MŰEGYETEM 1782

BME TMIT

RNN-based speech synthesis using a continuous sinusoidal model

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

Department of Telecommunications and Media Informatics

Budapest University of Technology and Economics, Budapest, Hungary

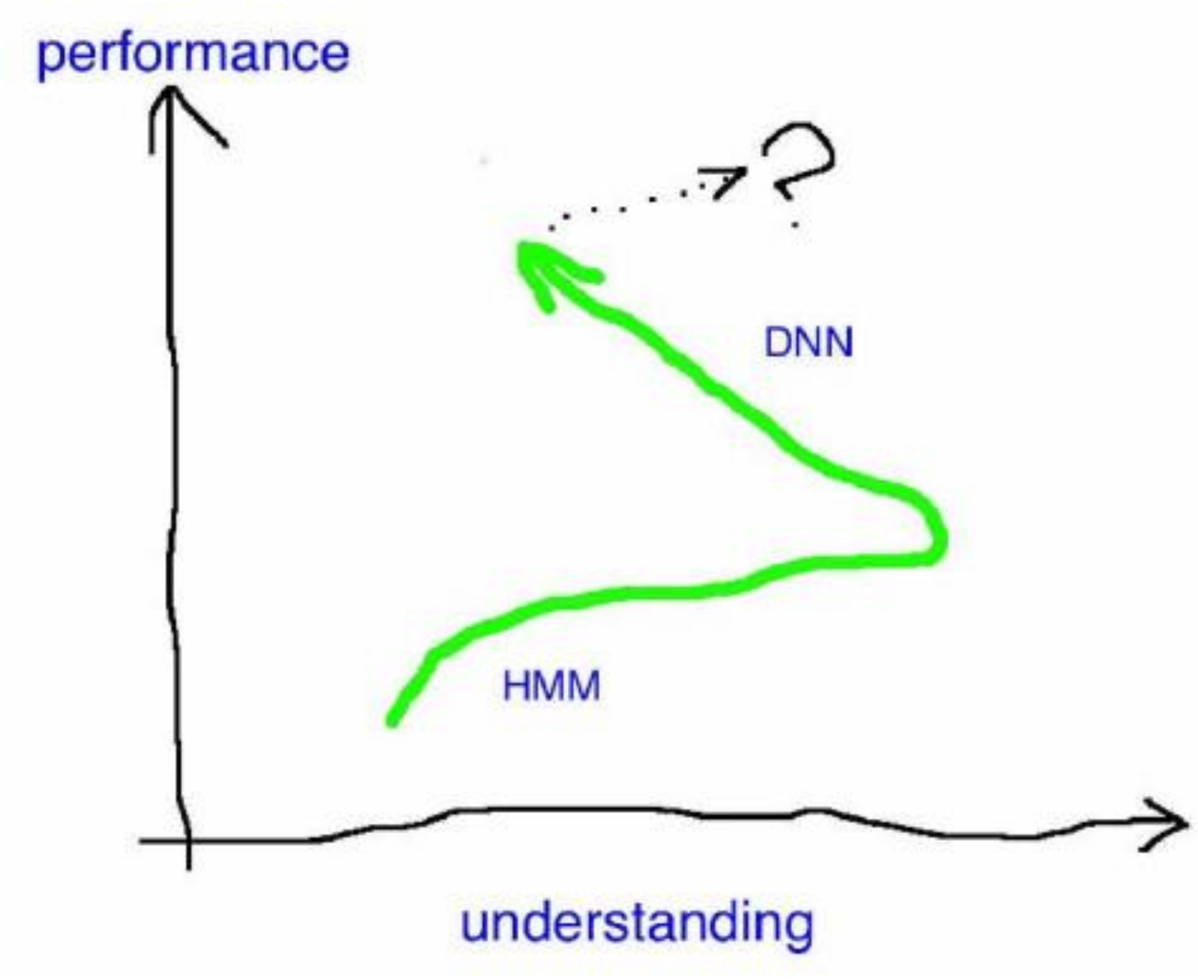
csapot@tmit.bme.hu

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

IJCNN 2019
Budapest, Hungary

NVIDIA GPU EDUCATION CENTER

Motivation



[Kawahara, interspeech2018]

Background

➢ Text-to-speech synthesis (TTS)

- Generating speech waveform from textual input
- Transmit data from a machine to a human user

➢ Vocoder

- Category of speech codec that analyzes and synthesizes human voice
- Provide a parametric representation of the speech signal suitable for coding and statistics

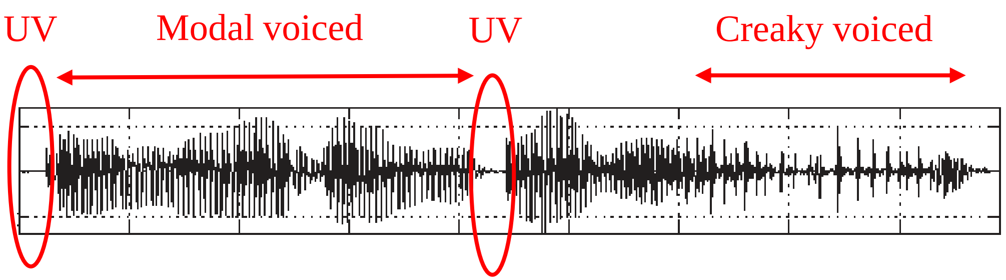
➢ Statistical Parametric Speech Synthesis

- Store statistics rather than waveforms
- Flexibility to change voice characteristics
- Smoothness and style adaptation

Problem formulation

➢ Key factors for quality degradation of speech synthesis:
1. Parametric vocoder (speech analysis & synthesis)
2. Acoustic modeling accuracy
3. Over-smoothing (sounds muffled)

➢ Vocoding issues:
1. Buzziness
2. Creaky voice
3. Real-time processing



❑ STRAIGHT and WORLD are the most widely used vocoders for statistical parametric speech synthesis (SPSS) as a baseline.

But

- STRAIGHT vocoder is too slow to be used in practice because it relies on high-order FFT for high-resolution spectral synthesis.

Hypotheses

- Refining the estimated contF0 algorithm by time-warping approach will
    - eliminate octave errors and isolated glitches.

- Continuous sinusoidal model (CSM)
    - is high quality and computationally efficient.

- Using sequence-to-sequence modeling with recurrent neural networks based Bi-LSTM in order to
    - predict acoustic features (contF0, MVF, and MGC),
    - which are then passed to a CSM to generate the synthesized speech,

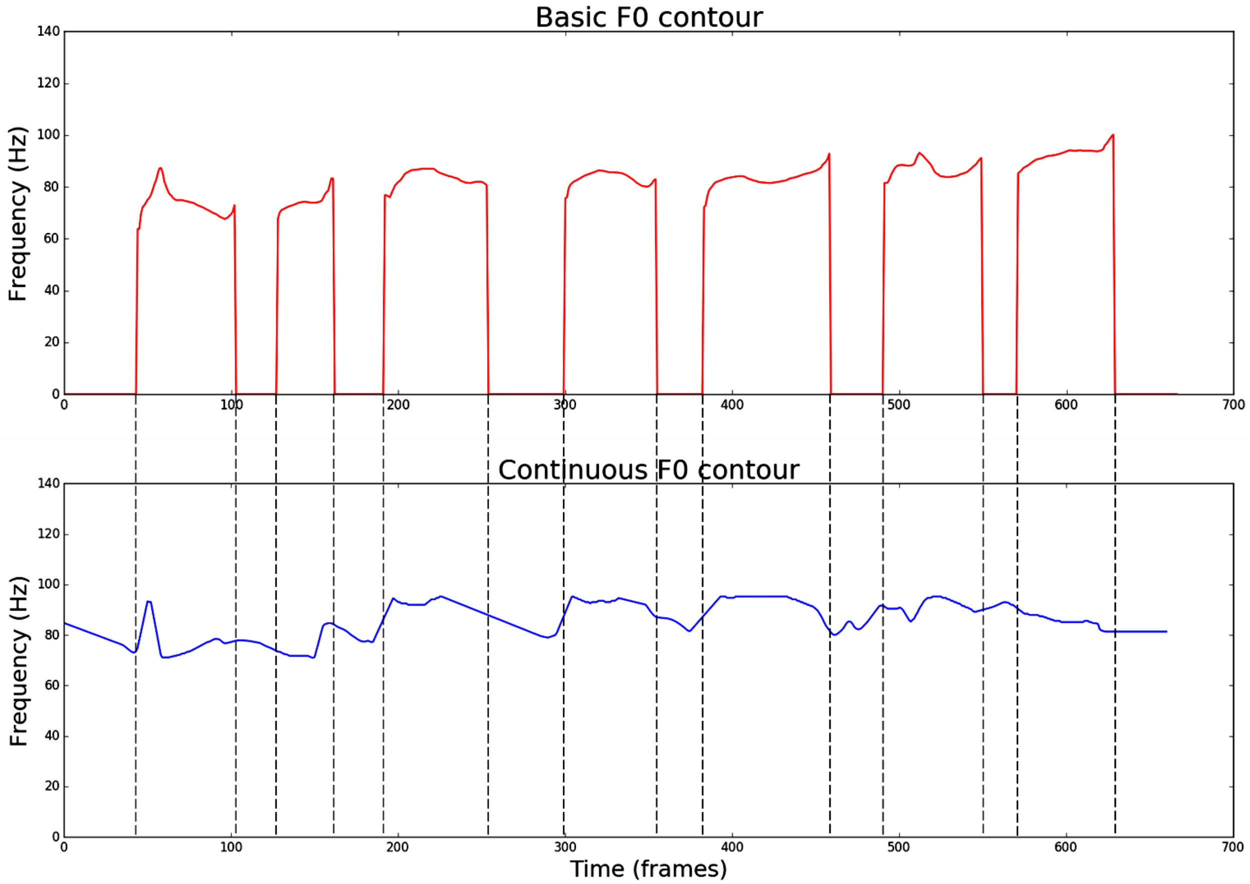
Proposed Methodology

1) Adaptive contF0 using Time-Warping

- Discontinuous F0 model (traditional)
  - Continuous (F0 > 0) in voiced regions
  - discontinuous (F0 = 0) in unvoiced regions
  - hard to model boundaries between voiced and unvoiced segments
  - difficult to handle mixed excitation

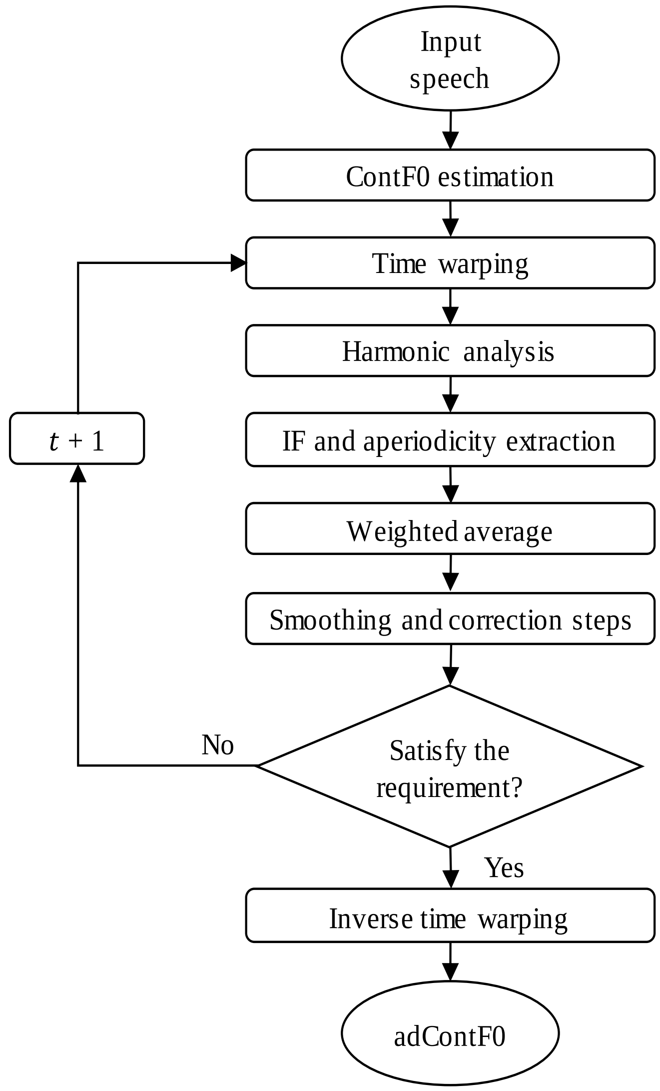
- Continuous F0 model
  - no voiced/unvoiced decision
  - decrease the disturbing effect of creaky voice
  - easier to handle mixed excitation

1) Adaptive contF0 using Time-Warping



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

1) Adaptive contF0 using Time-Warping



The estimating process of the adContF0 based on adaptive time-warping method

2) Continuous Sinusoidal Model (CSM)

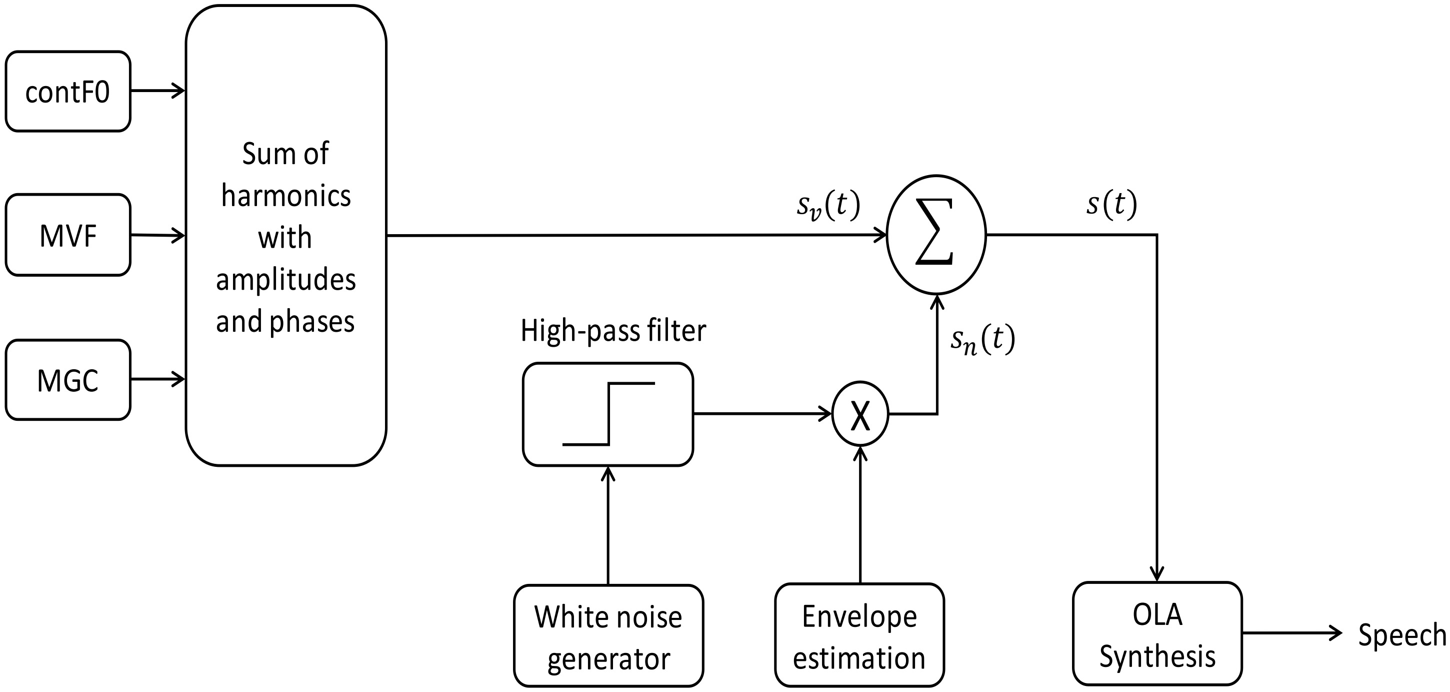
➢ Analysis step:

1. contF0 [Garner et al., 2013]: to model voiced and unvoiced sounds in a uniform way
2. MVF [Drugman and Stylianou, 2014]: to model the voiced/unvoiced characteristics of sounds.
3. MGC [Morise 1994]: Spectral envelope

2) Continuous Sinusoidal Model (CSM)

➢ Sinusoidal synthesis:

Decompose the speech frames into a harmonic/voiced component lower band and a stochastic/noise component upper band based on MVF values.



2) Continuous Sinusoidal Model (CSM)

$$s(t) = s_v(t) + s_n(t)$$

$$s_v^i(t) = \sum_{k=1}^{K^i} A_k^i(\mathrm{t}) \cos\left(w_k^i t + \emptyset_k^i(t)\right) \quad , \quad w_k^i = 2\pi k (contF0)^i$$

where $A_k(\mathrm{t})$ and $\emptyset_k(t)$ are the amplitude and phase at frame i, $t = 0, 1, \dots, N$ and N is the length of the synthesis frame. K is the time-varying number of harmonics that depends on the contF0 and MVF:

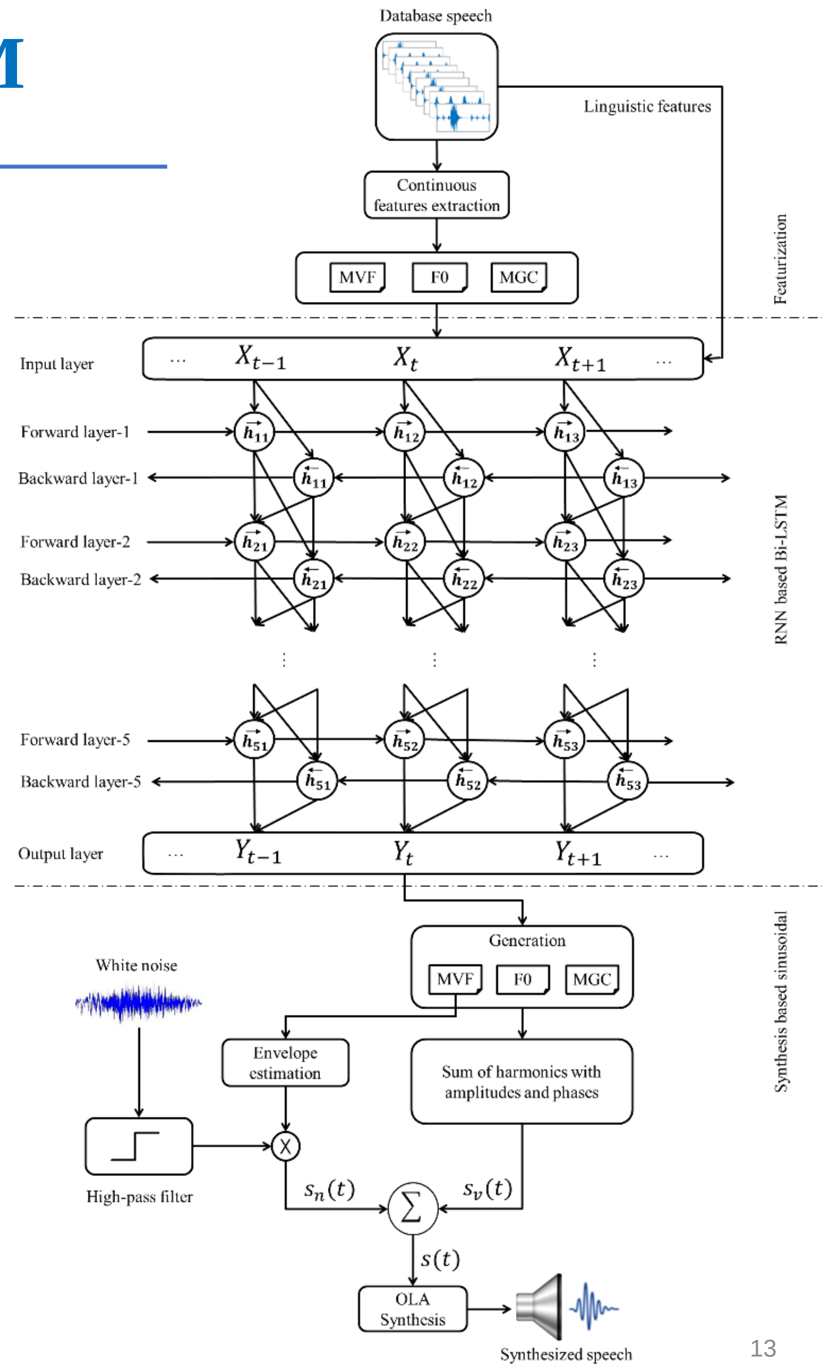
$$K^i = \begin{cases} round\left(\dfrac{MVF^i}{contF0^i}\right) - 1, & voiced\ frames \\ 0, & unvoiced\ frames \end{cases}$$

If the current frame is voiced, the synthesized noise part $n(t)$ is filtered by a high-pass filter $f_h(t)$ with cutoff frequency equal to the local MVF, and then modulated by its time-domain envelope $e(t)$. For unvoiced frames, the harmonic part is obviously zero and the synthetic frame is typically equal to the generated noise.

$$s_n^i(t) = e^i(t)\left[f_h^i(t) * n^i(t)\right]$$

3) RNN-TTS using Bi-LSTM

- Build a learning model to increase the quality of synthesized speech.
  - Recurrent neural network based Bi-LSTM



3) RNN-TTS using Bi-LSTM

For a given

- input vector sequence $x = (x_1, \ldots, x_T)$,
- hidden state vector sequence $h = (h_1, \ldots, h_T)$
- outputs vector sequence $y = (y_1, \ldots, y_T)$

The iterative process of the Bi-LSTM can be defined here as

- $\overrightarrow{h}_t = f\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right)$
- $\overleftarrow{h}_t = f\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right)$
- $y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y$

where a forward state sequence $\overrightarrow{h}$ (positive time direction), backward state sequence $\overleftarrow{h}$ (negative time direction); W is the connection weight matrix between two layers (e.g. W_{xh} is the weight matrix between input and hidden vectors), b is the bias vectors, and $f(\cdot)$ denotes an activation function which is defined as:

Results & Evaluation

Experimental Conditions

- English speaker from CMU-ARCTIC database [Kominek and Black, 2003]
    - SLT (American English, female)
    - AWB (Scottish English, male)
    - BDL (American English, male)
    - JMK (Canadian English, male)
- Waveform sampling rate of the database is 16 kHz
- vocoders:
    - Baseline
    - Anchor
    - Proposed
    - WORLD
- Metrics:
    - Log-Likelihood Ratio (LLR)
    - frequency-weighted segmental SNR
    - Log Spectral Distortion (LSD)

Experimental Conditions

- Neural Network Setting
  - 4 feed-forward hidden layers; each one has 1024 hyperbolic tangent units followed by a single Bi-LSTM layer with 385 units.
  - TANH function can yield lower error rates and faster convergence than a logistic sigmoid function.
  - In RNN-TTS:
    - 90% of these sentences were used for training and the rest were used for testing.
    - High performance NVidia Titan X GPU
    - Merlin: Open source neural network toolkit [Wu et al. 2016]

Definition of F0 Error Measures

1) Gross Pitch Error (GPE)
   is the percentage of incorrectly detected F0 values in voiced speech segments

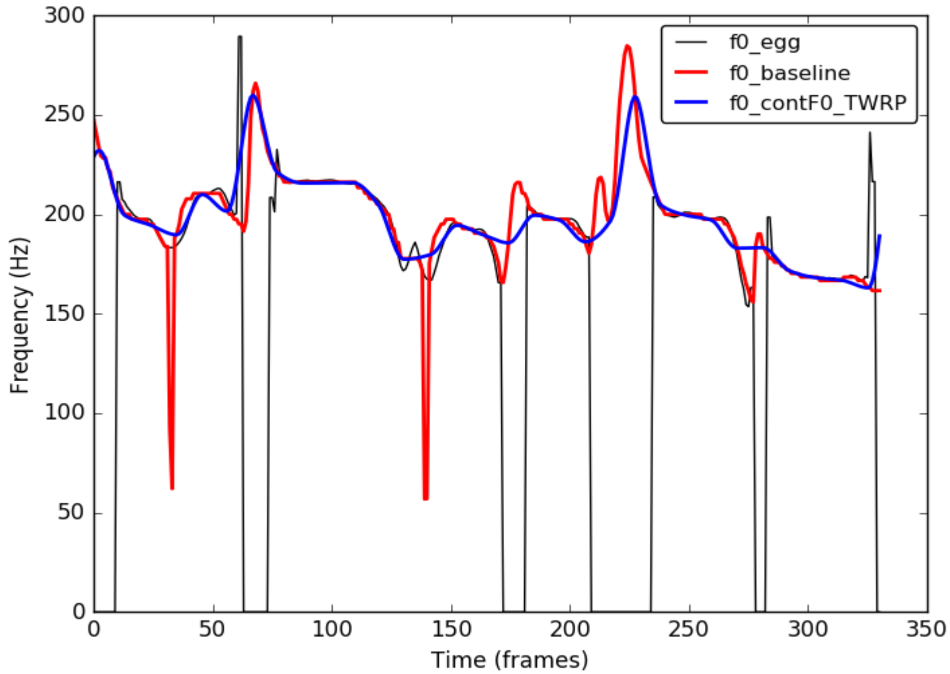
2) Mean Fine Pitch Errors (MPPE)
   is referred to all pitch errors that are not classified as GPE.

3) Standard Deviation of the Fine Pitch Errors (STD)
   is a measure of the accuracy of the F0 detector during voiced intervals

A) ContF0 Evaluation

Clean speech

Method	GPE %			MFPE			STD		
	BDL	JMK	SLT	BDL	JMK	SLT	BDL	JMK	SLT
baseline	12.754	9.850	7.677	3.558	3.428	4.421	4.756	4.513	6.764
contF0_TWRP	**8.294**	**8.777**	7.827	**2.764**	**3.024**	**3.656**	**3.873**	**4.188**	**5.788**

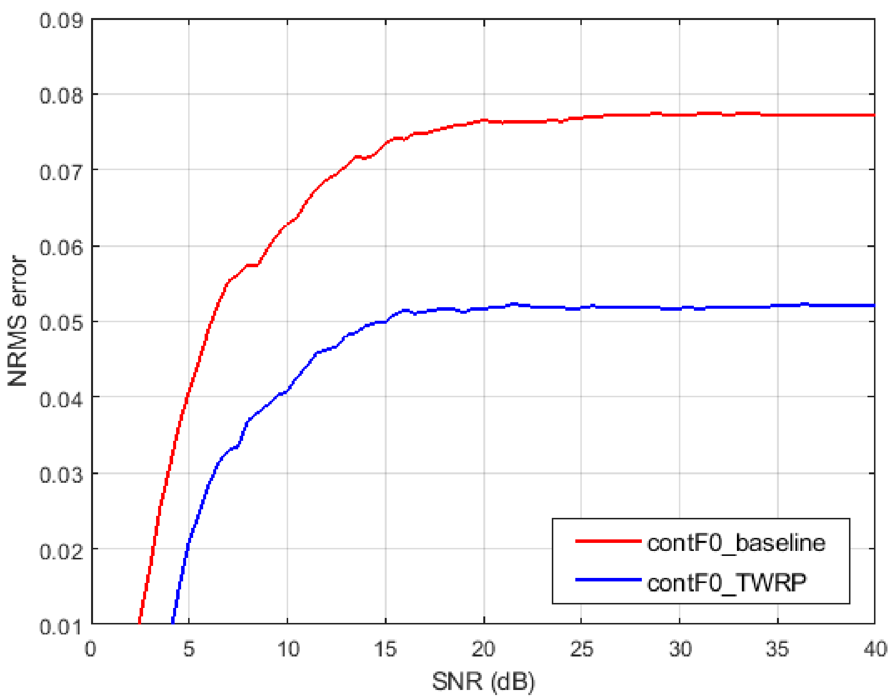


"Everything was working smoothly, better than I had expected.", from speaker SLT.

A) ContF0 Evaluation

Additive white noise

Method	GPE %			MFPE			STD		
	BDL	JMK	SLT	BDL	JMK	SLT	BDL	JMK	SLT
baseline	33.170	40.057	27.502	4.050	3.901	3.512	4.393	4.293	3.912
contF0_TWRP	**29.464**	**37.839**	**26.932**	**3.199**	**3.165**	**2.890**	**3.449**	**3.511**	**3.186**

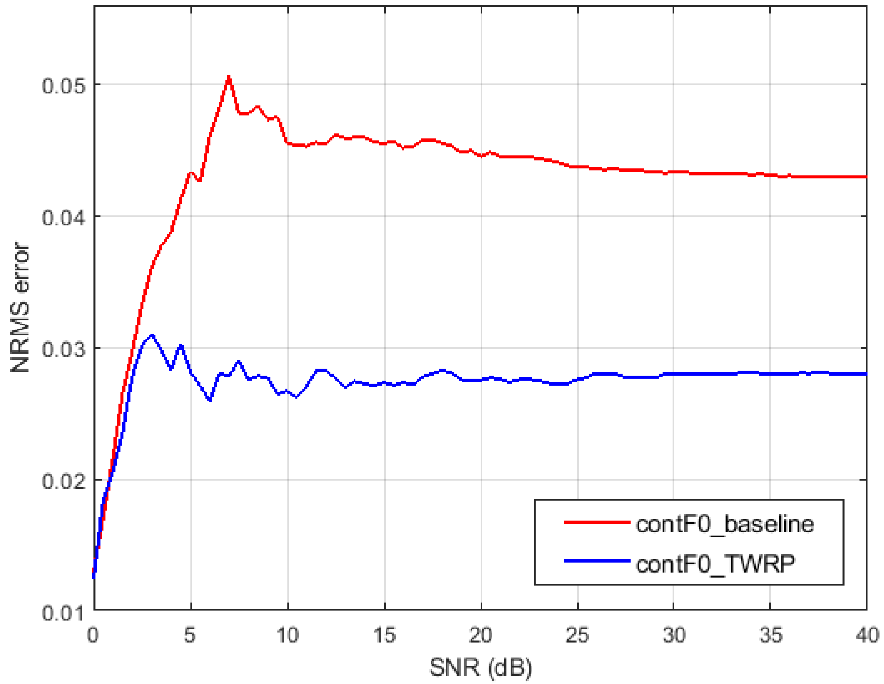


Influence of the SNR on the Average normalized RMSE (white noise).

A) ContF0 Evaluation

Pink noise

Method	GPE %			MFPE			STD		
	BDL	JMK	SLT	BDL	JMK	SLT	BDL	JMK	SLT
baseline	25.041	26.870	33.124	2.919	2.799	2.845	3.061	2.936	3.180
contF0_TWRP	**21.512**	**22.329**	**29.893**	**2.256**	**2.482**	**2.472**	**2.253**	**2.702**	**2.787**



Influence of the SNR on the Average normalized RMSE (pink noise)

B) Objective evaluation of BiLSTM & CSM vocoder

Metrics	Model	AWB	SLT
LLR	Baseline	1.4309	1.6966
	Proposed	**1.4178**	**1.6791**
	WORLD	1.5008	1.7516
$fwSNR_{seg}$	Baseline	2.514	1.1882
	Proposed	2.4972	**1.2278**
	WORLD	**2.5802**	0.81389
LSD	Baseline	**2.0739**	**2.2254**
	Proposed	2.0995	2.2391
	WORLD	2.108	2.3373

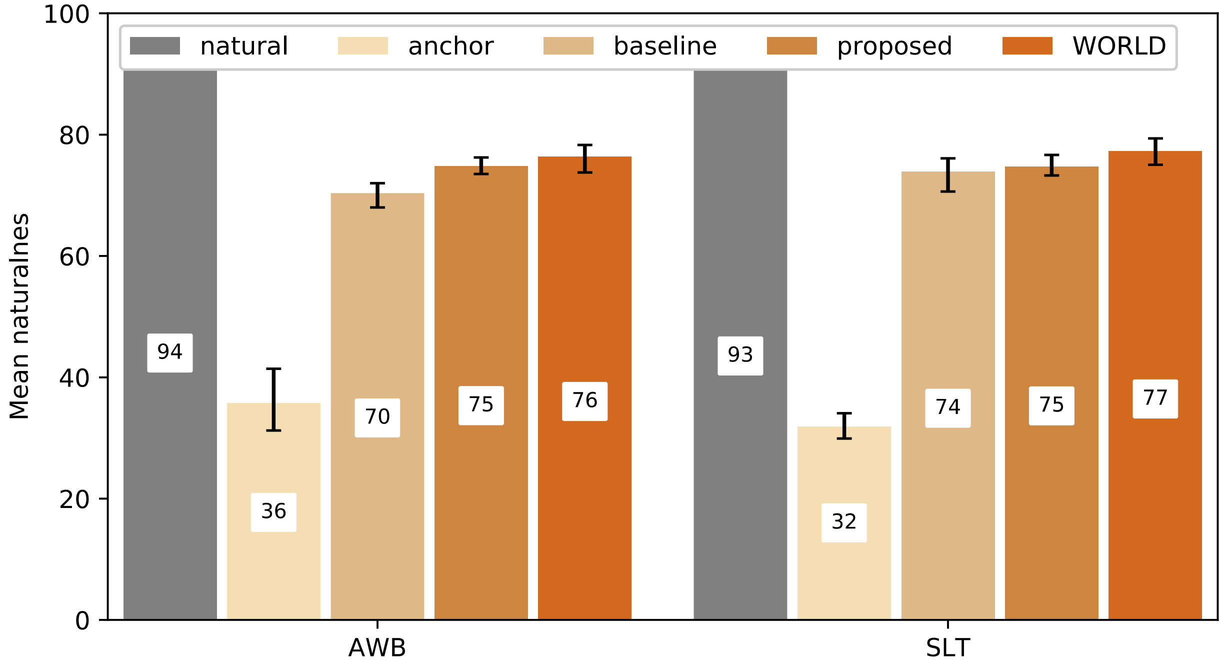
- The proposed vocoder based sinusoidal model succeeded in the Bi-LSTM training.
- CSM framework provides satisfactory results in terms of naturalness and intelligibility comparable to the high-quality WORLD.

C) Subjective evaluation of BiLSTM & CSM vocoder

- MUSHRA: enables evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons.
- reference: natural speech
- anchor: pulse-noise excitation
- 100 utterances were included in the test (2 speaker x 5 types x 10 sentences)
- 13 participants (5 males, 8 females) with an age range of 24-38 years
- The test took 17 minutes to fill

http://smartlab.tmit.bme.hu/ijcnn2019_vocoder

C) Subjective evaluation of BiLSTM & CSM vocoder



- WORLD was slightly preferred over CSM (not significant).
- This means that CSM based RNN-TTS is closer to the level of the state-of-the-art high quality vocoder than the baseline system.

Summary and Future plans

- ✓ Continuous Sinusoidal Model (CSM) generates higher speech quality.

- ✓ The proposed vocoder was not found to be significantly different from the WORLD system.

- ✓ Continuous vocoder has fewer parameters
    - computationally feasible
    - suitable for real-time operation

- ✓ For future work, the authors plan to apply the proposed sinusoidal model into voice conversion

MŰEGYETEM 1782

BME TMIT

Thank you for your attention !

csapot@tmit.bme.hu

Key references

- Garner, P. N., Cernak, M., and Motlicek, P., "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 102-105, 2013.
- Drugman, T., and Stylianou, Y., "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," IEEE Signal Processing Letters, vol. 21, no. 10, p. pp. 1230–1234, 2014.
- Morise, M., “CheapTrick, a spectral envelope estimator for high-quality speech synthesis,” Speech communication, 67, pp. 1-7, 2015.
- Kominek, J., and Black, A.W., "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.
- Wu, Z., Watts, O., King, S., "Merlin: An Open Source Neural Network Speech Synthesis System" in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA.

SmartLab

Intelligent Interactions

http://smartlab.tmit.bme.hu

NVIDIA

GPU EDUCATION CENTER