Project acronym: **TTS4ALL**

Project title: **TTS in low resource scenarios: data management, methodology, models, evaluation**

Project leaders: Yannick Estève and Fethi Bougares

Project description (General project goals):

Text-To-Speech (TTS) is the process that enables machines to convert written text into audible speech. Traditionally, TTS systems have been implemented using various technologies, including concatenative synthesis, formant synthesis, and statistical parametric TTS. Recently, significant advancements in speech synthesis models have been achieved thanks to deep learning approaches and the development of self-supervised speech models. These models have shown remarkable improvements in generating natural-sounding speech, even in situations with limited training data and complex linguistic contexts.

TTS in complex linguistic contexts and Low-Resource scenarios presents several challenges that need to be addressed. Obviously, one of the primary issues is the scarcity of data. The amount of available data is often limited, and the audio quality may not be sufficient to train a robust TTS system. Effective tools and methodologies for recording and collecting speech data are crucial. Ensuring the quality of this collected data is a major concern, as poor-quality data can significantly impact the performance of the TTS system. Additionally, natural language processing (NLP) tools may not perform well or may not exist for certain target languages. This is critical for TTS, as good phonetizers are needed to convert text into phonetic input. Aligned audio-text data, obtained through manual or automatic transcription, is also essential. One potential solution to overcome these challenges is to use trained discrete speech units. However, further study is needed to determine the most relevant units for processing new low-resource spoken languages or considering isolated speech attributes like accent. This could involve exploring novel algorithms that can adapt to the unique characteristics of low-resource languages, ensuring that the synthesized speech remains intelligible and natural-sounding.

Another critical aspect of developping TTS in Low-Resource scenarios is Speech Quality Evaluating. Native speakers are typically needed to evaluate the intelligibility, comprehensibility, expressivity, and audio quality of the synthetic speech signal. However, finding such native speakers for human evaluation can be challenging, especially for low-resource languages. This makes it important to develop automatic tools that perform well in these languages. Automatic evaluation metrics could include objective measures such as Mel-Cepstral Distortion (MCD) and Perceptual Evaluation of Speech Quality (PESQ), which can provide quantitative assessments of speech quality. Additionally, advances in machine learning could lead to the development of more sophisticated evaluation models that can mimic human perception more accurately.

Objectives

The project aims to effectively train and evaluate TTS systems in a situation of scarce training data and complex linguistic contexts. We aim to set up an effective data collection, preparation and evaluation protocols that are adapted to the situation above-mentioned. We will also explore effective strategies for training TTS models for spoken languages without written form or dialects without standardized writing systems. Besides that, we will also address the use of Self-Supervised Learning (SSL) for building TTS and investigate SSL layers in order to find where linguistic content and emotions are encoded.

Furthermore, we will benefit from our multidisciplinary and highly-skilled team to build TTS for additional applications that include speech pseudonymization and streaming TTS. Speech pseudonymization is an area lacking existing resources and previous studies. It involves altering the linguistic content of recorded natural speech to protect the speaker's identity while maintaining the intelligibility of the utterance. This could be particularly useful in scenarios where privacy is a concern, such as in legal or children protection contexts. Streaming TTS is also an emerging topic, which allows for speech generation as symbolic inputs (text or discrete tokens) are provided. This could be particularly useful for integrating TTS with the output of a textual Large Language Model (LLM) or for simultaneous speech translation. Streaming TTS could enable real-time applications where immediate feedback is required, such as in conversational agents or live broadcasting.

Organization

Here below, an initial organization of the project into 5 work packages (see Figure 1). We expressly point that all the WPs are interconnected and team members will participate in multiple tasks from different WPs.

- **WP0 Data Collection and Preparation (lead: Sarah Samson Juan)**: This work package focuses on planning and implementing data collection, preparation, and management strategies. It will be primarily active before the workshop session.

- **WP1 Data Qualification and System Evaluation (lead: Meysam Shamsi)**: This work package addresses the quality of Text-to-Speech (TTS) data and its impact on system accuracy. It also aims to develop new evaluation metrics and improve existing metrics for TTS systems.

- **WP2 TTS for Low-Resource Languages and Accents (lead: Aghilas Sini)**: This work package explores effective strategies for training TTS models for low-resource languages and accents, including scenarios of spoken languages without written form or dialects without standardized writing system.

- **WP3 SSL Speech Representation for TTS (lead: Paulin Melatagia)**: This work package investigates the optimal Self-Supervised Learning (SSL) configuration for building effective TTS models and examines the information representation in SSL layers for TTS applications.

- **WP4 TTS Applications (lead: Salima Mdhaffar)**: This work package involves training TTS models for innovative applications, such as real-time audio generation (streaming) based on text or discrete tokens, and linguistic content pseudonymization to protect sensitive information.

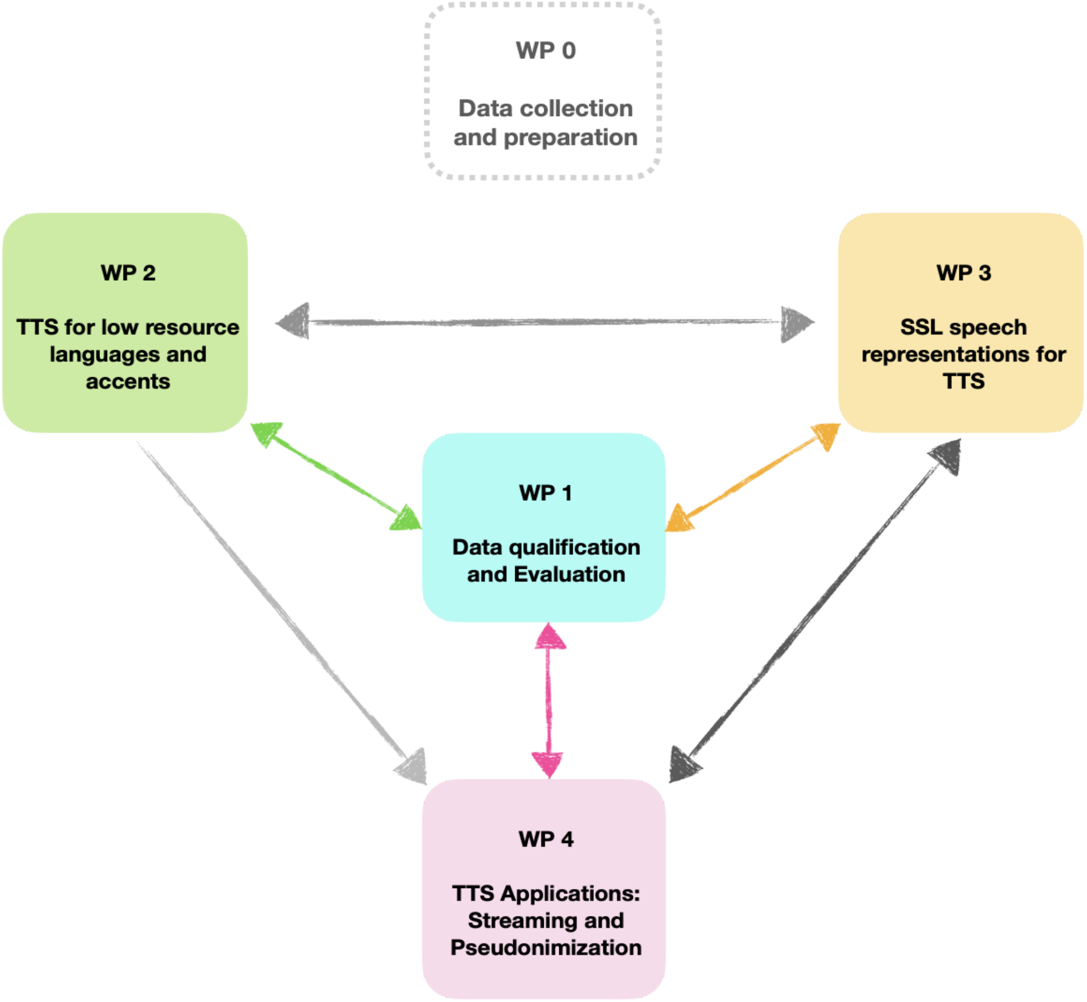
WP 0
Data collection and preparation
WP 2
TTS for low resource languages and accents
WP 3
SSL speech representations for TTS
WP 1
Data qualification and Evaluation
WP 4
TTS Applications: Streaming and Pseudonimization

Figure 1. Interactions between work packages. The direction of the arrow x→y means WPx feeds WPy by providing data, algorithm or scientific points of interest