A novel codebook-based excitation model for use in speech synthesis

Tamás Gábor Csapó, Géza Németh

Budapest University of Technology and Economics, Hungary

CogInfoCom 2012, Kosice
December 2-5

Contents

- CogInfoCom and text-to-speech
- Statistical parametric speech synthesis
- Speech coding
- Novel excitation model
- Conclusions and applications

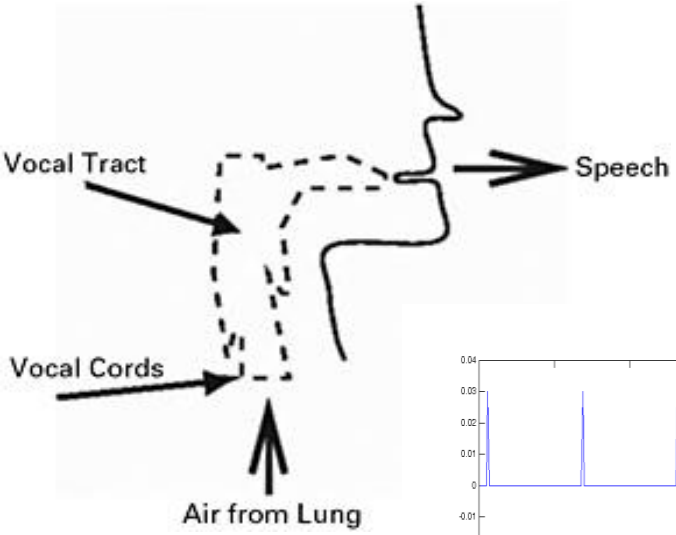
CogInfoCom and text-to-speech

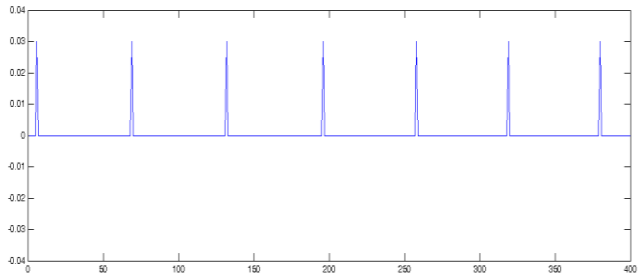
- Speech, text-to-speech synthesis (TTS)
  - one of the main modalities of human-human communication
  - important in human-computer communication
  - natural inter-cognitive sensor-bridging communication mode
  - applications like talking robot, car speech interface and telesurgery
  - helpful for the vision impaired and blind people to access information

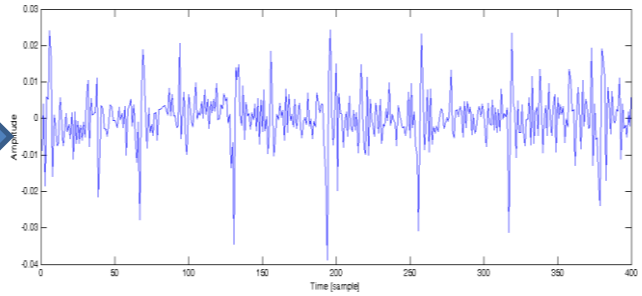
Statistical parametric speech synthesis

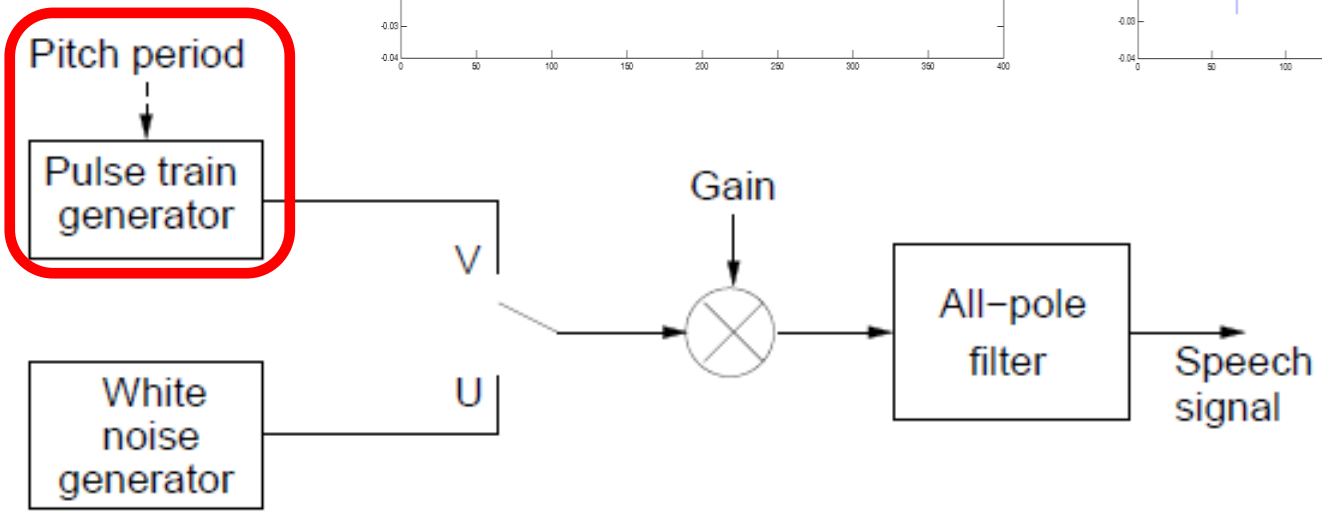
- State-of-the art speech synthesis technique
- Parametric
  - Speech signal is encoded to parameters
  - Parameters are decoded to speech
- Sub-problem: speech coding
  - Speech encoding & decoding in an effective way
  - Speech excitation modeling

Speech excitation models







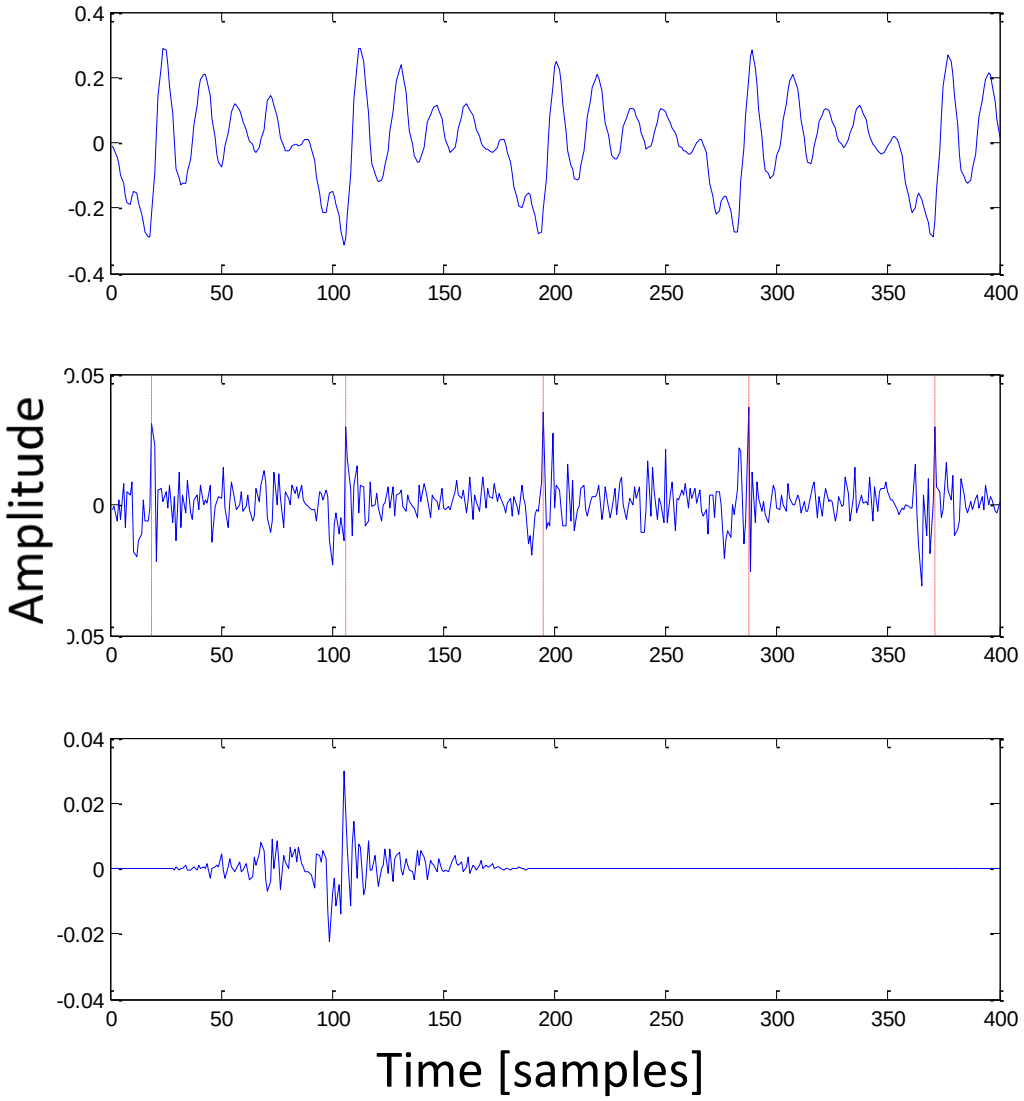


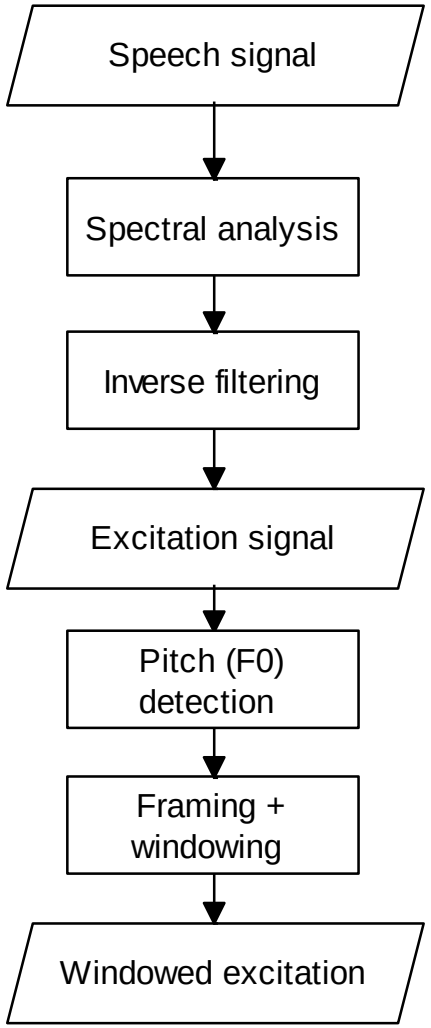
Csapó, Németh: A novel codebook-based excitation model for TTS

Methods

- Speech coding = encoding + decoding
- Novel excitation model
  - Fit to the machine learning in TTS
  - Codebook of excitation frames
  - Phoneme-dependent excitation
  - Flexible
  - Different voice qualities (modal, breathy, whispered)

Speech encoding

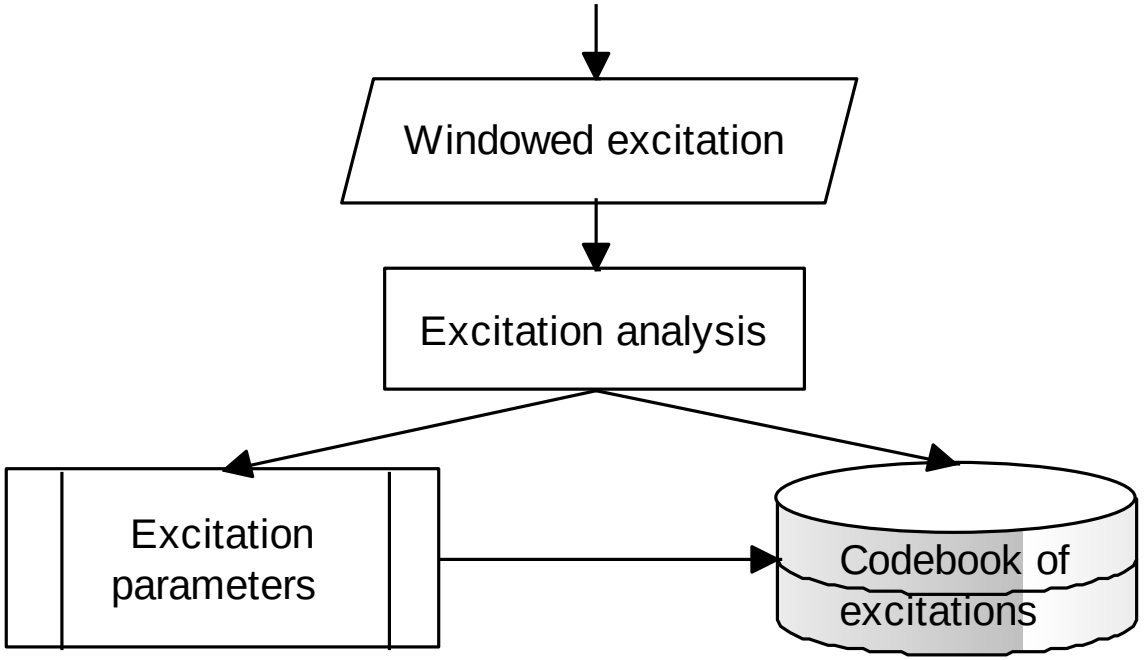


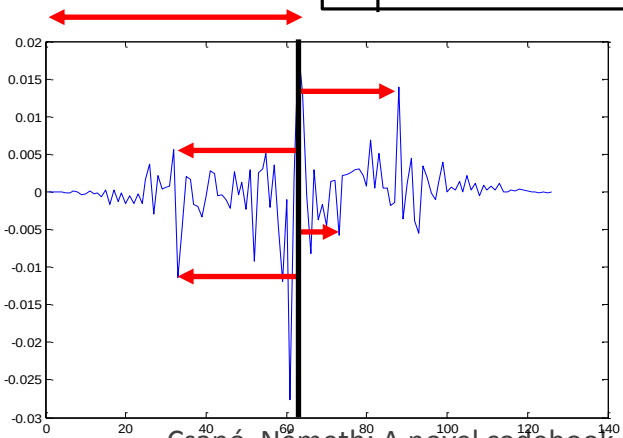


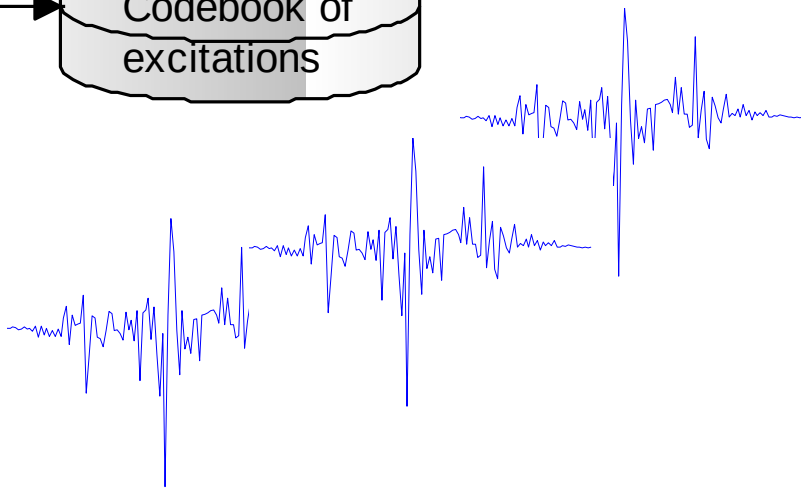
Pitch tracking

- Pitch (F0) detection
  - Snack ESPS method
  - autocorrelation based
- Glottal Closure Instant detection
  - SEDREAMS method
  - detect moments of high energy in the excitation signal
  - highly reliable and noise robust

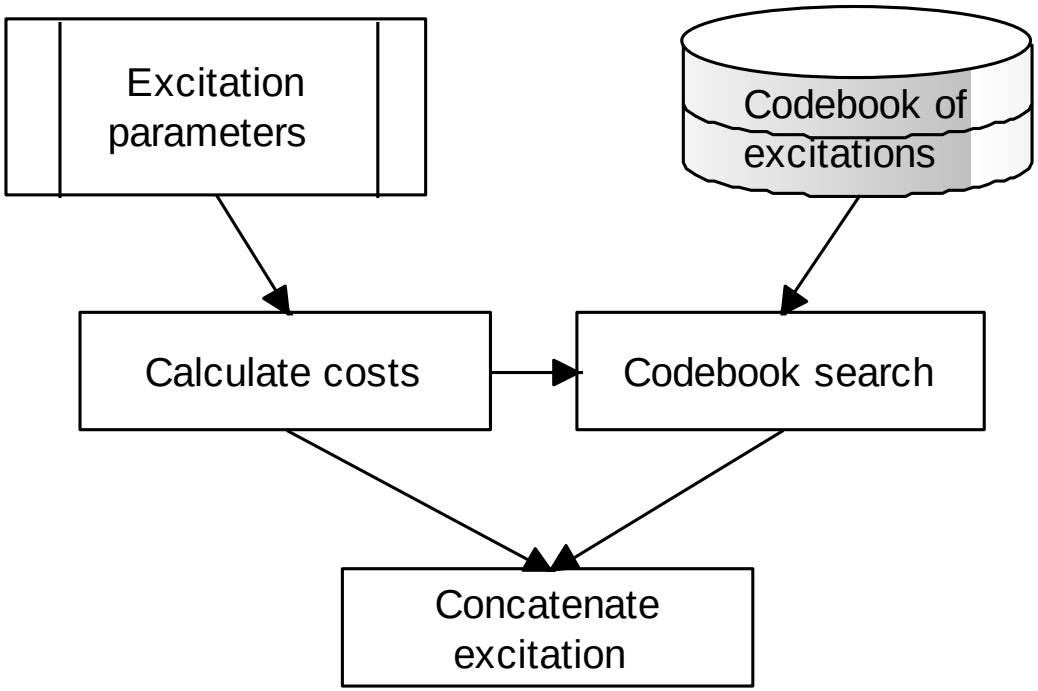
Codebook of excitations







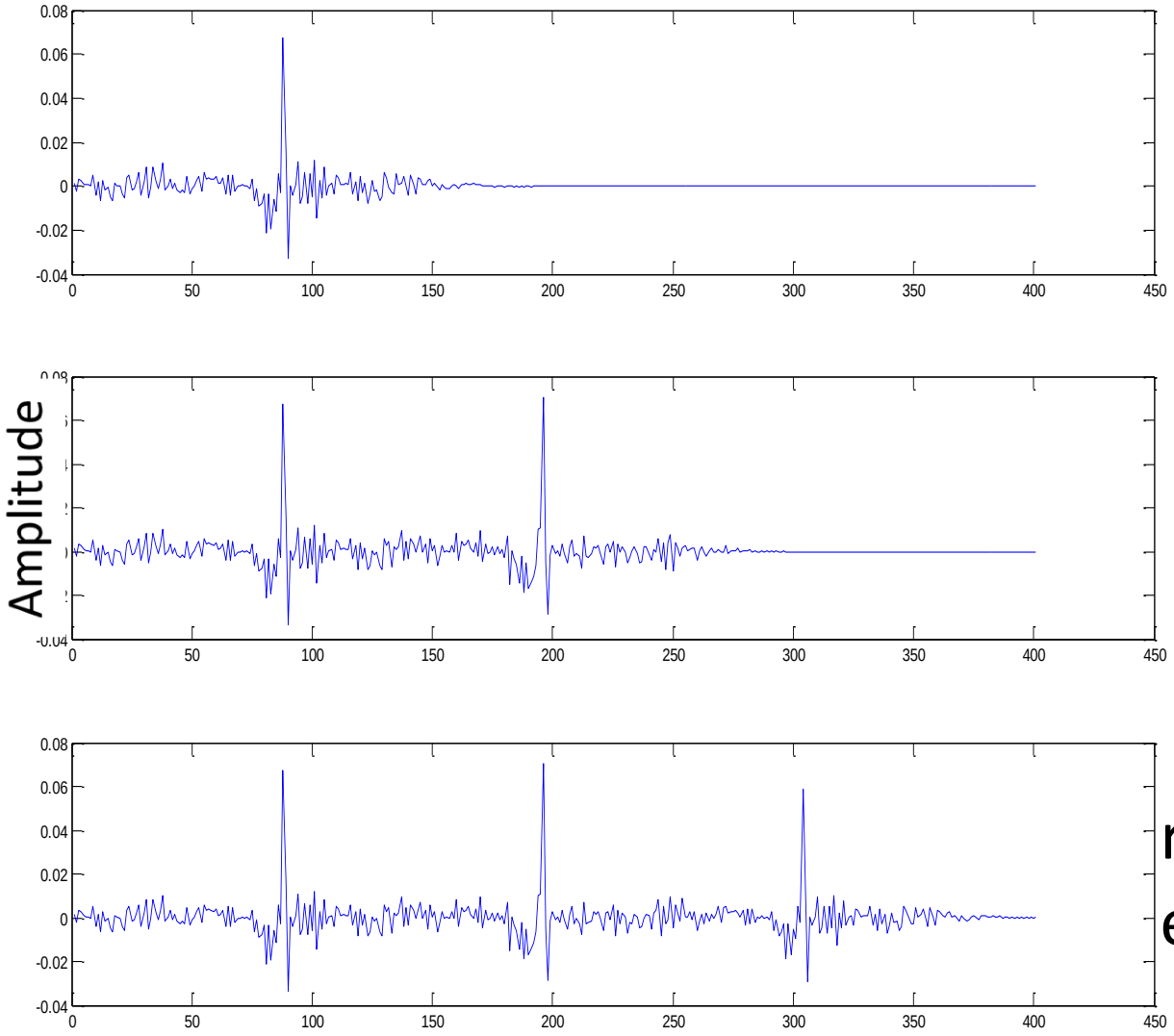
Speech decoding /1



Speech decoding /2

- Unit selection from the codebook
  - Well-known technique in TTS
- Target cost
- Concatenation cost
- Weights: $C_{total'} = w * C_{concatenation}^2 + C_{target}^2$
  - $w: \{0.01 \ldots 100\}$
  - Find optimal weight, $w \sim 1$

Concatenation: overlap-and-add



1 frame of windowed excitation

+ add 1 frame of windowed excitation

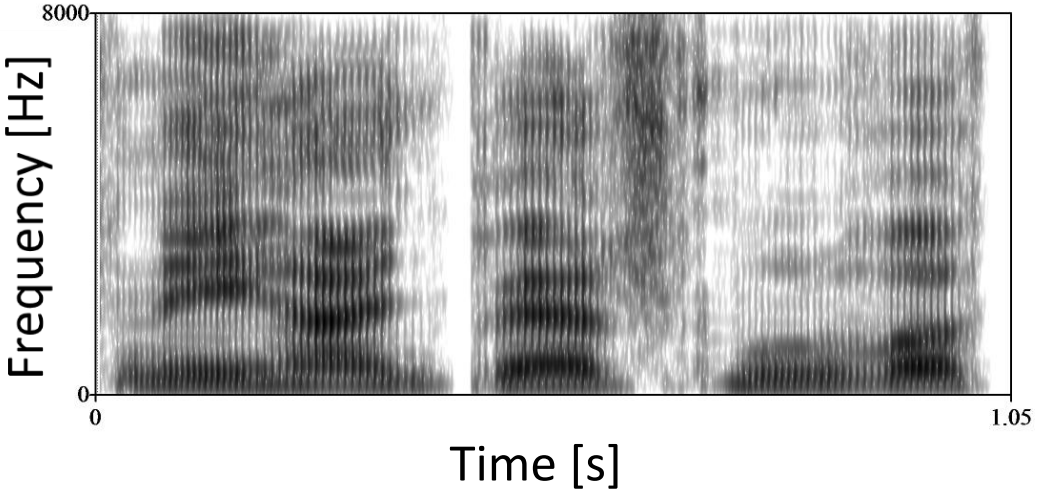
reconstructed excitation

Speech decoding /3

- Parameters
  - F0, places of impulses, gain
- Reconstructed excitation
- Spectral filtering
  - Mel-Generalized Cepstrum
- Speech signal

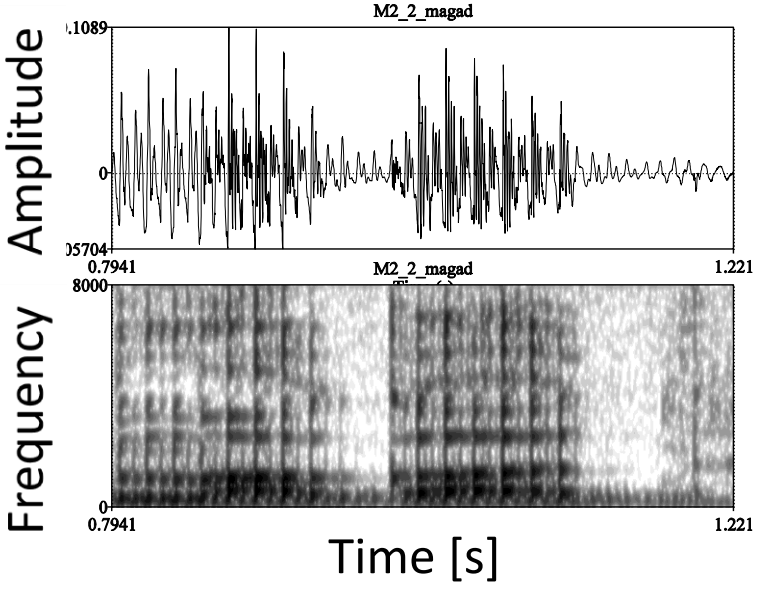
Results

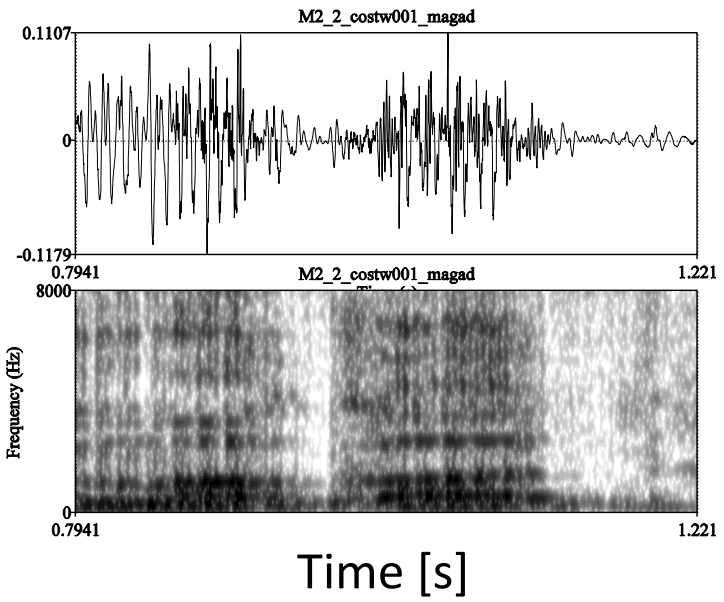
- Samples:
  - Original speech
  - Coded with simple pulse-noise
  - Coded with novel excitation model



Discussion

- Informal listening tests
- Male vs. Female speaker
- Creaky voice





Summary

- Novel excitation model
- Good quality for speech coding
- Modify parameters
- Modify prosody (e.g. intonation, F0)

90 %

110 %

Applications

- Text-to-speech synthesis with the novel excitation model
- TTS in CogInfoCom context
  - Cognitive Infocommunications can gain from better speech-driven human-machine interfaces
  - Natural communication modality between infocommunication systems and users
- Synthesize different voice qualities (e.g. breathy, whispered)

Tamás Gábor Csapó, Géza Németh: A novel codebook-based excitation model for use in speech synthesis

csapot@tmit.bme.hu

This research is partially supported by the Paelife (Grant No AAL-08-1-2011-0001) and the CESAR (Grant No 271022) projects.

CESAR
CEntral and South-east europeAn Resources

PaeLife