# Continuous Wavelet Vocoder-based Decomposition of Parametric Speech Waveform Synthesis

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

malradhi@tmit.bme.hu

Budapest University of Technology and Economics
Budapest, Hungary

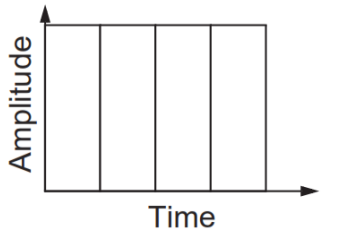
September 01, 2021

# Motivation

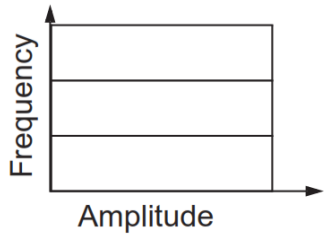
- Fourier Transform decomposes a signal into infinite length sines and cosines.
  - losing all time-localization information.

- Short-Time Fourier Transform (STFT) have a fixed width.
  - Can’t vary the window size to determine accurately either time or frequency.

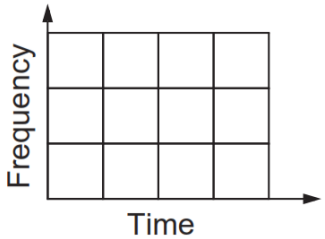
- Wavelet Analysis breaking up of a signal into shifted, shrinked, and scaled function.
  - windowing technique with variable-sized regions.



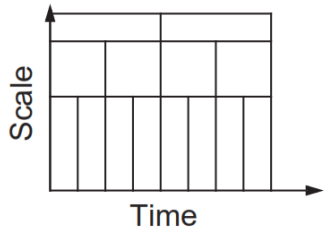
**Time Domain (Shannon)**



**Frequency Domain (Fourier)**



**STFT (Gabor)**



**Wavelet Analysis**

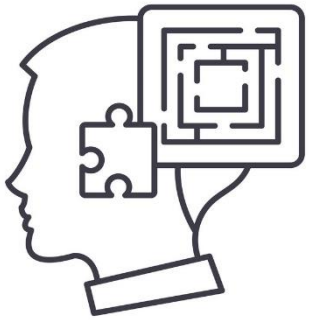
# Problem formulation

- Source-filter models
  - over-smoothed spectra
  - buzzy synthesized TTS

- Neural models
  - large quantity of voice data
  - difficult to use in real-time

# In this study …

- present an updated synthesizer to:
  - characterize and decompose speech features
  - retain the fine fundamental frequency
  - generate natural-sounding synthetic speech



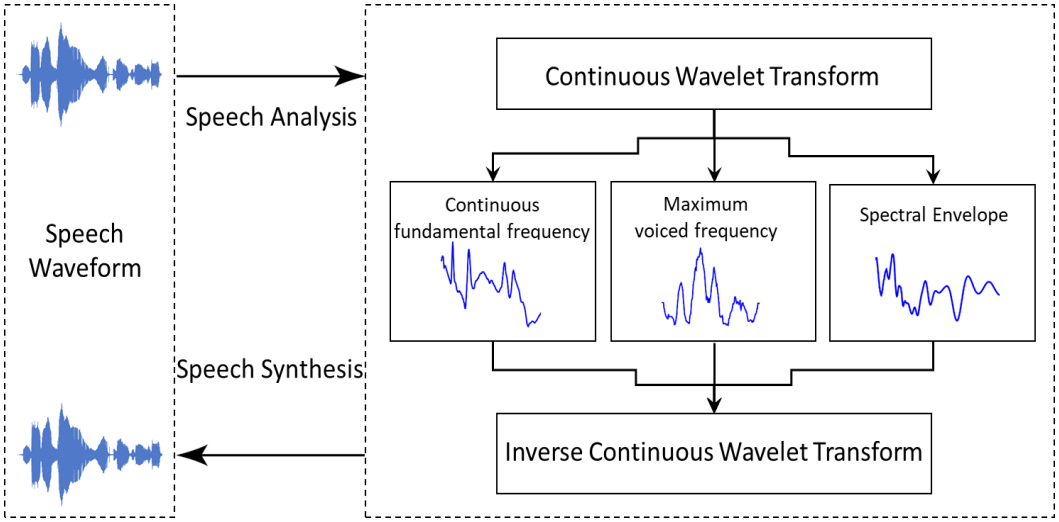
# Methodology

## ➢ Continuous Wavelet Transform (CWT)

- ❑ It is the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet.

$$C(scale, position) = \int_{-\infty}^{\infty} f(t)\psi(scale, position, t)dt$$

- ❑ decomposes a multi-level representation of contF0, MVF, and spectral envelope.



# Experimental conditions

## Speech Corpus

**English** speaker from CMU-ARCTIC database [Kominek and Black, 2003]

- 4 male and 2 female
- 1132 sentences with sampling rate 16 kHz

## Reference Systems

- WaveNet [Oord et al., 2016]
- WORLD [Morise et al., 2016]
- Continuous [Al-Radhi et al., 2017]
- Anchor

# Results

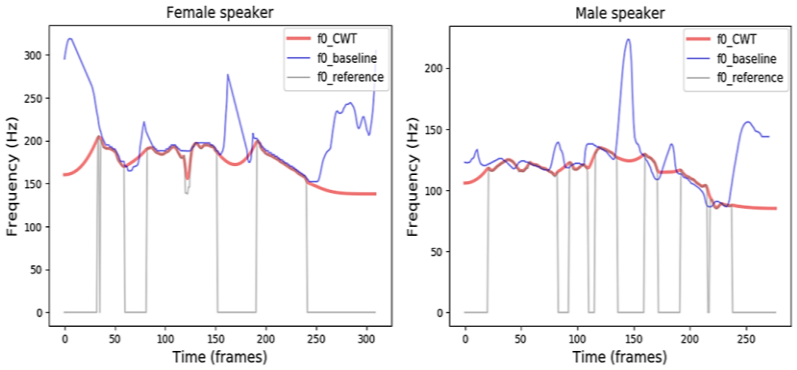
- mel-cepstrum distortion

| MCD (dB) | Male | Female |
|---|---|---|
| Baseline | 4.03 | 4.13 |
| WaveNet | 4.74 | 4.97 |
| WORLD | 3.31 | 3.27 |
| Proposed | 3.47 | 3.42 |

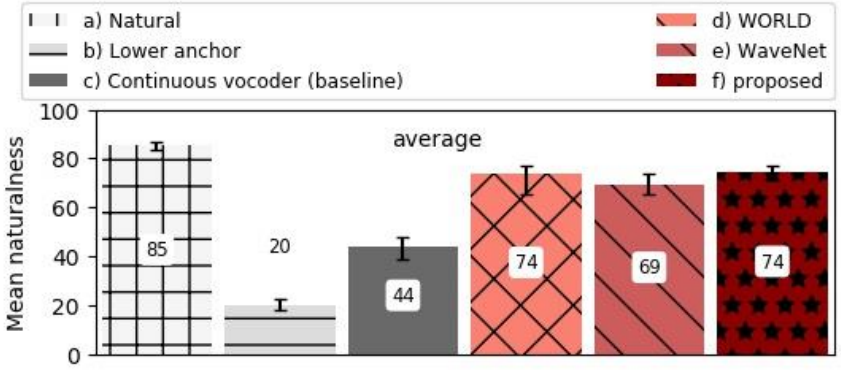
- F0 root mean square error

| RMSE (dB) | Male | Female |
|---|---|---|
| Baseline | 4.37 | 4.31 |
| WaveNet | 4.14 | 4.67 |
| WORLD | 3.42 | 3.51 |
| Proposed | 3.85 | 3.98 |

- continuous F0 estimated by CWT




- sound quality of synthesized speech



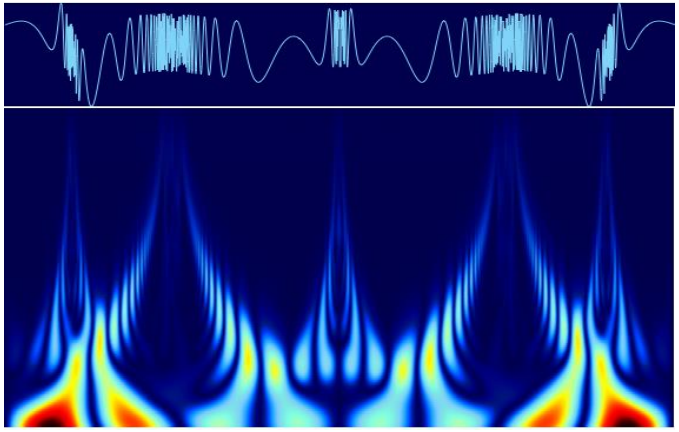

- Samples https://malradhi.github.io/cwt_vocoder/

# Summary

- ✓ Synthetic speech was produced with continuous wavelet transform technique.
- ✓ WaveNet model did not perform well with CMU-ARCTIC corpus (tested with 6 hours of recorded speech).
- ✓ Proposed system was able to generate a natural-sounding synthetic speech and superior to WaveNet vocoder.

# We'd love to talk to you!

malradhi@tmit.bme.hu



**Wavelet Vocoder**

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

NVIDIA
GPU
EDUCATION
CENTER