PhD thesis proposal
Incremental sequence-to-sequence mapping for speech generation using deep neural networks

June 17, 2019

1 Context and objectives

In recent years, deep neural networks have been widely used to address sequence-to-sequence (S2S) learning. S2S models can solve many tasks where source and target sequences have different lengths such as: automatic speech recognition, machine translation, speech translation, text-to-speech synthesis, etc. Recurrent, convolutional and transformer architectures, coupled with attention models, have shown their ability to capture and model complex temporal dependencies between a source and a target sequence of multidimensional discrete and/or continuous data. Importantly, end-to-end training alleviates the need to previously extract handcrafted features from the data by learning hierarchical representations directly from raw data (e.g. character string, video, speech waveform, etc.).

The most common models are composed of an encoder that reads the full input sequence (i.e. from its beginning to its end) before the decoder produces the corresponding output sequence. This implies a latency equals to the length of the input sequence. In particular, for a text-to-speech (TTS) system, the speech waveform is usually synthesized from a complete text utterance (e.g. a sequence of words with explicit begin/end-of-utterance markers). Such approach cannot be used in a truly interactive scenario, in particular by a speech-handicapped person to communicate “orally”. Indeed, the interlocutor has to wait for the complete utterance to be typed before being able to listen to the synthetic voice, hence limiting the dynamics and naturalness of the interaction.

The goal of this project is to develop a general methodology for incremental sequence-to-sequence mapping, with application to interactive speech technologies. It will require the development of end-to-end classification and regression neural models able to deliver chunks of output data on-the-fly, from only a partial observation of input data. The goal is to learn an efficient policy that leads to an optimal trade-off between (variable) latency and accuracy of the decoding process. Possible strategies to decode the output data as soon as possible include: (i) Predicting online “the future” of the output sequence from “the past
and present” of the input sequence, with an acceptable tolerance to possible errors, or (2) learn automatically from the data an optimal “waiting policy” that prevents the model to output data when the uncertainty is too high. The developed methodology will be applied to address two speech processing problems: (i) Incremental Text-to-Speech synthesis in which speech is synthesized while the user is typing the text (possibly with a variable latency), and (ii) Incremental speech enhancement/inpainting in which portions of the speech signal are unintelligible because of sudden noise or speech production disorders, and must be replaced on-the-fly with reconstructed portions.

2 Work plan

The proposed working plan is the following:

- Bibliographic work on S2S neural models, in the context of speech recognition, speech synthesis, and machine translation as well as their incremental (low-latency) variations
- Investigating new architectures, losses, and training strategies toward incremental S2S models.
- Implementing and evaluating the proposed techniques in the context of end-to-end neural TTS systems (the baseline system may be a neural TTS trained with past information/left-context only).
- Implementing and evaluating the proposed techniques in the context of speech enhancement/inpainting, first on simulated noisy speech and then on pathological speech.

3 Requirements

We are looking for an outstanding and highly motivated PhD candidate to work on this subject. Following requirements are mandatory:

- Engineering degree and/or a Master’s degree in Computer Science, Signal Processing or Applied Mathematics.
- Solid skills in Machine Learning. General knowledge in natural language processing and/or speech processing.
- Excellent programming skills (mostly in Python and deep learning frameworks).
- Good oral and written communication in English.
- Ability to work autonomously and in collaboration with supervisors and other team members.
4 Work context

Grenoble Alpes Univ. offers an excellent research environment with ample computing facilities, as well as remarkable surroundings to explore over the weekends. The PhD project will be funded by the Grenoble Artificial Intelligence Institute (MIAI). The PhD candidate will work both at GIPSA-lab (CRISSP team) and LIG-lab (GETALP team). The duration of the PhD is 3 years. The salary is between 1770 and 2100 euros gross per month (depending on complementary activity or not).

5 How to apply?

Applications should include a detailed CV; a copy of their last diploma; at least two references (people likely to be contacted); a cover letter of one page; a one-page summary of the Master thesis; the two last transcripts of notes (Master or engineering school). Applications should be sent to thomas.hueber@gipsa-lab.fr, laurent.girin@gipsa-lab.fr and laurent.besacier@imag.fr. Applications will be evaluated as they are received: the position is open until it is filled, with deadline on July 10th, 2019.