# A Continuous Vocoder using Sinusoidal Model for Statistical Parametric Speech Synthesis

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

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary

{malradhi,csapot,nemeth}@tmit.bme.hu

SPECOM 2018
Leipzig, Germany

September 20, 2018

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

NVIDIA GPU EDUCATION CENTER

# Outline

- **Background**
  - Speech synthesis
  - Vocoder
- **Reseach goal**
  - Hypotheses
- **Methodology**
  - Continuous sinsoidal model
  - Smoothing of the contF0
- **Experimental evaluation**
  - Objective test
  - Subjective test
- **Future research direction and conclusion**

# Background

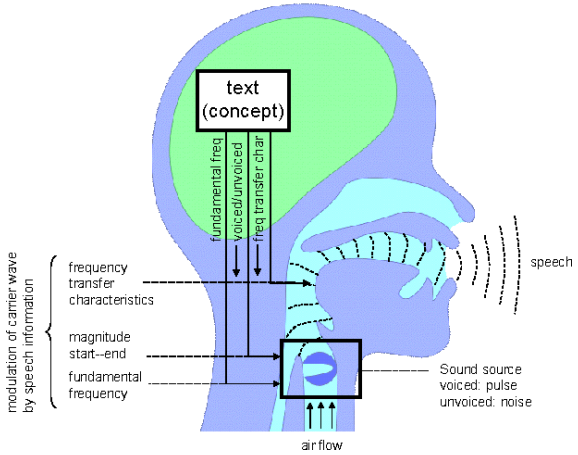
- **Speech synthesis** is the computer generated human speech.

- **Vocoder**
  - is a category of voice codec that analyzes and synthesizes the human voice signal.
  - provide a parametric representation of the speech signal suitable for coding and statistics

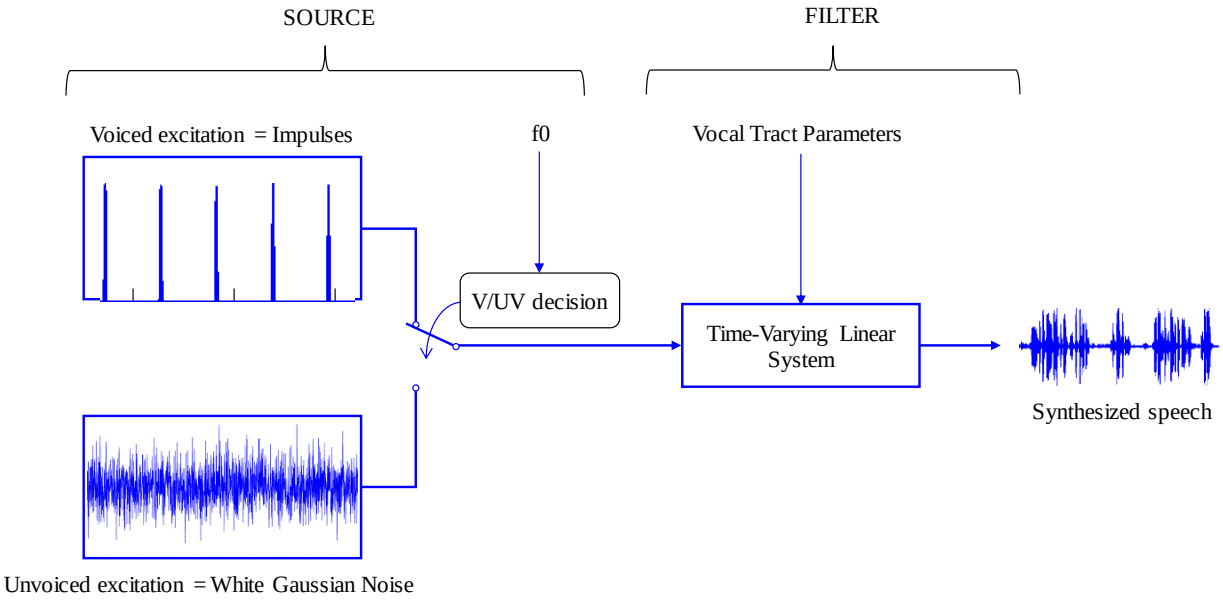
- **Text-to-speech synthesis (TTS)**
  - Generating speech waveform from textual input
  - Transmit data from a machine to a human user

- **Statistical Parametric Speech Synthesis**
  - Flexibility to change voice characteristics
  - Store statistics rather than waveforms
  - Smoothness and style adaptation

# Simplified vocoder of speech production




Human speech production




Basic pulse-noise excitation

- Voiced source: Generated by vocal cords vibrations, periodic, F0
- Unvoiced source: Generated without vibrations, no pitch

# 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

# Challenges

- Despite large improvements, Speech synthesis can still sound a little unnatural.

- Natural speech need considerable resources in terms of data storage and processing power.

- STRAIGHT is the most widely used vocoder 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.

# Goal

There is a need for simple and computationally feasible algorithms:

- to construct a vocoder whose parameters can be controlled to achieve high quality synthesized speech.

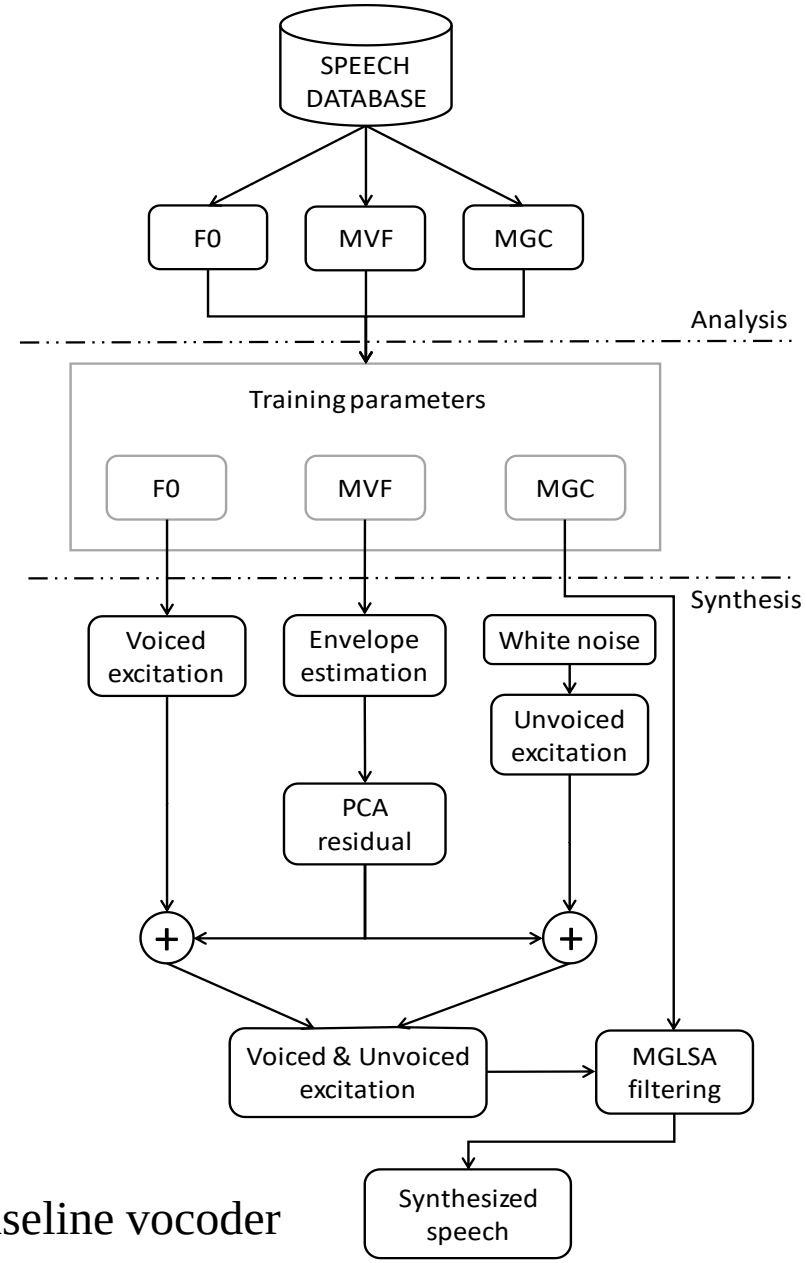
# Baseline

# Continuous vocoder

# 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



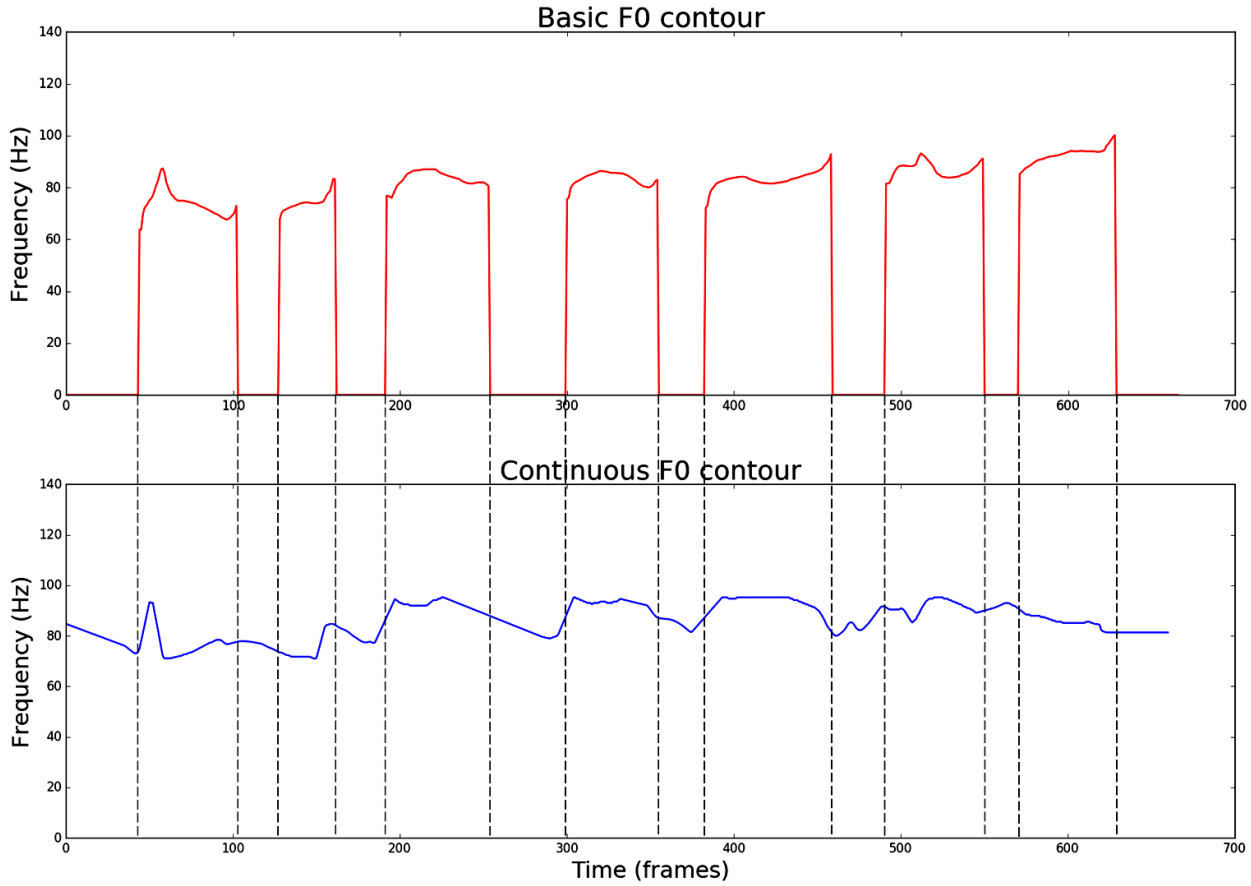

Workflow of the baseline vocoder

# Continuous vocoder

- 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

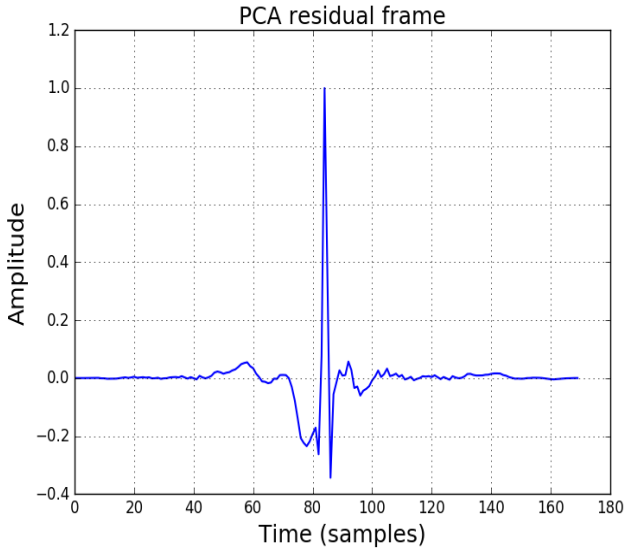
# Continuous vocoder

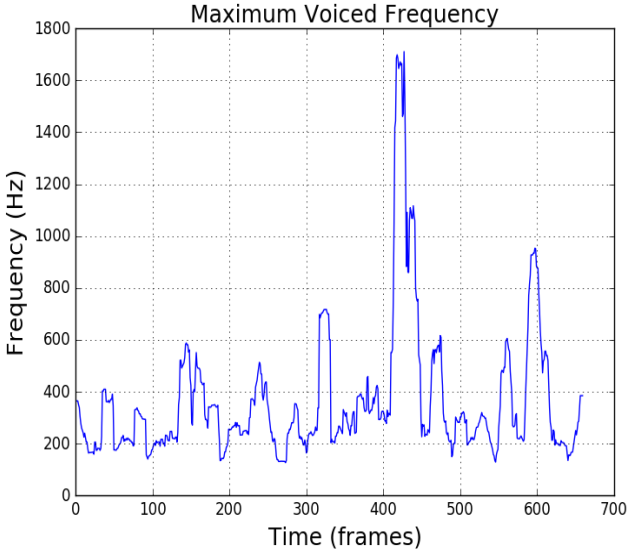



*"The girl faced him, her eyes shining with sudden fear."*

# Continuous vocoder

- 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."*

# Problem formulation

- There is a lack of naturalness and still not achieving a high-quality speech synthesis compared to the well-known vocoders (e.g. STRAIGHT or WORLD).

- The estimated contF0 contours have high unwanted voiced component in the unvoiced speech sounds.

# Hypotheses

- Using sinusoidal synthesis model that is applicable in statistical frameworks will be superior to the baseline vocoder (source-filter model).

- Smoothing the estimated contF0 algorithm by a post-processing phase will eliminate octave errors and isolated glitches.

# Proposed vocoder

## Continuous Sinusoidal Model (CSM)

# Continuous Sinusoidal Model (CSM)

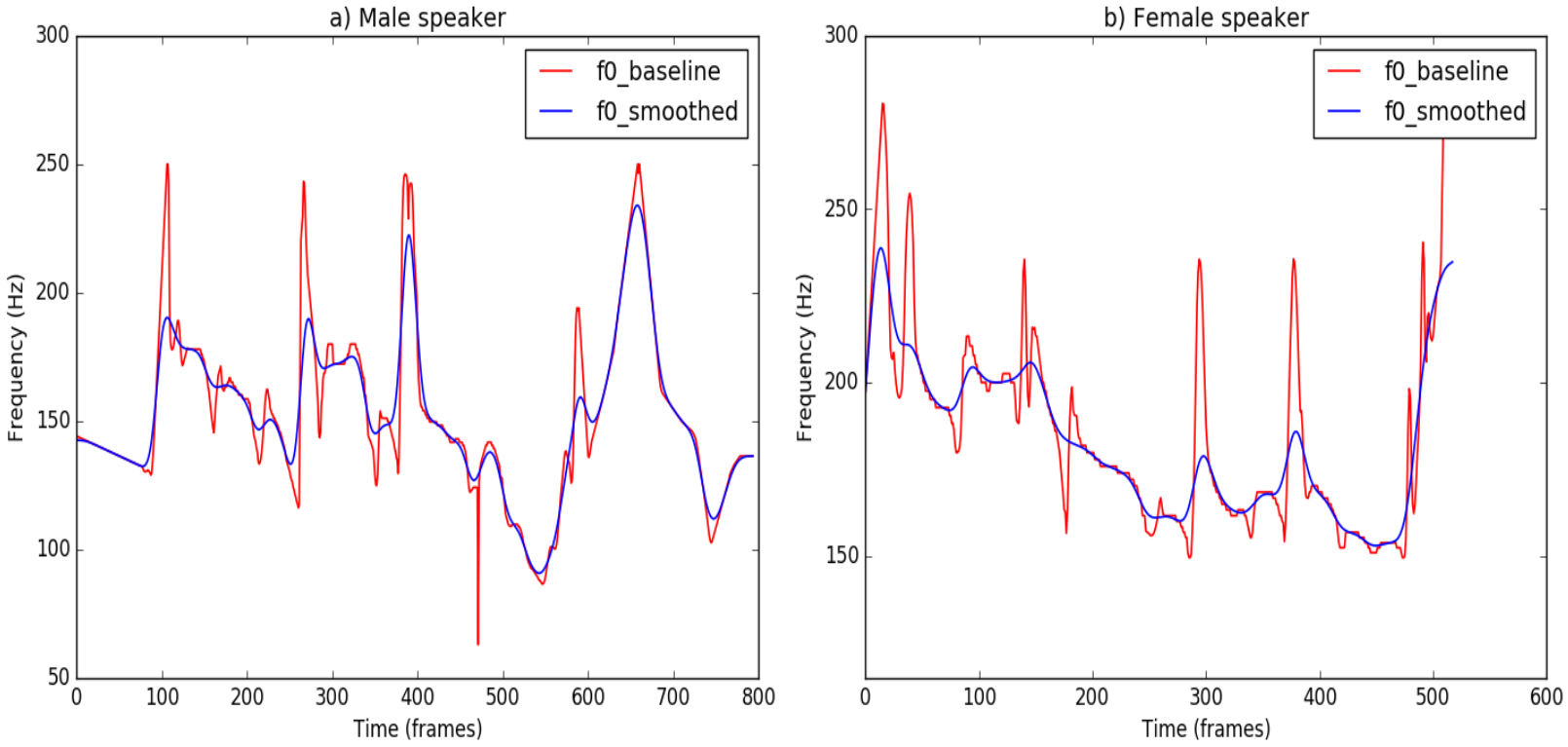
## ➢ Analysis step:

Three parameters of the analysis part from our previous source-filter model have been also extracted and used for this study.

1. contF0: Pitch values to have a small amount of noise variation if they are not estimated well
   - leads to extra buzziness.
   - undependable feature measurements.
2. MVF
3. MGC

# Smoothing of the contF0 contour

- Combine two smoothing steps in contF0
  - median filter using 0.1s window to ignore isolated outliers while preserving both the fine-grained variations and the sharpness of true step transitions
  - linear smoother (zero-phase filtering) with Hanning window is applied to remove higher-frequency resonance effects and hence suppress the noisiness of the measurement

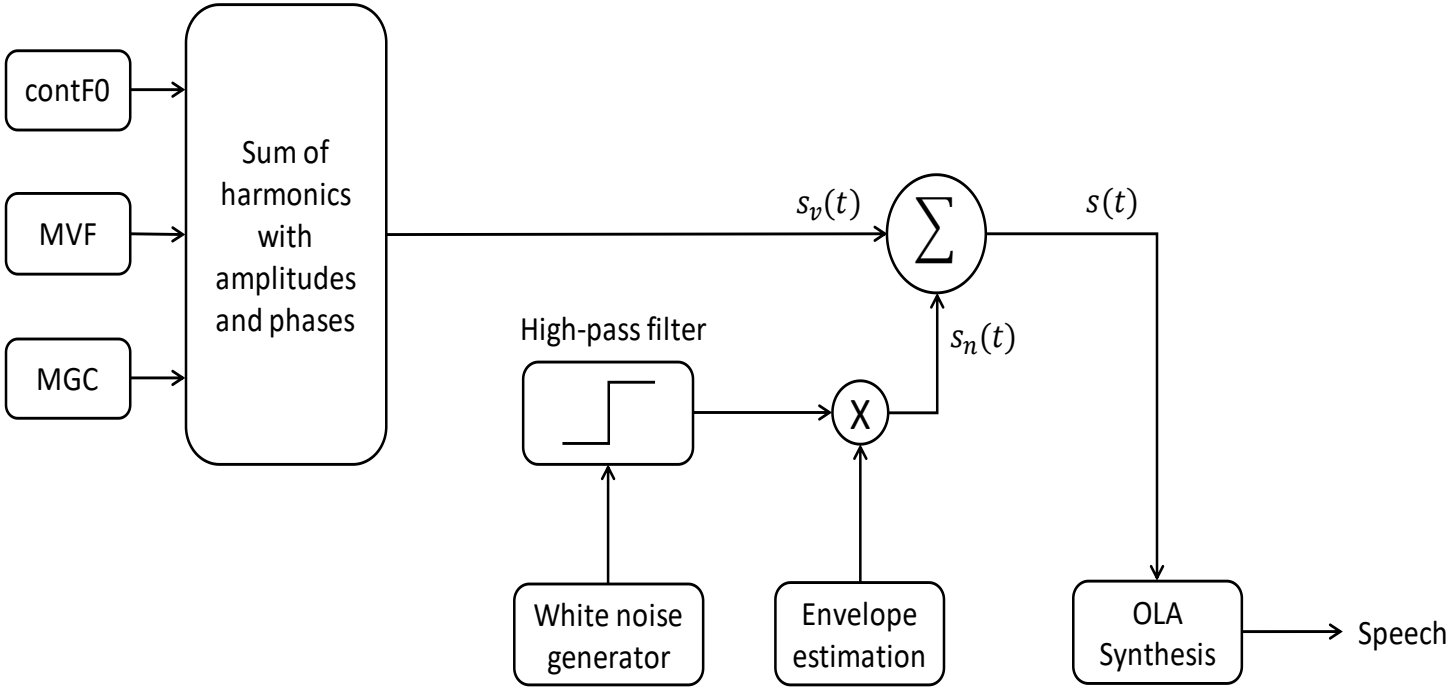



"Sometimes her dreams were filled with visions."

# Continuous Sinusoidal Model (CSM)

## ➢ Sinusoidal synthesis:

The novelty behind this step is to decomposes the speech frames into a harmonic/voiced component lower band and a stochastic/noise component upper band based on MVF values.



# 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 $i$, $t = 0, 1, \ldots, N$ 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 $e(t)$ [Al-Radhi et al., 2017]. For unvoiced frames, the harmonic part is obviously zero and the synthetic frame is typically equal to the generated noise.

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

# Evaluation

# Experimental Conditions

- English speaker from CMU-ARCTIC database [Kominek and Black, 2003]
  - SLT (American English, female)
  - AWB (Scottish English, male)
- Waveform sampling rate of the database is 16 kHz
- 100 sentences from each speaker were analyzed and synthesized with below vocoders:
  - Baseline
  - Proposed
  - STRAIGHT
  - WORLD
- Metrics:
  - Itakura-Saito distance
  - frequency-weighted segmental SNR
  - Extended Short-Time Objective Intelligibility

# A) Objective evaluation

| Vocoder | IS | | fwSNRseg | | ESTOI | |
|---|---|---|---|---|---|---|
| | AWB | SLT | AWB | SLT | AWB | SLT |
| Baseline | 0.148 | 0.447 | 6.987 | 7.940 | 0.517 | 0.676 |
| Proposed (CSM) | **0.058** | **0.082** | **9.560** | **11.034** | **0.749** | **0.867** |
| WORLD | **0.016** | **0.014** | **13.312** | 13.336 | **0.808** | **0.951** |
| TANDEM-STRAIGHT | 0.065 | 0.042 | 11.840 | **14.641** | 0.772 | 0.933 |

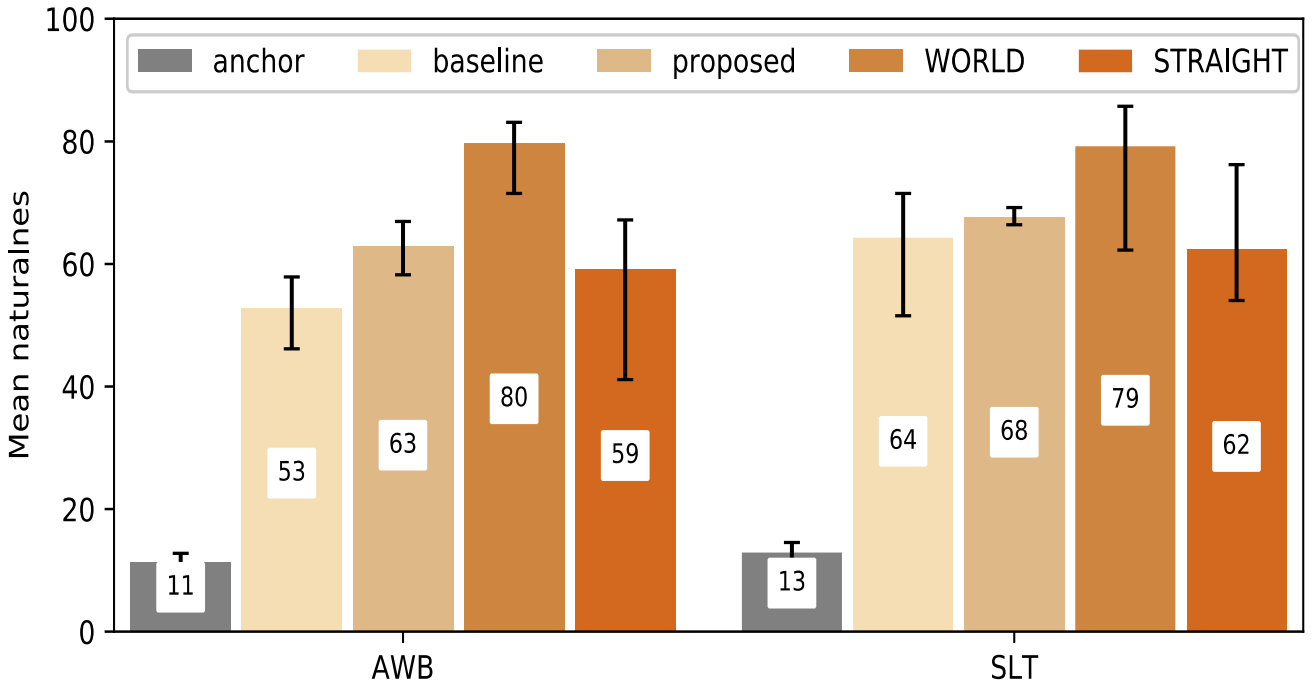
| Vocoder | Parameter per frame | Excitation |
|---|---|---|
| Proposed (CSM) | F0: 1 + MVF: 1 + MGC: 24 | Mixed |
| WORLD | F0: 1 + Band aperiodicity: 5 + MGC: 60 | Mixed |
| TANDEM-STRAIGHT | F0: 1 + Aperiodicity: 2048 + Spectrum: 2048 | Mixed |

- Continuous Sinusoidal Model (CSM) has few parameters and is computationally feasible; therefore, it is suitable for real-time operation.

# B) Subjective evaluation

- 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
- 60 utterances were included in the test (2 speaker x 6 types x 5 sentences)
- 13 participants (7 males, 6 females) with an age range of 20-42 years
- The test took 15 minutes to fill

# B) Subjective evaluation




- CSM was slightly preferred over TANDEM-STRAIGHT (not significant), showing that the sinusoidal extension of our vocoder is similar to state-of-the-art high quality vocoders.

Online samples:
http://smartlab.tmit.bme.hu/specom2018

# Summary and Future plans

- ✓ Continuous Sinusoidal Model (CSM) generates higher output speech quality.
- ✓ The proposed vocoder were preferred over TANDEM-STRAIGHT.
- ✓ Continuous vocoder has fewer parameters
  - computationally feasible
  - suitable for real-time operation
- ✓ For future work, the authors plan to train and evaluate all continuous parameters (F0, MVF, and MGC) using deep learning algorithm such as feed-forward and recur-rent neural networks to test the proposed vocoder in SPSS.

# Key reference


- 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.
- Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," Interspeech, vol. Stockholm, no. , pp. 434-438, 2017.
- Tokuda K., Kobayashi T. ,Masuko T., and Imai S., "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in Proc. of the ICSLP, p. 1043–1046, 1994.
- Kominek, J., and Black, A.W., "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.
- Ma J., Hu Y., and Loizou P., "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," Acoustical Society of America, vol. 125, no. 5, pp. 3387-3405, 2009.

# Thank you very much for your attention !

malradhi@tmit.bme.hu

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

NVIDIA
GPU
EDUCATION
CENTER