Towards Parametric Speech Synthesis Using Gaussian-Markov Model of Spectral Envelope and Wavelet-Based Decomposition of F0

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

malradhi@tmit.bme.hu

30th EUROPEAN SIGNAL PROCESSING CONFERENCE
EUSIPCO
BELGRADE/SERBIA
29 AUG – 2 SEP 2022

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

SIGNAL PROCESSING NEVER SLEEPS
EUSIPCO 2022
30th European Signal Processing Conference
BELGRADE // SERBIA

Motivation

Concatenative speech synthesis

• High intelligibility, but requires huge database, less natural and emotionless

Statistical parametric speech synthesis

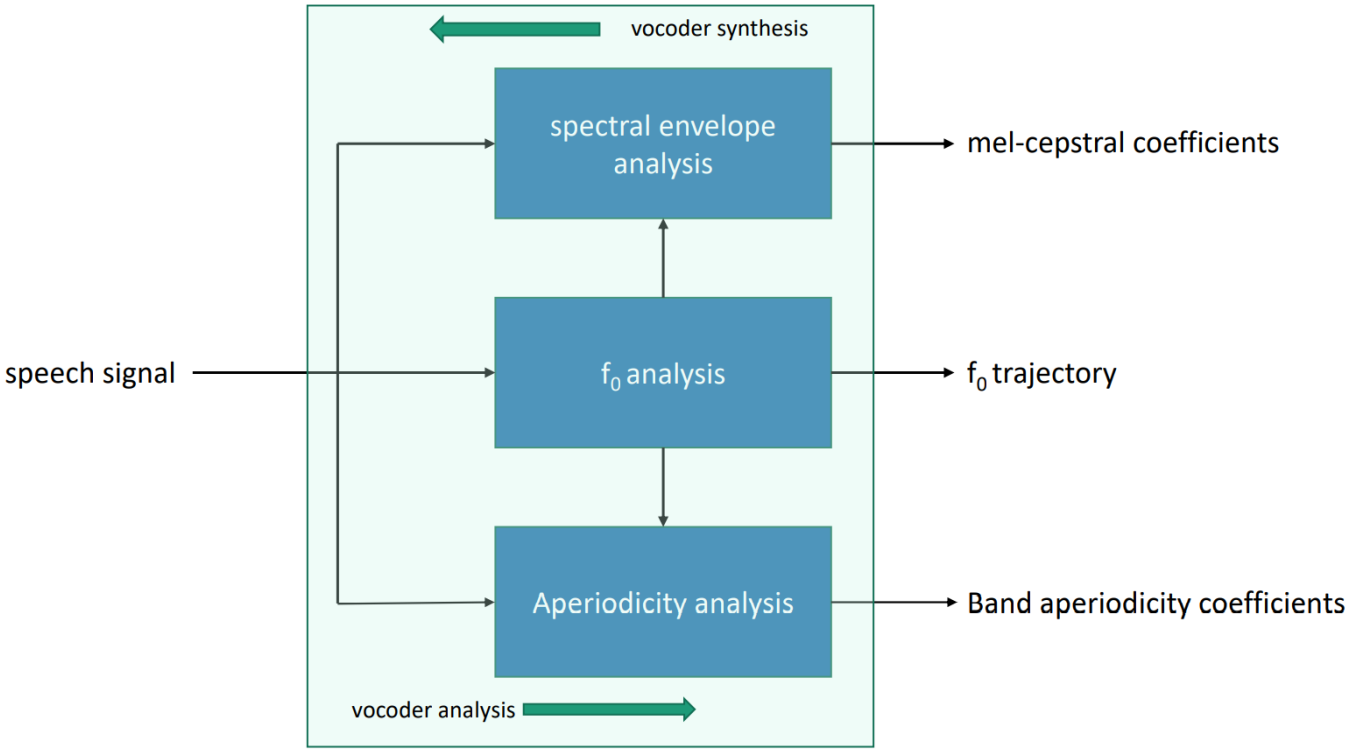
• Lower data cost and more flexible, but lower quality and robotic

Neural network based end-to-end speech synthesis

• Huge quality improvement, less human preprocessing and feature development

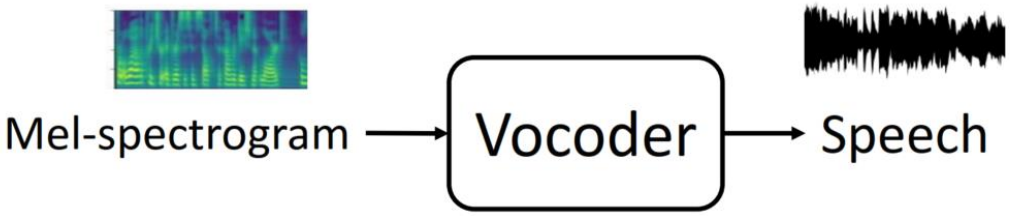
Vocoder

- Statistical parametric speech synthesis
  - STRAIGHT, Phase vocoder, PSOLA, sinusoidal model, Continuous, WORLD



Vocoder

- Neural vocoder
  - WaveNet, ParallelWaveNet
  - SampleRNN, WaveRNN, LPCNet
  - GAN-based model
  - Flow-based model
  - Diffusion-based model



Problem formulation

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

- Neural models
  - large quantity of voice data
  - Too many adaptation parameters
  - difficult to use in real-time

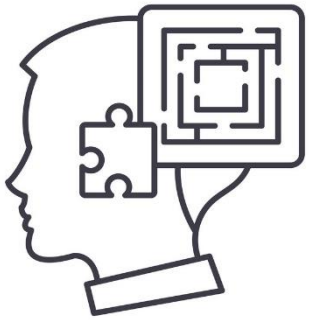
- Challenges

  - The parameters need to be small enough for each speaker to reduce memory usage while maintaining high voice quality
    - e.g., each user/voice with 100MB, 1M users, total memory storage = 100PB! [Xu-Tan, 2021]

In this study …

Gaussian-Markov model

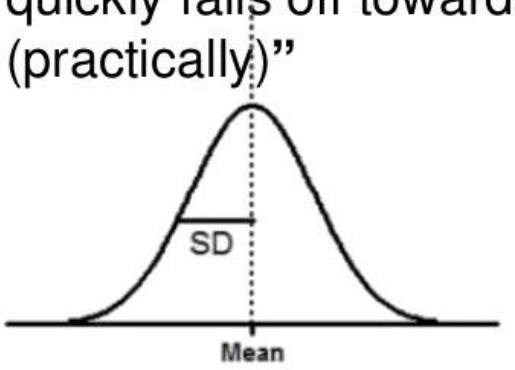
- present an updated synthesizer to:
  - characterize and decompose speech features
  - retain the fine spectral envelope and fundamental frequency
  - generate natural-sounding synthetic speech



Background

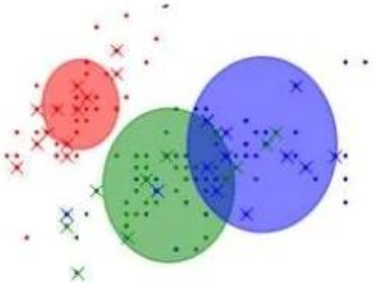
- Gaussian

"Gaussian is a characteristic symmetric "bell curve" shape that quickly falls off towards 0 (practically)"



- Mixture Model

"mixture model is a probabilistic model which assumes the underlying data to belong to a mixture distribution"



Background

➢ Gaussian Mixture Modeling (GMM)

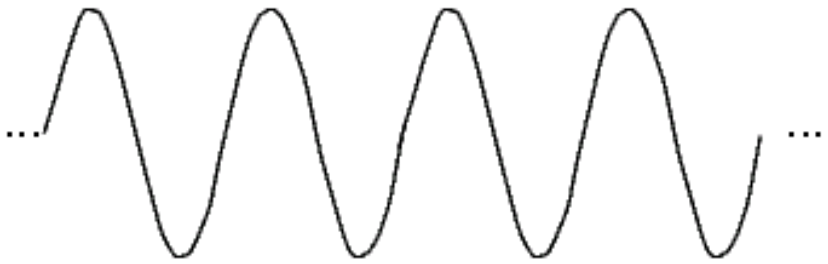
- Provides parametric representation of smoothened spectra.
- Can be used to extract formant like features
  - Gaussian mean → formant frequency, amplitude → formant peak, variance → formant bandwidth.
- Abrupt spectral variations results in abrupt variations in Gaussian parameters.

Background

- Wavelet transform

"is a tool that cuts up data, functions or operators into different frequency components, and then studies each component with a resolution matched to its scale"

Dr. Ingrid Daubechies, Lucent, Princeton U.



Sine Wave

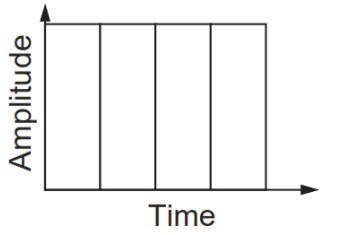
Wavelet (db10)

Background

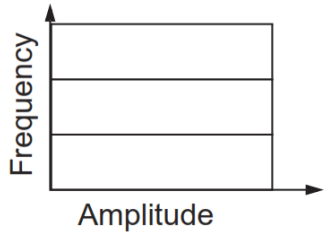
- Fourier Transform decomposes a signal into infinite length sines and cosines.
  - losing all time-localization information.

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

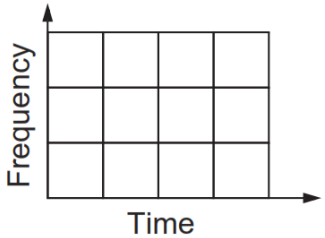
- Wavelet Analysis breaking up of a signal into shifted, shrinked, and scaled function.
  - windowing technique with variable-sized regions.



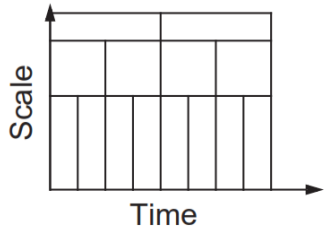
Time Domain (Shannon)



Frequency Domain (Fourier)



STFT (Gabor)



Wavelet Analysis

Methodology

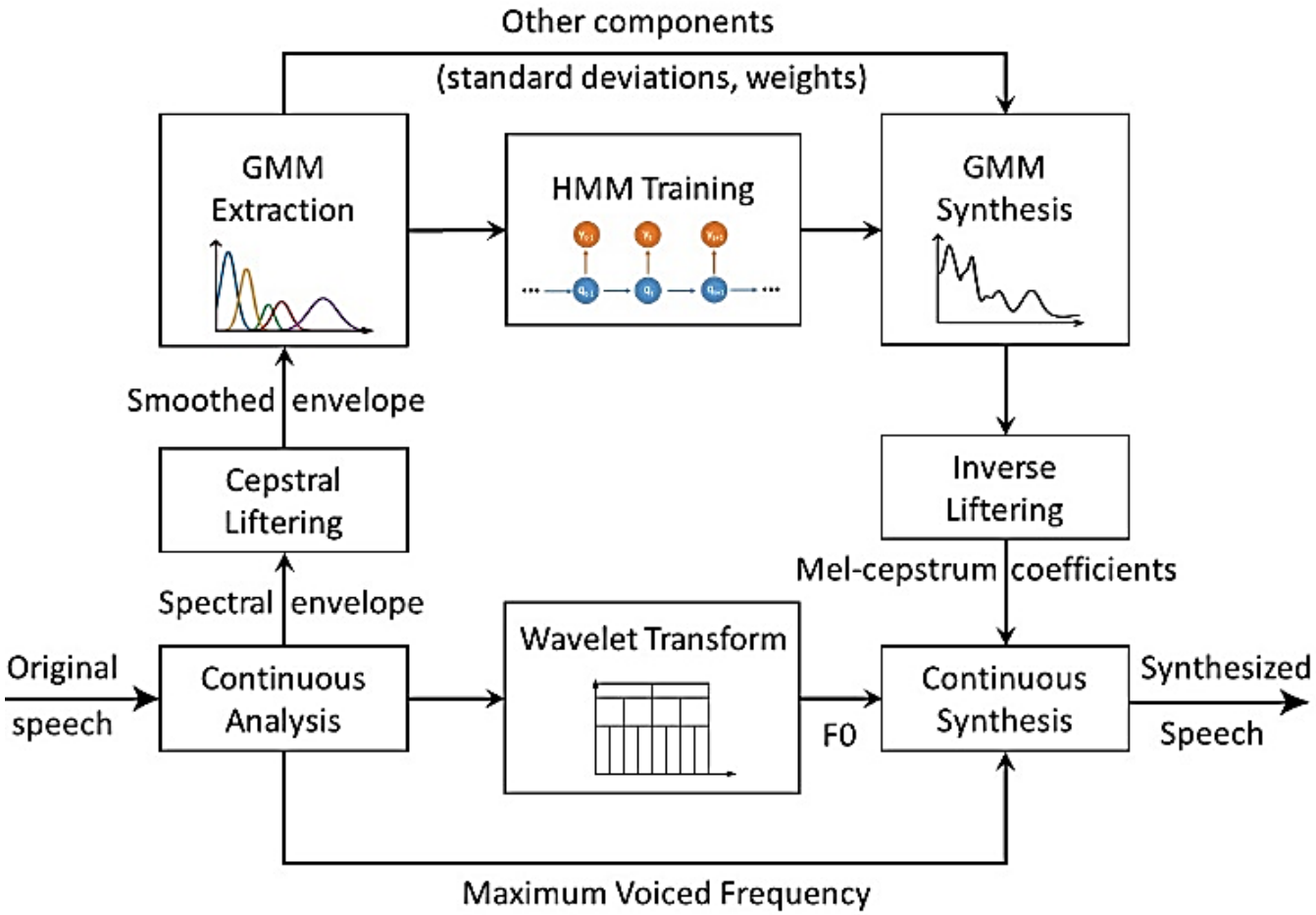
➢ Spectral Envelope Approximation with Gaussian-Markov Model

1. extract spectral envelope using CheapTrick algorithm [Morise, 2015].
2. cepstral liftering of the magnitude spectra is applied to remove the unwanted high-quefrency effects of the excitation from the spectrum.
   - This can yield better fits and smoother formant trajectories.
3. approximate the modified spectral envelope with a GMM [Nguyen and Akagi, 2009].
   - GMM parameters are estimated by minimizing a loss function of the observed spectral envelope $H(\omega)$, and the GMM $G(\omega)$ expressed by

$$G(\omega) = \sum_{k=1}^{K} \frac{w_k}{\sqrt{2\pi\sigma_k^2}} exp\left[-\frac{(\omega - \mu_k)^2}{2\sigma_k^2}\right]$$

4. apply HMM to provide an effective framework for modelling time-varying spectral vector sequences
   - HMM is fit using the Baum-Welch algorithm and decoded using the Viterbi algorithm

Methodology



Overview of proposed analysis-synthesis framework using GMM and CWT-based approximation of speech features

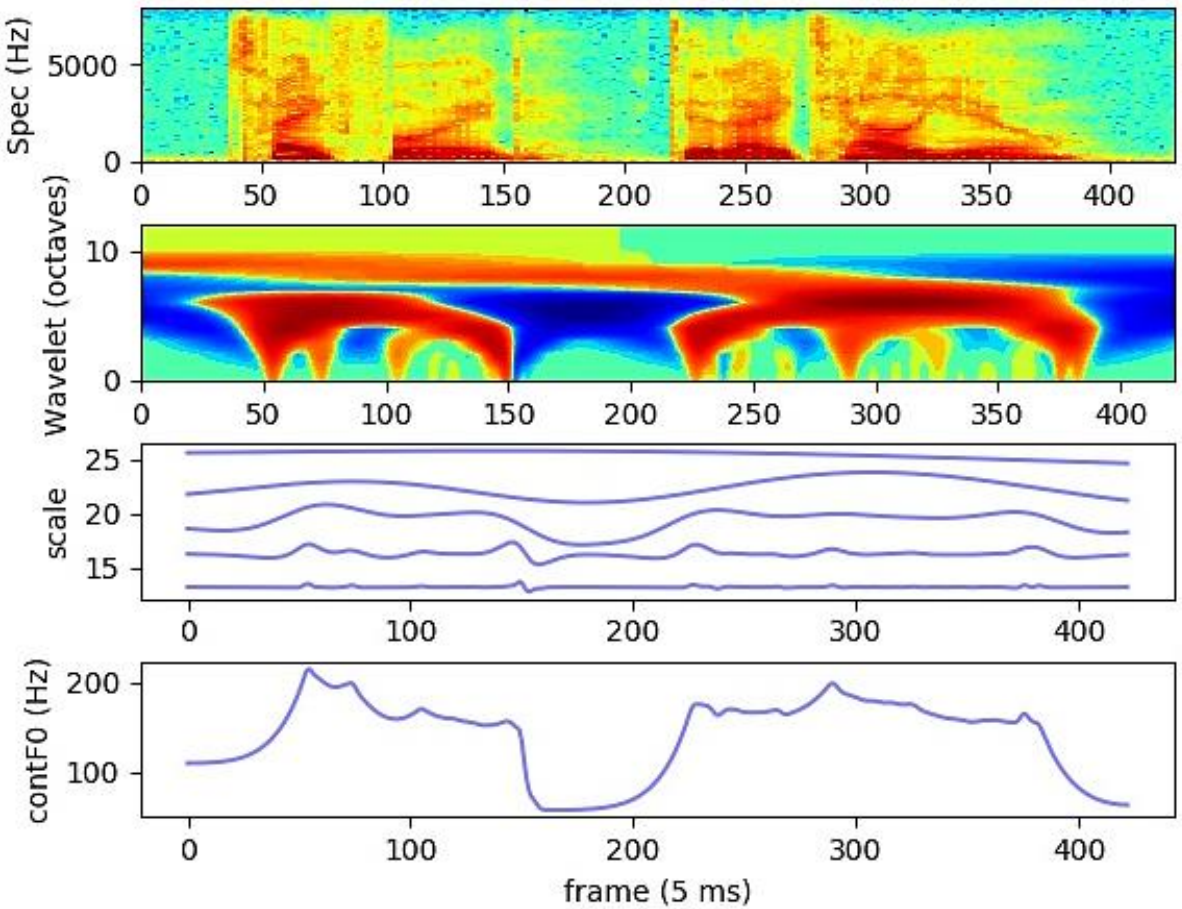
Methodology

➢ Wavelet-Based Decomposition of Continuous F0

- F0 is not well defined for the unvoiced segments of the speech and the silent intervals
  - this makes the direct wavelet analysis impossible.
  - the continuous F0 is used in this study.
- applied a continuous pitch estimation algorithm [Garner et al., 2013] which used:
  - Bayesian approach that naturally yields estimates for unvoiced segments, along with variances for all estimates
  - Kalman smoother to the sequence of estimates and variances to give a sequence of pitch estimates.
- applied Continuous Wavelet Transform (CWT) [Al-Radhi et al., 2021]
  - This leads to high-frequency resolution with CWT at low frequencies and high time resolution at high frequencies.

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

Methodology



Top pane shows a spectrogram of the speech signal, the second pane depicts the continuous wavelet transform with Mexican hat mother wavelet of F0, the third pane shows the scales, and bottom pane gives the modified continuous F0.

Experimental conditions

Speech Corpus

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

- 1 male and 1 female
- 1132 sentences per each speaker

Systems

- WaveNet [Oord et al., 2016]
- WaveRNN [Kalchbrenner et al., 2018]
- NSF [Wang et al., 2019]
- STRAIGHT [Kawahara et al., 1999]
- Continuous [Al-Radhi et al., 2017]
- Anchor

Results

➢ Objective Evaluation

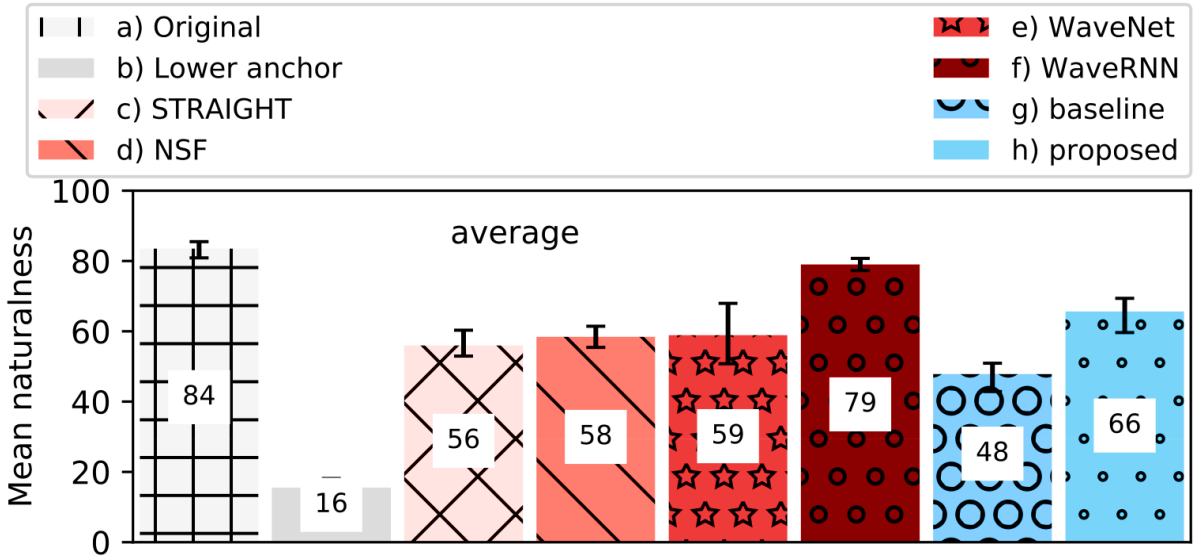
Systems	MCD (dB)	
	Male	**Female**
Baseline (continuous vocoder)	4.086	4.194
STRAIGHT	3.792	3.925
NSF	3.671	3.650
WaveNet	3.785	3.924
WaveRNN	3.428	3.589
Proposed	**3.399**	**3.564**

COMPARISON OF MEL-CEPSTRUM DISTORTION BETWEEN SPECTRAL FEATURES OF NATURAL AND SYNTHESIZED SPEECH.

Results

➢ Listening test

- MUSHRA-like (Multi-Stimulus test with Hidden Reference and Anchor) listening test
- All sentences appeared in randomized order (different for each listener)
  - 10 listeners



http://smartlab.tmit.bme.hu/eusipco2022

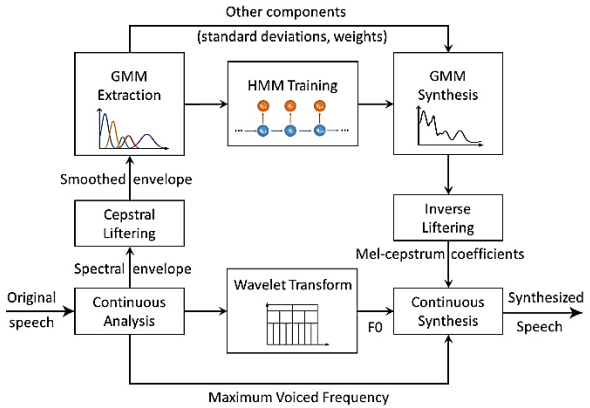
Key Notes and Summary

- TTS technology evolves from concatenative synthesis, statistical parametric synthesis, and neural based end-to-end synthesis.
- Improving the quality while reducing the cost is always the goal of TTS
  - Quality: Intelligibility, naturalness, robustness, expressiveness and controllability
  - Cost: Engineering cost (end-to-end), serving cost (inference speedup), data cost (low resource)
- We use the Gaussian-Markov model toward robust learning of spectral envelope and wavelet-based statistical signal processing to characterize and decompose F0 features.
- Gaussian-Markov model of spectral envelope generate a natural-sounding synthetic speech.
- Wavelet-based decomposition of F0
- The proposed a novel updated vocoder, which is a simple signal model to train and easy to generate waveforms.
- Our method presents a good alternative technique to other systems for the reconstruction of speech.

Thanks for your attention!

ACKNOWLEDGMENT

- APH-ALARM project (contract 2019-2.1.2-NEMZ-2020-00012) funded by the European Commission and the National Research, Development and Innovation Office of Hungary.
- National Laboratory of Infocommunication and Information Technology, carried out by BME and IdomSoft Ltd., supported by the Ministry of Innovation and Technology and the National Research, Development and Innovation Office.
- Bolyai János Research Fellowship of the Hungarian Academy of Sciences and by the ÚNKP-21-5 (identifier: ÚNKP-21-5-BME-352).



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

malradhi@tmit.bme.hu