined as speech carrying an emotion, or spoken as in spontaneous speech, or also as speech with emphasis set on some words.

Principal activities

Missions

Deep learning approaches lead to good speech synthesis quality, however the main scientific and technological barrier remains the necessity of having a speech corpora corresponding to the speaker and the target style conditions, here expressive speech. This thesis aims at investigating approaches to overcome this barrier. More precisely, the objective is to propose and investigate approaches allowing expressive speech synthesis for a given speaker voice, using both the neutral speech data of that speaker, or the corresponding neutral speech model, and expressive speech data from other speakers. This will avoid lengthy and costly recording of specific ad hoc expressive speech corpora (e.g., emotional speech data from the target voice speaker).

Let recall that three main steps are involved in parametric speech synthesis: the generation of sequences of basic units (phonemes, pauses, etc.) from the source text; the generation of prosody parameters (durations of sounds, pitch values, etc.); and finally the generation of acoustic parameters, which leads to the synthetic speech signal. All the levels are involved in expressive speech synthesis: alteration of pronunciations and presence of pauses, modification of prosody correlates and modification of the spectral characteristics.

The thesis will essentially focus on the two last points, i.e., a correct prediction of prosody and spectral characteristics to produce expressive speech through deep learning-based approaches. Some aspects to be investigated include the combined use of only the neutral speech data of the target voice speaker and expressive speech of other speakers in the training process, or in an
adaptation process, as well as data augmentation processes.

The baseline experiments will rely on neutral speech corpora and expressive speech corpora previously collected for speech synthesis in the Multispeech team. Further experiments will consider using other expressive speech data, possibly extracted from audiobooks.

Bibliography


Compétences

Required qualifications

Master in automatic language processing or in computer science

Skills and profile

Background in statistics, and in deep learning

Experience with deep learning tools

Good computer skills (preferably in Python)

Experience in speech synthesis is a plus

Avantages sociaux

- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- French courses

Rémunération

Gross Salary per month: 1982€ brut per month (year 1 & 2) and 2085€ brut/month (year 3)