GlottDNN

A full-band glottal vocoder for statistical parametric speech synthesis

USER MANUAL
Version 0.1

January 18, 2017
Manu Airaksinen & Lauri Juvela

Contents

1 Introduction

1.1 Background

1.2 Licence and distribution

1.3 Authors & contact

1.4 Acknowledgements

2 Quick-start guide

2.1 Installation

The vocoder C++ code has the following library dependencies:

- `GSL` (GNU scientific library), for basic linear algebra and FFT etc.
- `libsndfile`, for reading and writing audio files
- `libconfig`, for reading structured configuration files

Usually the best way to install the dependencies is with the system package manager.

Alternatively, you may download the source code for the libraries and compile them yourself. The

https://www.gnu.org/software/gsl/

http://www.mega-nerd.com/libsndfile/

http://www.hyperrealm.com/libconfig/

Additionally, this package uses the C++ wrappers for GSL provided at http://gslwrap.sourceforge.net. libsndfile1-dev libconfig++-dev

We recommend to use the GitHub version to get the latest updates, and for the ease of other people contributing to the development `git clone https://github.com/ljuvela/GlottDNN.git`

2.2 Configuration

2.3 Running Analysis

2.4 Running Synthesis

3 GlottDNN Analysis

3.1 Technical description

TODO: Rundown of block diagram etc.

3.2 Spectral (vocal tract) estimation

3.2.1 Quasi-closed phase analysis (QCP)

3.2.2 Iterative adaptive inverse filtering (IAIF)

3.2.3 Frequency-warped time-weighted linear prediction (WWLP)

3.2.4 Quadrature mirror filter (QMF) sub-band analysis

3.3 Harmonic-to-noise ratio (HNR) estimation

3.4 Analysis features

4 GlottDNN Synthesis

4.1 Technical description

TODO: Rundown of block diagram etc.

4.2 Glottal excitation generation

4.2.1 DNN-based excitation generation

4.2.2 Pulses-as-features excitation

4.2.3 Library pulse excitation

4.3 Training of excitation DNN

5 Text-to-speech (TTS) pipeline integration

6 Configuration file explained

6.1 General shared parameters

SAMPLING_FREQUENCY : Sampling frequency should match that of the wav file

FRAME_LENGTH : Analysis frame length (in ms)

UNVOICED_FRAME_LENGTH : Analysis frame length in unvoiced frames. Shorter frames can better capture plosives and other impulse-like unvoiced events.

F0_FRAME_LENGTH : Frame length used for fundamental frequency analysis.

FRAME_SHIFT : Frame rate (in ms)

LPC_ORDER_VT : LPC order for the vocal tract filter

LPC_ORDER_GLOT : LPC order for the glottal source

HNR_ORDER : Number of ERB bands for Harmonic-to-noise ratio

DATA_TYPE : Data type for saving and reading parameters. Valid types are ”ASCII” / ”DOUBLE” / ”FLOAT”

6.2 Pulse extraction related parameters

MAX_PULSE_LEN_DIFF : Percentage of how much pulse length can differ from F0. Pulses are searched iteratively until the nearest pulse fulfilling the length condition is found.

PAF_PULSE_LENGTH : Pulses-as-features length in samples. If interpolation is not used, this should be large enough to fit two pitch periods at the lowest F0.

USE_PULSE_INTERPOLATION : If true, two pitch-period pulses are interpolated to fill the feature vector. If false, the pulse is only centered at GCI.

USE_WAVEFORMS_DIRECTLY : If true, the speech waveform is used directly instead of the inverse filtered waveform.

USE_FOUR_PERIOD_PULSES : If true, Four pitch-periods are used instead of two.

PAF_WINDOW : Select the windowing function applied to the pulse at analysis. Valid options are ”NONE”/”HANN”/”COSINE”/”KBD”

USE_PAF_ENERGY_NORM : Normalize the pulse to unit energy.

Appendices

A Configuration file description

B Excitation DNN file format