Continuous vocoder in feed-forward deep neural network based speech synthesis

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh

Budapest University of Technology and Economics, Hungary

DOGS 2017
Novi Sad, Serbia

Nov 23, 2017

SmartLab
Intelligent Interactions
http://smartlab.tmit.bme.hu

Background

- Text-to-speech synthesis (TTS)
  - Generating speech waveform from textual input
  - Transmit data from a machine to a human user

- Statistical Parametric Speech Synthesis (SPSS) training
  - Flexibility due to the statistical modeling process
    - Hidden Markov-models (HMM)
    - Deep Neural Networks (DNN)
  - Speech signal is analyzed to parameters and synthesized to speech, using vocoder

Background

- Key factors for quality degradation [Zen et al., 2009]
  - Parametric vocoder (speech analysis & synthesis)
  - Acoustic modeling accuracy
  - Over-smoothing (parameter generation)
- Vocoding issues
  - Buzziness
  - Modeling of rare events (creaky voice)
  - Real-time processing
- Research goal
  - Construct a simple and flexible vocoder whose parameters can be controlled to achieve high quality synthesized speech.

Baseline: Continuous vocoder

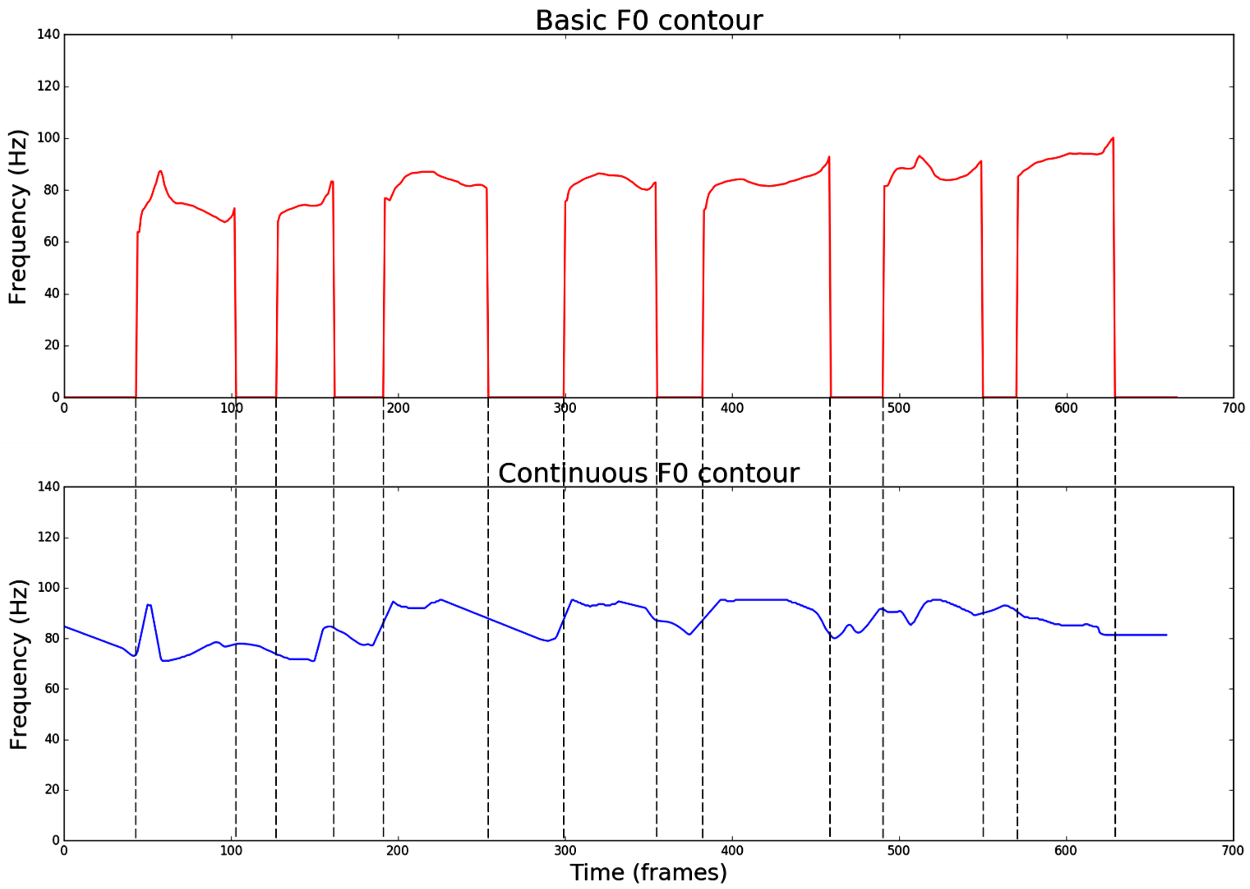
- Analysis
  - Linear Prediction residual-based excitation [Csapó et al., 2016]
  - Continuous fundamental frequency (F0) algorithm [Garner et al., 2013]
  - Maximum Voiced Frequency (MVF) [Drugman and Stylianou, 2014]
  - Standard Mel-Generalized Cepstral (MGC) [Tokuda et al. 1994]
- Statistical training of HMMs
  - Decision tree-based context clustering [Zen et al., 2007]
- Synthesis
  - Voiced and unvoiced excitation component added together according to MVF

Continuous vocoder: Motivation I

- Basic F0 model
  - continuous in voiced regions
  - discontinuous in unvoiced regions
  - hard to model boundaries between voiced and unvoiced segments
  - difficult to handle mixed excitation

- Continuous F0 model
  - no voiced/unvoiced decision
  - decrease the disturbing effect of creaky voice
  - easier to handle mixed excitation

Continuous vocoder: Motivation I



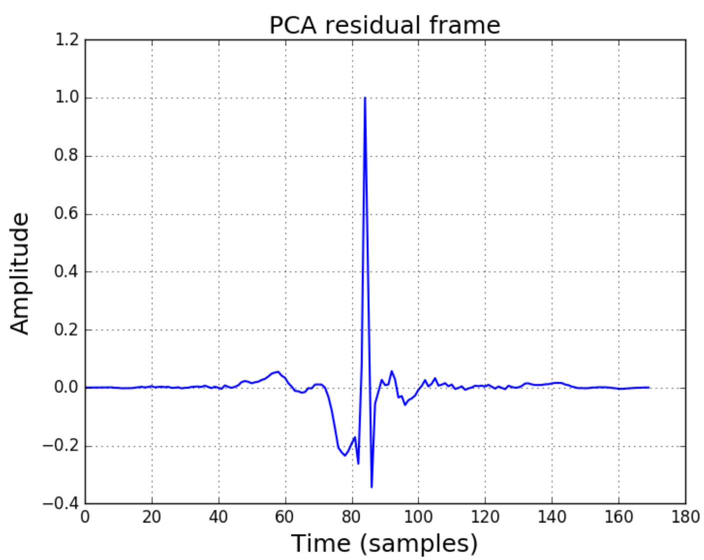
"The girl faced him, her eyes shining with sudden fear."

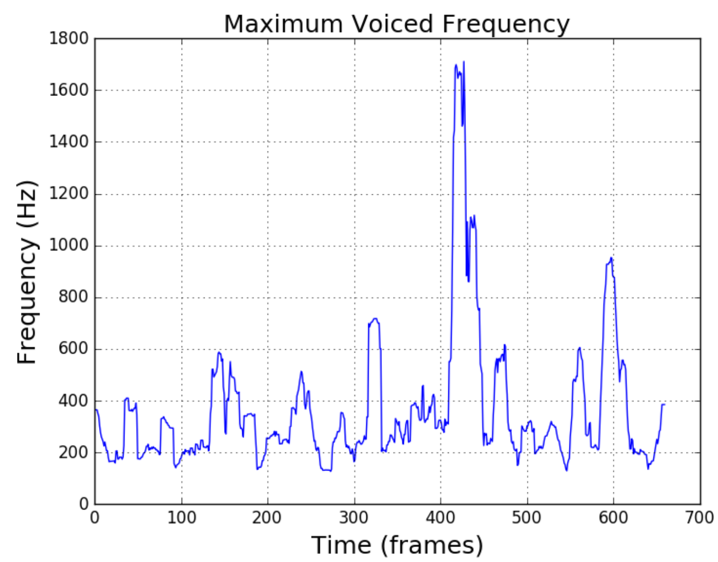
SmartLab
Intelligent Interactions

NVIDIA GPU EDUCATION CENTER

Continuous vocoder: Motivation II

- MVF to model the voiced/unvoiced characteristics of sounds
  - Excitation parameter
- To overcome simple impulse based excitation
  - Principle Component Analysis (PCA) residual frames overlap-added depending on the continuous F0





"The girl faced him, her eyes shining with sudden fear."

SmartLab Intelligent Interactions
NVIDIA GPU EDUCATION CENTER

Continuous vocoder: Problem formulation

- The noise component is not accurately modeled in modern vocoders
  - even in the widely used STRAIGHT vocoder [Degottex et al. 2016]

- There is a lack of voiced component in higher frequencies of the baseline vocoder
  - mixed excitation not probably modeled
  - causes buzzy speech in unvoiced segments

- The high-frequency noise component is time-aligned with the F0 periods [Stylianou 2001]

Objectives: Cases study

1. Extension of a Continuous vocoder [Csapó et al., 2016] based SPSS for advanced modeling

    a) Shaping the high-frequency component by adding True envelope modulated noise to the voiced excitation

    b) Refinement of speech spectral estimation

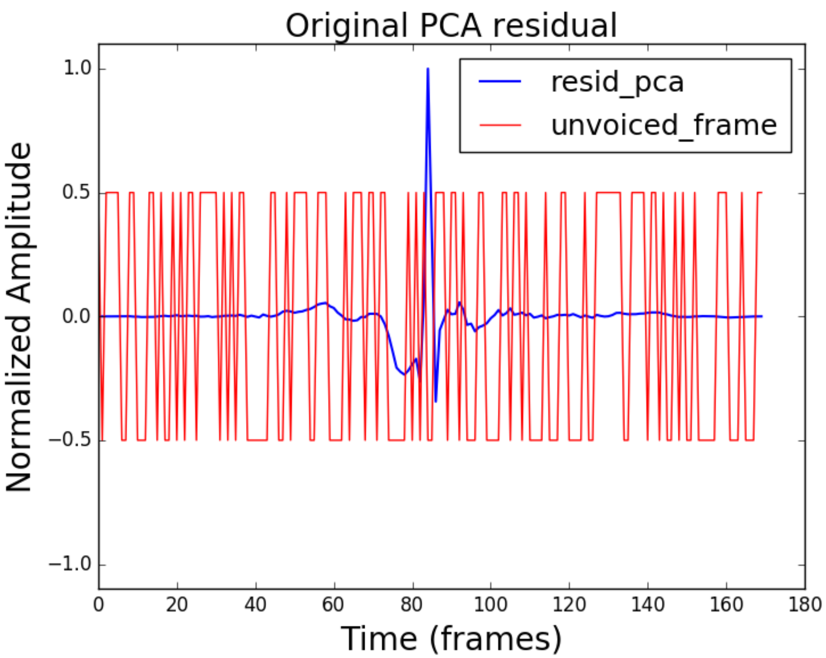
    c) Build a learning model to increase the quality of synthesized speech

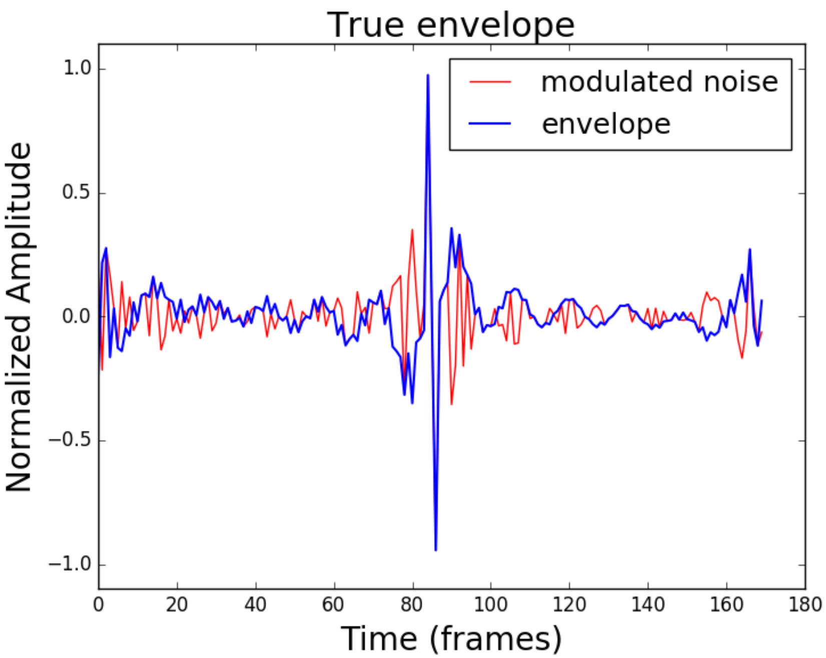
    - feed-forward Deep neural networks (DNNs)

2. Evaluation between Continuous and WORLD vocoders

Proposed I: Adding envelope modulated noise

- Estimating the True time envelope of the speech residual signal to the voiced and unvoiced excitation





Proposed II: Spectral envelope refinement

- CheapTrick [Morise 2015]
  - an accurate and temporally stable spectral envelope estimation

- Spectral Distortion Evaluation
  - Root Mean Square (RMS) Log Spectral Distance (LSD) evaluation was carried out.

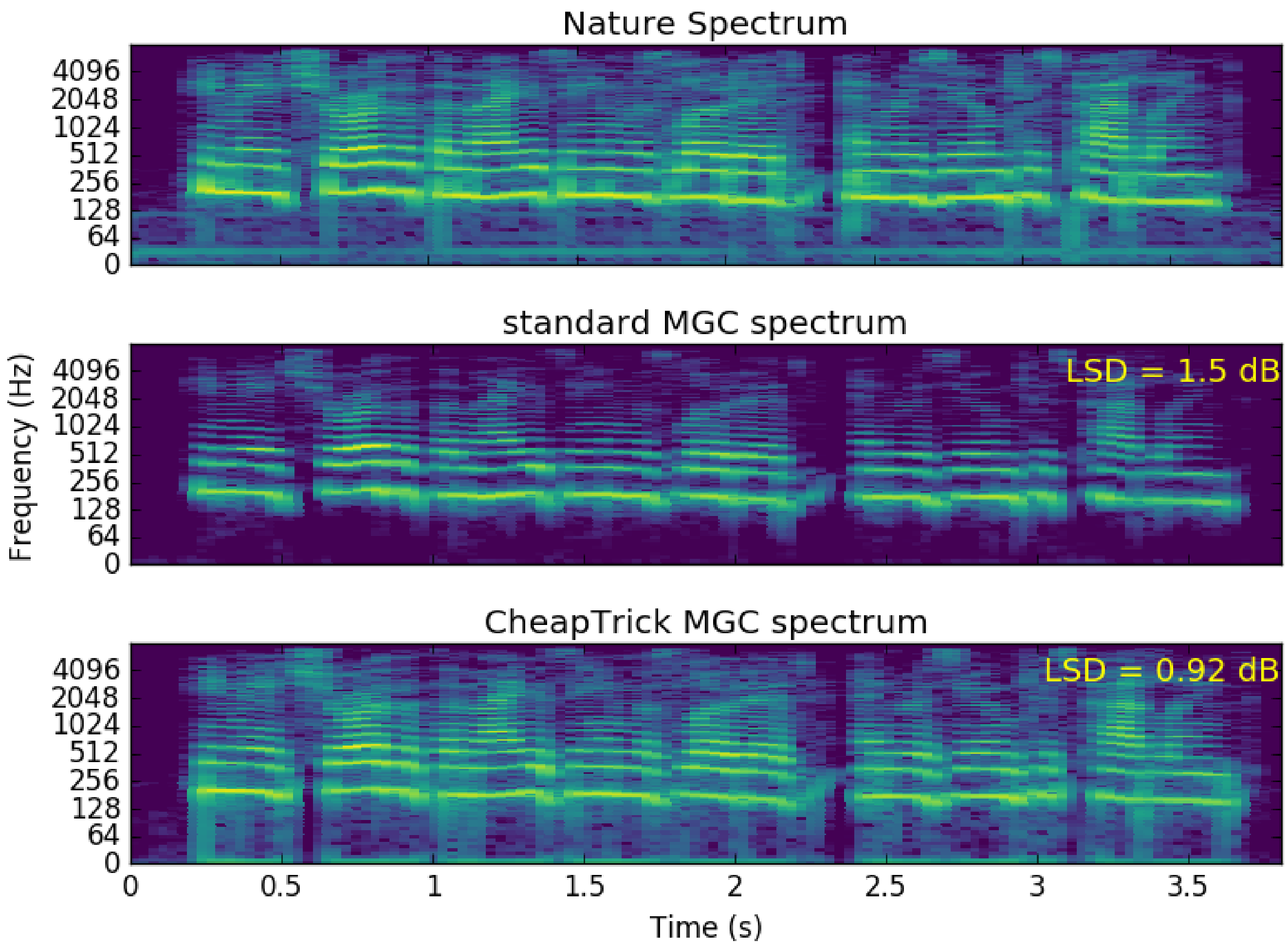
$$LSD = \sqrt{\frac{1}{N}\sum_{k=1}^{N} mean\left[\log P(f_k) - \log \hat{P}(f_k)\right]^2}$$

Where $P(f_k)$ and $\hat{P}(f_k)$ are spectral power magnitudes of the nature and synthesis speech respectively, defined at N frequency points.

SmartLab Intelligent Interactions

NVIDIA GPU EDUCATION CENTER

Proposed II: Spectral distortion evaluation



“He made sure that the magazine was loaded, and resumed his paddling.”

SmartLab Intelligent Interactions

NVIDIA GPU EDUCATION CENTER

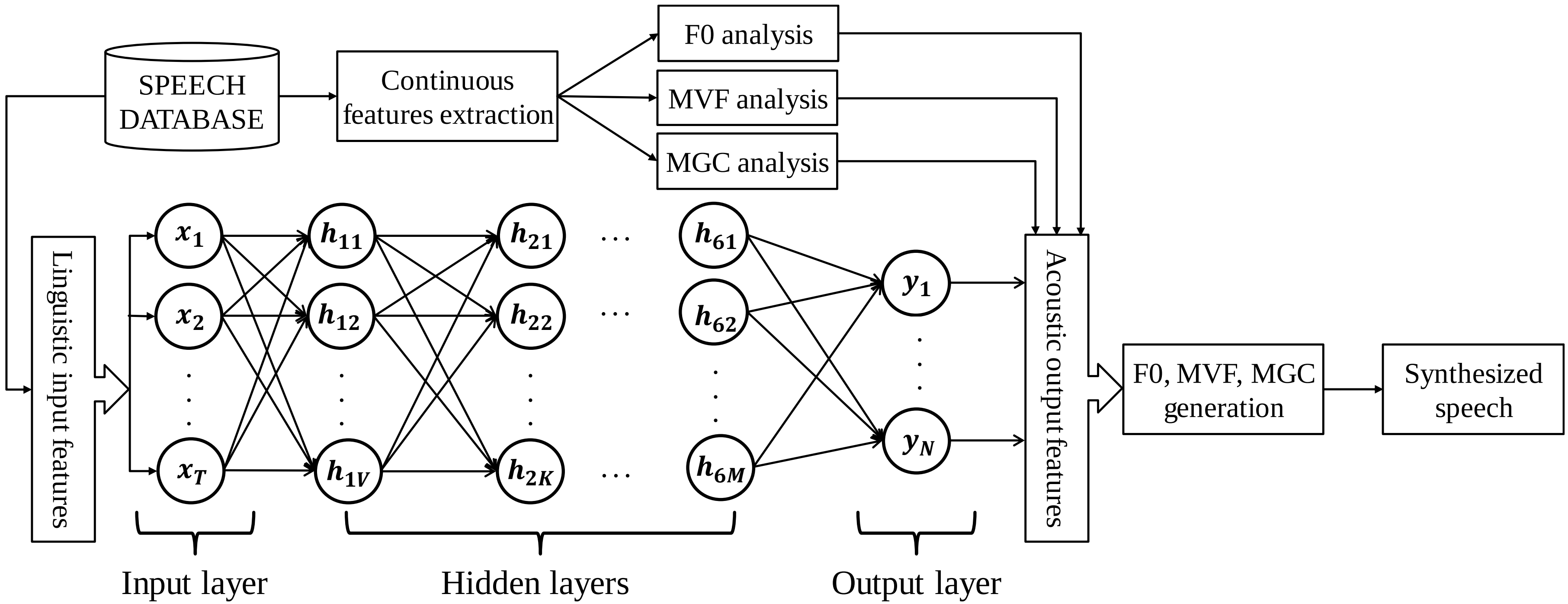
Proposed III: Acoustic modeling

- Feed-Forward deep neural network (DNN)
  - Baseline system was successfully used with HMM based TTS
  - However, HMMs often generate over-smoothed, and muffled synthesized speech
  - DNNs can be viewed as a replacement for the decision tree used in HMM-TTS systems
  - DNN-based acoustic models offer an efficient and distributed representation of complex dependencies between linguistic and acoustic features

- DNN topology
  - 6 feed-forward hidden layers; each one has 1024 hyperbolic tangent units
  - TANH function can yield lower error rates and faster convergence than a logistic sigmoid function.

Proposed III: Acoustic modeling

- Continuous vocoder applied with DNN



Data for the evaluation

- English speaker from CMU-ARCTIC database [Kominek and Black, 2003]
  - SLT (American English, female)
- Waveform sampling rate of the database is 16 kHz
- 90% of these sentences were used for training and the rest were used for testing
- 100 sentences from each speaker were analyzed and synthesized with the baseline and proposed vocoders.
- High performance NVidia Titan X GPU
- Merlin: Open source neural network toolkit [Wu et al. 2016]

Continuous vocoder capability

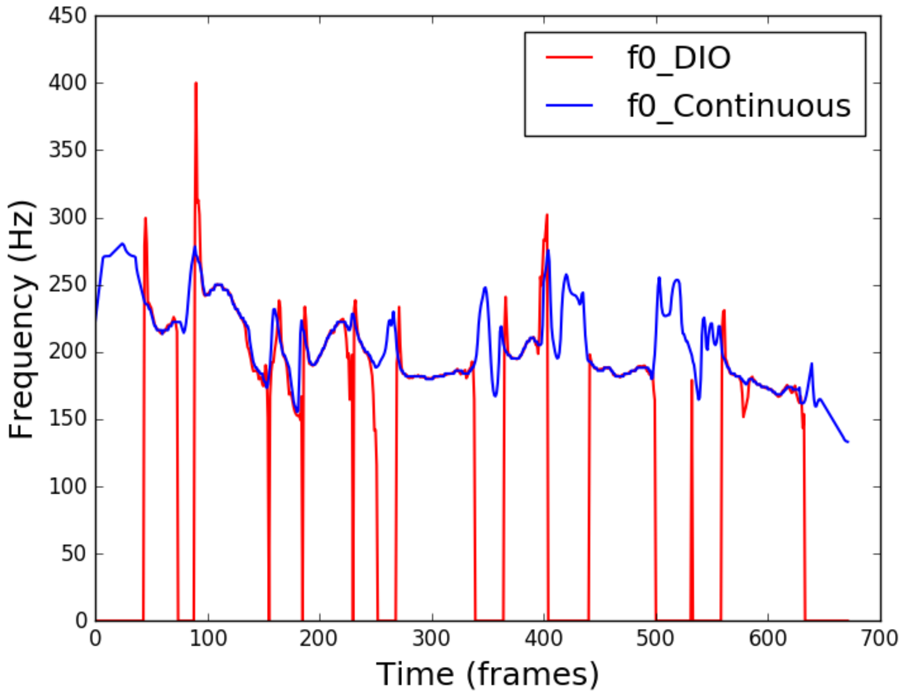
- WORLD vocoder [Morise et al. 2016] was chosen for comparison with our optimized vocoder
    - F0: Distributed Inline-filter Operation (DIO)
    - Band aperiodicity: Definitive Decomposition Derived Dirt-Cheap (D4C)
    - Spectral envelope: CheapTrick

- We found that WORLD vocoder can make V/UV decision errors (V/UV error = 5.35%)
    - setting voiced that should be unvoiced, or vice versa
    - errors at boundaries (at the V/UV or UV/V transitions)
    - Thus, often synthesizes speech with clicks

- Continuous vocoder make V/UV decision errors = 0%
    - voicing feature is modeled by the continuous MVF parameter

Continuous vocoder capability: parameters

Vocoder	Parameters per frame	Excitation
Continuous	F0: 1 + MVF: 1 + MGC: 60	Mixed
WORLD	F0: 1 + Band aperiodicity: 5 + MGC: 60	Mixed



SmartLab Intelligent Interactions

NVIDIA GPU EDUCATION CENTER

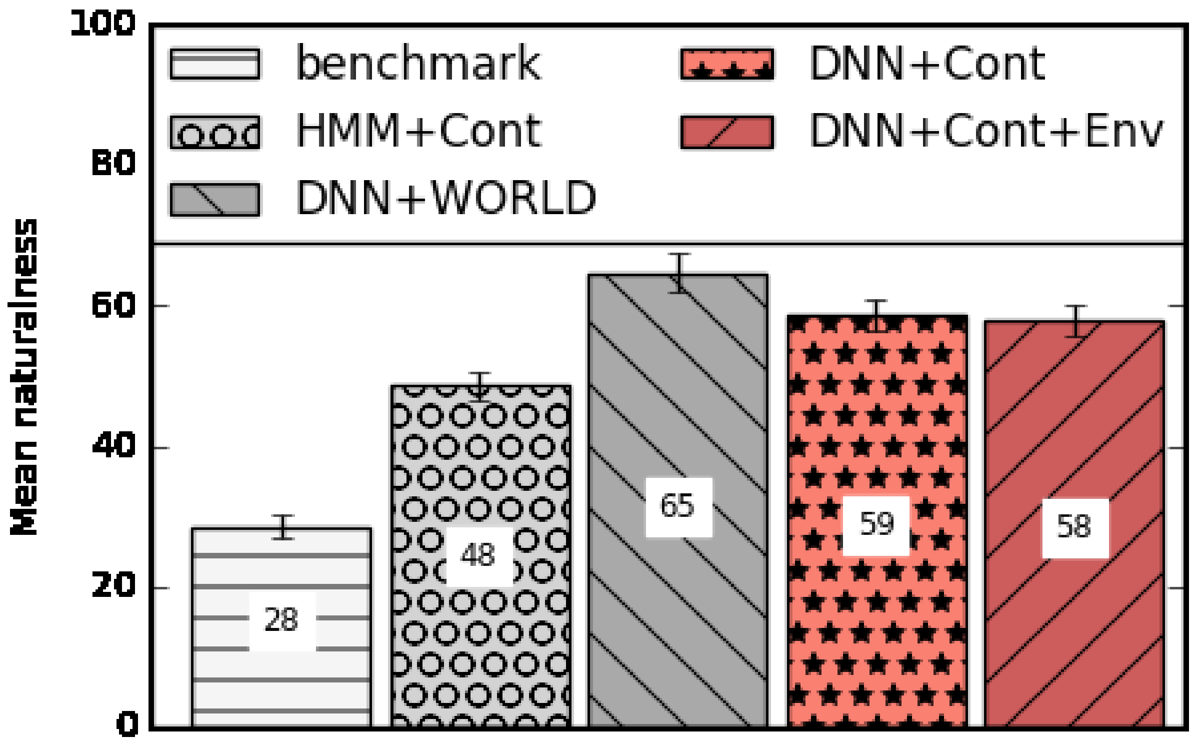
“Author of the danger trail, Philip Steels, etc.”

Subjective evaluation

- MUSHRA: enables evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons.

- MUSHRA listening test
  - reference: natural speech
  - benchmark: impulse-noise excitation
  - 15 sentences were selected from the SLT
  - 6 types x 15 sentences
    - 90 utterances were included in the test
  - 9 participants (7 males, 2 females)
  - The test took 20 minutes to fill

Subjective evaluation



- DNN-TTS with the Continuous vocoder is more natural than the HMM-TTS
- Continuous vocoder still not rated better than WORLD vocoder

SmartLab Intelligent Interactions

NVIDIA GPU EDUCATION CENTER

Online samples:
http://smartlab.tmit.bme.hu/dogs2017_vocoder_dnn

Summary and Future plans

- Modulated noise component
  - Further control the time structure of the high-frequency noise component
- Spectral approach
  - Improved when using the CheapTrick algorithm
- Acoustic modeling
  - DNN-TTS using the continuous parameters was rated better than an earlier HMM-TTS system
- Continuous vocoder has few parameters
  - computationally feasible
  - suitable for real-time operation

- To further reduce the buzziness caused by vocoding, add a Harmonics-to-Noise Ratio parameter to: Analysis phase, statistical learning phase, synthesis phase.

Thanks for listening!