Speech Prediction using an Adaptive Recurrent Neural Network with Application to Packet Loss Concealment

Reza Lotfidereshgi, Philippe Gournay

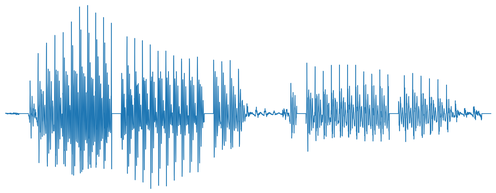
Speech and Audio Research Group, Université de Sherbrooke

Objectives

The goal of this research is to:

- Design a predictor to capture all sorts of linear and nonlinear dependencies between speech signal samples.
- Avoid feature extraction and use an end-to-end predictor which operate directly on samples.
- Actively train the predictor based on the recent history of the signal since it contains the most relevant information.
- Achieve high subjective quality.

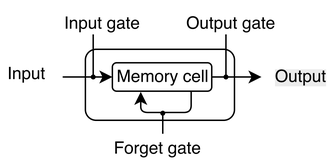
Packet Loss Concealment



A sample signal with some lost packets

Packet Loss Concealment (PLC) is a straightforward and typical application for speech prediction.
The purpose of PLC in a Voice over Packet Network (VoPN) speech communication system is to provide a replacement for unavailable (either lost or overly delayed) speech packets.

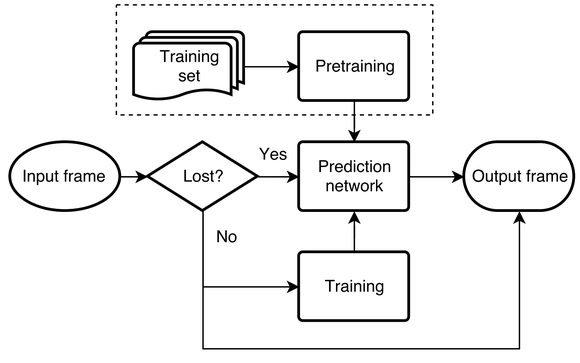
LSTM Network



Structure of a LSTM block

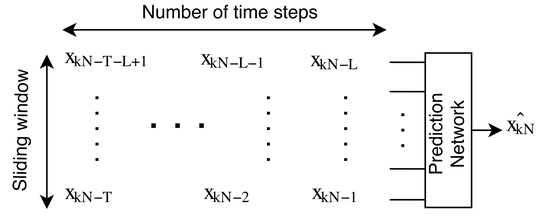
An LSTM network is used as predictor. It is composed of blocks, each block containing different gates that control the flow of information.

Overview of the System



Flowchart of the proposed PLC algorithm

Inputs of the Predictor

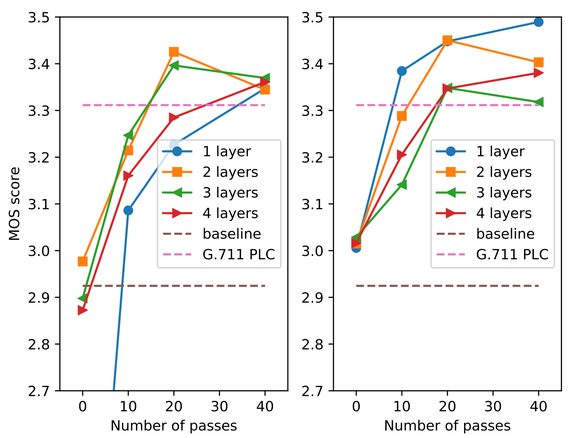


Inputs of the proposed predictor

The prediction network operates directly on speech samples. A sliding window of consecutive samples is fed to the network at every time step.

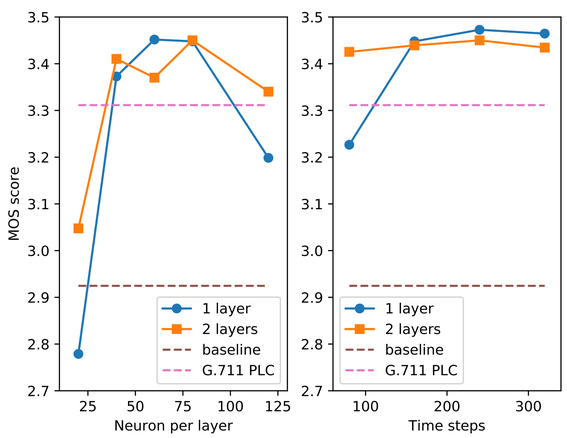
Performance Evaluation

A variety of configurations for the proposed speech predictor are explored. The experiments are done on a subset of the TIMIT database. The MOS score is obtained using the Perceptual Evaluation of Speech Quality (PESQ) software tool.



MOS score as a function of number of passes

Multiple training passes are made over a small training set of the most recent history of the signal. The left panel corresponds to 80 time steps and the right one to 160 time steps.
In both cases, both the number of neurons per layer and the size of the sliding window are set to 80.



MOS score as a function of number of neurons

The results shown above were obtained with a varying number of neurons per layers (left panel) and the number of time steps (right panel).
In both cases, the number of passes over the recent history is equal to 20.

Important Results

- Pretraining is not enough to allow the network to predict speech efficiently.
- A single layer of only 40 LSTM neurons performs reasonably well when online training is used.
- Increasing the number of time steps beyond 160 does not seem to increase performance significantly.

Conclusion and Outlooks

- Since the local variability of the signal is limited compared to the variability of speech in general, a small network structure is effective. This allows the network to operate directly on speech samples.
- With proper setting, the proposed predictor was shown to outperform the standard ITU-T G.711 Appendix I PLC algorithm.
- Results were obtained using a completely speech-agnostic system. In fact, the proposed predictor is a very general tool that could benefit other applications in speech processing, and that could apply to other correlated yet highly dynamic types of data.
- Different neural network configurations, different pretraining and training procedures, or even different types of neural networks may further improve performances.

Contact Information

- Web: http://www.gel.usherbrooke.ca/audio/
- Email: Reza.Lotfi.Dereshgi@USherbrooke.ca
  Philippe.Gournay@USherbrooke.ca
- Phone: +1 (819) 821-8000

UNIVERSITÉ DE SHERBROOKE