Deep learning-based speech coding and synthesis in adverse conditions.

Projet : Vokkero 2023
Type : Internship, 6 months, start of 2021
Offre : vogo-bernin-pfe-2
Contact : r.vincent@vogo.fr

Keywords : Neural vocoding, deep-learning, speech synthesis, training dataset, normalisation.

Résumé : The project consists in evaluating the performances of the LPCNet neural vocoder for speech coding and decoding under adverse conditions (noisy environment, varied speech style, etc.) and in proposing learning techniques to improve the quality of synthesis.

1 L’entreprise VOGO, le Gipsa-lab

Vogo is an SME based in Montpellier, south of France : www.vogo-group.com. Vogo is the first Sportech listed on Euronext Growth and develops solutions that enrich the experience of fans and professionals during sporting events. Its brand Vokkero is specialized in the design and production of radio communication systems : www.vokkero.com. It offers solutions for teams working in very noisy environments and is notably a world reference in the professional sports refereeing market.

Gipsa-lab is a CNRS research unit joint with Grenoble-INP (Grenoble Institute of Technology), and Université Grenoble Alpes. With 350 people, including about 150 doctoral students, Gipsa-lab is a multidisciplinary research unit developing both basic and applied researches on complex signals and systems. Gipsa-lab is internationally recognised for the research achieved in Automatic Control, Signal and Images processing, Speech and Cognition, and develops projects in the strategic areas of energy, environment, communication, intelligent systems, Life and Health and language engineering.

2 Le projet Vokkero 2023

Every 3 years, Vokkero renews its Hardware (radio, cpu) and Software (rte, audio processing) platforms, in order to design new generations of products. The project extends over several years of study and it is within this framework that the internship is proposed. In the form of a partnership with the Gipsa-lab, the project consists in the study of speech coding using « neural networks » approaches, in order to obtain performances not yet reached by classical approaches. The student will work at the GIPSA-lab in the CRISSP team of the Speech and Cognition cluster under the supervision of Olivier PERROTIN, research fellow at CNRS, and at the R&D of Vogo Bernin, with Rémy VINCENT, project leader on the Vogo side.
3 Context & Objectives

The project consists in evaluating the performances of the LPCNet neural vocoder for speech coding and decoding under adverse conditions (noisy environment, varied speech style, etc.) and in proposing learning techniques to improve the quality of synthesis.

3.1 Context

Vocoders (voice coders) are models that allow a speech signal to be first reduced to a small set of parameters (this is speech analysis or coding) and then reconstructed from these parameters (this is speech synthesis or decoding). This coding/decoding process is essential in telecommunication applications, where speech is coded, transmitted and then decoded at the receiver. The challenge is to minimise the quantity of information transmitted, while keeping the quality of the reconstructed speech signal as high as possible. Current techniques use high-quality speech signal models, with a constraint on algorithmic complexity to ensure real-time processes in embedded systems. Examples of Codecs widely used are Speex (Skype) and its little brother, Opus (Zoom). A few orders of magnitude: OPUS converts a sampled stream at 16kHz into a bitstream at 16kbits (i.e. a compression ratio of 1:16), the reconstructed signal is also at 16kHz and has 20ms of latency.

Since 2016 a new type of vocoder has emerged, called neural vocoder. Based on deep neural network architectures, these are able to generate a speech signal from the classical input parameters of a vocoder, without a priori knowledge of an explicit speech model, but using machine learning. The first system, Google’s WaveNet [1], is capable of reconstructing a signal almost identical to natural speech, but at a very high computation cost (20 seconds to generate a sample, 16,000 samples per second). Since then, models have been simplified and are capable of generating speech in real time (WaveRNN [2], WaveGlow [3]). In particular, the LPCNet neural vocoder [4, 5], also developed by Mozilla, is able to convert a 16kHz sampled stream into a 4kbits bitstream, and reconstruct a 16kHz audio signal. This mix of super-compression combined with bandwidth extension leads to much higher equivalent compression ratios than 1:16!

However, the ability of these systems to generate high-quality speech has only been evaluated following training on large and homogeneous databases, i.e. 24 hours of speech read by a single speaker and recorded in a quiet environment [6]. On the other hand, in the application of Vokkero, speech is recorded in adverse conditions (very noisy environment), and presents a significant variability (spoken voice, shouted voice, multiplicity of referees, etc.). Is a neural vocoder trained on a read speech database capable of decoding speech of this type? If not, is it possible to train the model on such data, while they are only available in small quantities?

The aim of this internship is to explore the limits of the LPCNet vocoder in application to the decoding of referee speech. Various learning strategies (curriculum training, transfer learning, learning on augmented data, etc.) will then be explored to try to adapt pre-trained models to our data.

3.2 Tasks

The student will evaluate the performance of a pre-trained LPCNet vocoder on referee speech data, and will propose learning strategies to adapt the model to this new data, in a coding/re-synthesis scena-
Deep learning-based speech coding and synthesis in adverse conditions.

1. Get familiar with the system, performance evaluation on an audio-book database (baseline);
2. Evaluation of LPCNet on the Vokkero database and identification of the limits (ambient noise, pre-treatments, voice styles, etc.);
3. Study of strategies to improve system performance by data augmentation:
   — Creation of synthetic and specific databases: noisy atmospheres, shouted voices;
   — Recording campaigns on Vokkero systems, in anechoic rooms and/or real conditions if the sanitary situation allows it;
   — Comparison of the 2 approaches according to various learning strategies to learn a new model from this data.

3.3 Required Skills

The student is expected to have a solid background in speech signal processing and an interest in Python development. Experience in programming deep learning models in Python is a plus. The student is expected to show curiosity for research, scientific rigour in methodology and experimentation, and show autonomy for technical and organisational aspects. Depending on the candidate’s motivation, and subject to obtaining funding, it is possible to pursue this topic as a PhD thesis.

The student will be able to subscribe to the company’s insurance system, will have luncheon vouchers and will receive a monthly gratuity of 800€.

Références