# Residual-based Excitation with Continuous F0 Modeling in HMM-based Speech Synthesis

Tamás Gábor Csapó[1], Géza Németh[1], Milos Cernak[2]

csapot@tmit.bme.hu

[1]Budapest University of Technology and Economics

[2]Idiap Research Institute

SLSP 2015
Budapest

Nov 24, 2015

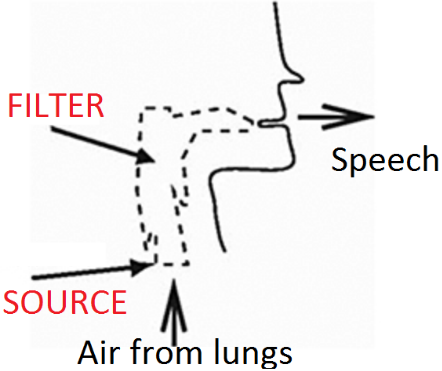
# HMM-based speech synthesis

# HMM-based speech synthesis

- State-of-the-art Text-To-Speech (TTS) synthesis technique [Zen et al., 2009]
- Statistical
  - Generative models with maximum likelihood criterion
  - Hidden Markov-models (HMM)
- Parametric
  - Excitation and spectral modeling
  - Speech signal is encoded to parameters
  - Parameters suitable for statistical modeling
  - Parameters are decoded to speech

# Excitation models in HMM-TTS

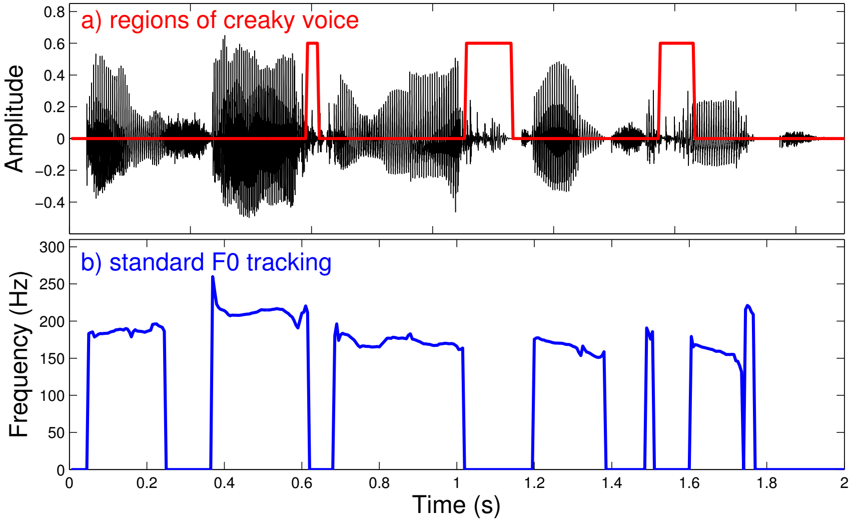
- Goal: model human speech production
- Source-filter separation [Fant, 1960]
- Excitation model types [Hu et al., 2013]
  - Impulse-noise
  - Mixed excitation
  - Glottal source
  - Harmonic plus noise
  - Sinusoidal
  - Residual-based



# Effect of creaky voice

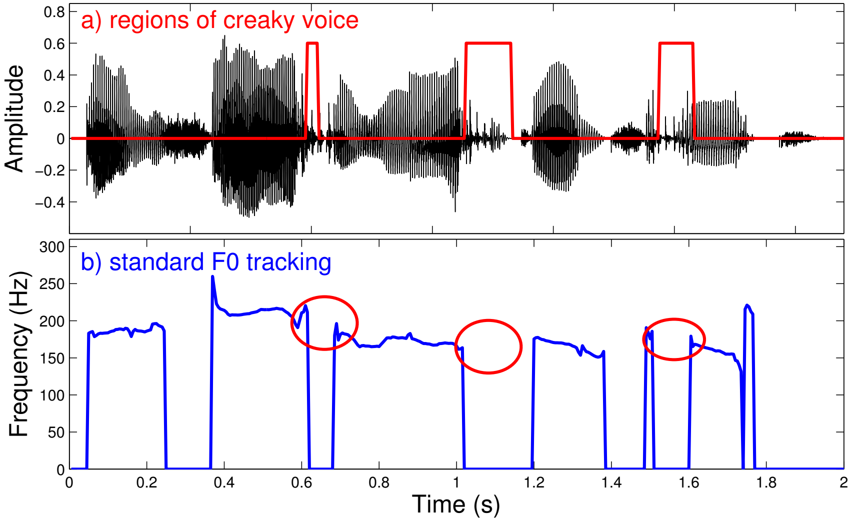
- Creaky voice
  - Irregular vibration of vocal folds
  - Abrupt changes in F0 (fundamental frequency, pitch) and/or amplitudes
  - Perceived as rough voice
  - Up to 15% of vowels of natural speech
- Effect of creaky voice on HMM-TTS
  - Can cause problems for standard speech analysis methods (e.g. F0 tracking and spectral analysis)
  - Voiced / unvoiced error is learned during training
  - Audible distortions in synthesized sentences

# Creaky voice sample




*'Eggshell is not good to eat.'* (sample)

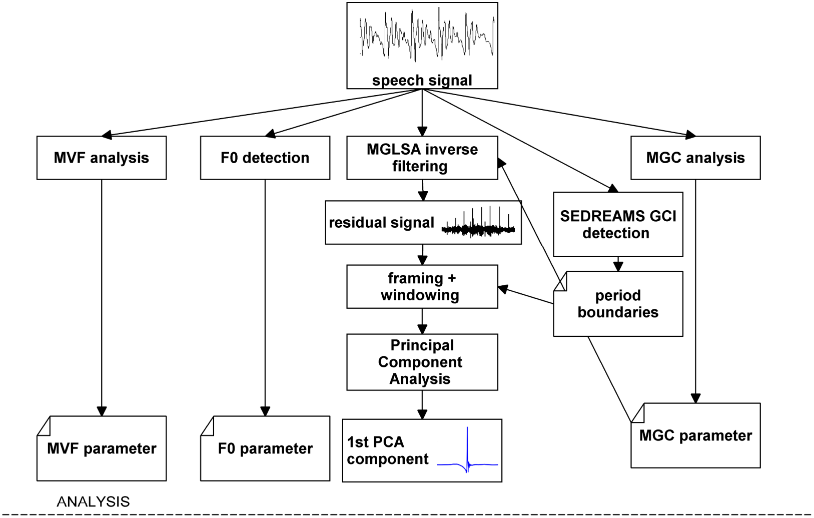
# Creaky voice sample



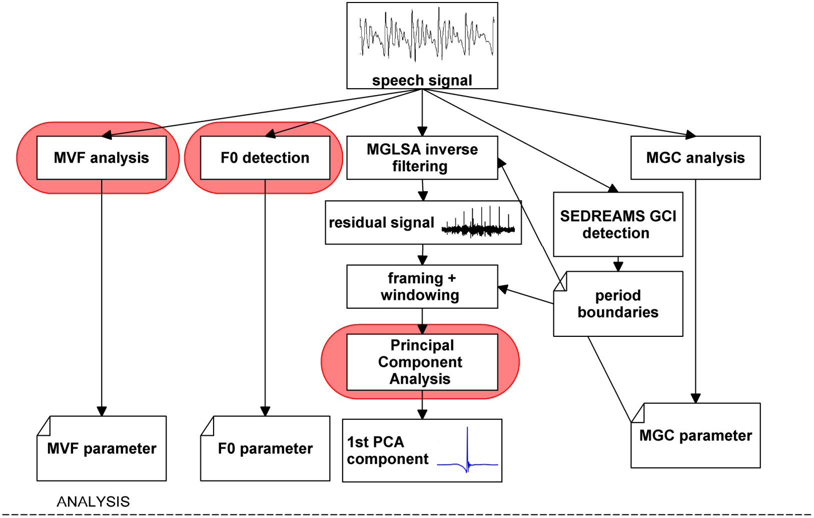

*'Eggshell is not good to eat.'* (sample)

# Proposed residual-based excitation model

# Block diagram of analysis



# Block diagram of analysis

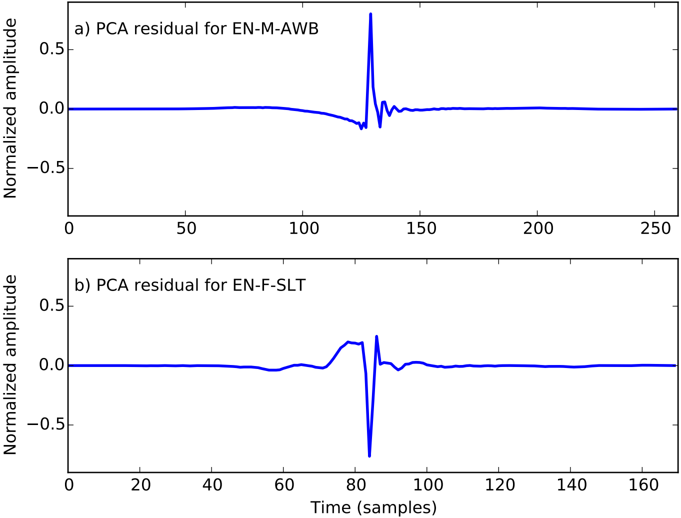


# Analysis: PCA-based residual

- Inverse filtered residual
- Pitch synchronous framing

- Earlier excitation models:
  - Store frames in a codebook
  - Select frames from codebook during synthesis
- Proposed model:
  - Window and resample frames to fixed length
  - Apply Principal Component Analysis (PCA)
  - Use first PCA component later

# Analysis: PCA-based residual



# Analysis: continuous F0 modeling

- Traditional F0 trackers
  - F0 is discontinuous, jumps occur at voiced-unvoiced transitions
  - HMMs can model continuous functions efficiently
  - Multi-Space Distribution (MSD) necessary for traditional F0 [Tokuda et al., 2002]
- Simple continuous pitch tracker 'F0cont' [Garner et al., 2013]
  - Standard autocorrelation
  - No voiced/unvoiced decision
  - Kalman smoothing-based interpolation
  - Interpolates F0 in regions of creaky voice
  - No need for MSD during training

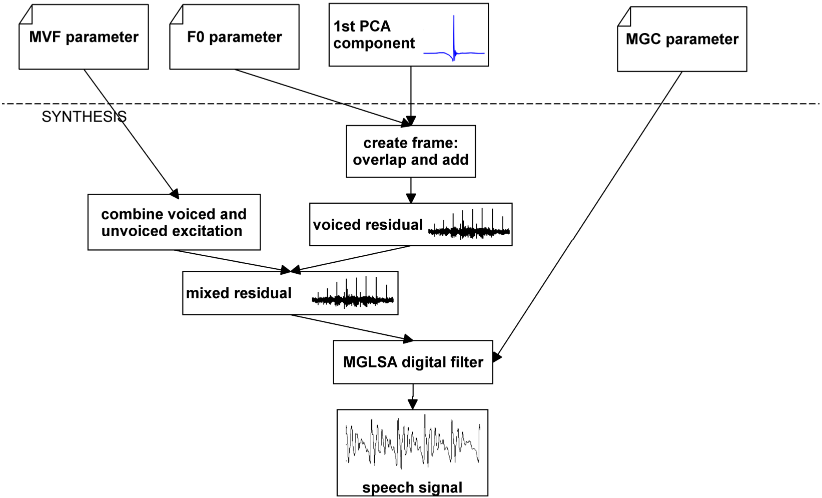
# Analysis: Maximum Voiced Frequency

- Divide spectrum to two frequency bands
  - Lower frequency band: voiced
  - Higher frequency band: unvoiced
- Earlier excitation models:
  - Boundary between frequency bands fixed (at 6 kHz)
- Proposed excitation model:
  - Boundary between frequency bands varying
  - Maximum Voiced Frequency (MVF) [Drugman and Stylianou, 2014]

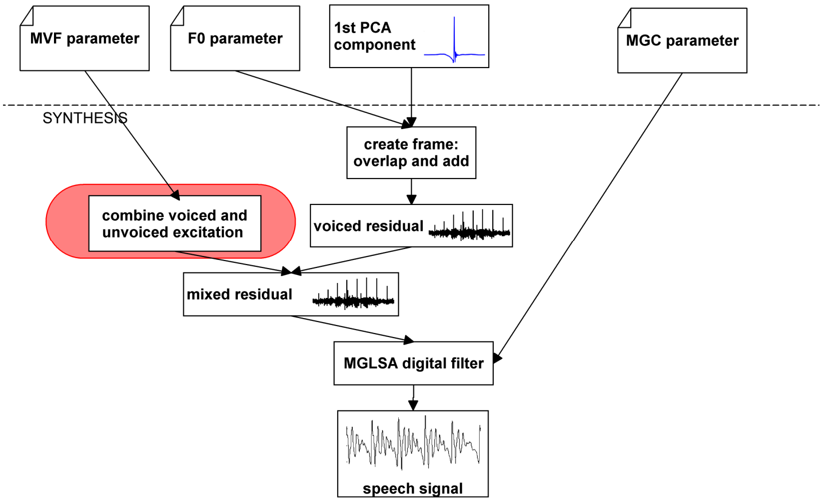
# Training with proposed model

- Parameters calculated for each 25 ms frame
  - MGC: Mel-Generalized Cepstrum
  - F0cont: continuous pitch track
  - MVF: Maximum Voiced Frequency
- Decision tree-based context clustering and Context dependent labeling [Zen et al., 2007]
- Independent decision trees for all the parameters and duration using a maximum likelihood criterion

# Block diagram of synthesis



# Block diagram of synthesis



# Synthesis features

- PCA residual overlap-added according to F0cont
- Voiced and unvoiced excitation component added together according to MVF
- MVF models voicing
  - for unvoiced sounds, the MVF is low (around 1 kHz)
  - for voiced sounds, the MVF is high (above 4 kHz)
  - for mixed excitation sounds, the MVF is in between (e.g. for voiced fricatives, MVF is around 2-3 kHz)
- Spectral filtering according to MGC

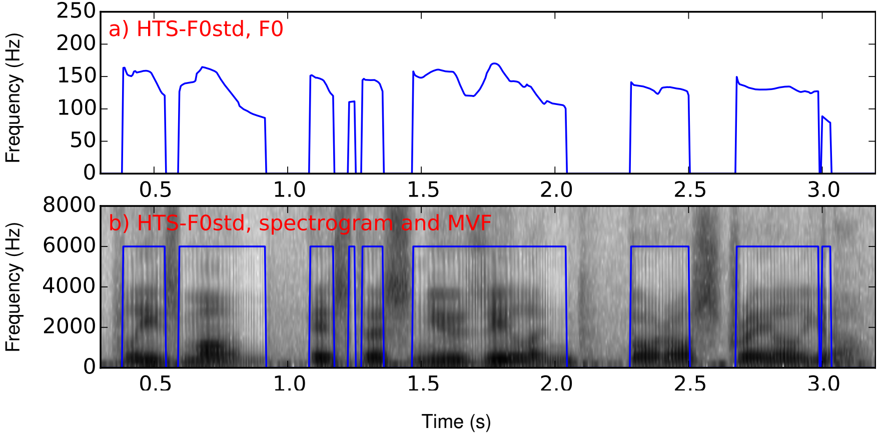
# Evaluation

# Data

- Two English speakers from CMU-ARCTIC database [Kominek and Black, 2003]
  - EN-M-AWB (Scottish English, male)
  - EN-F-SLT (American English, female)
  - Both produced irregular phonation frequently, mostly at the end of sentences
- 16 kHz sampling
- 1132 sentences from each speaker, single speaker training
- Text processing using the Festival TTS front-end (e.g. phonetic transcription, labeling, etc.)

# System A: HTS-F0std (baseline)

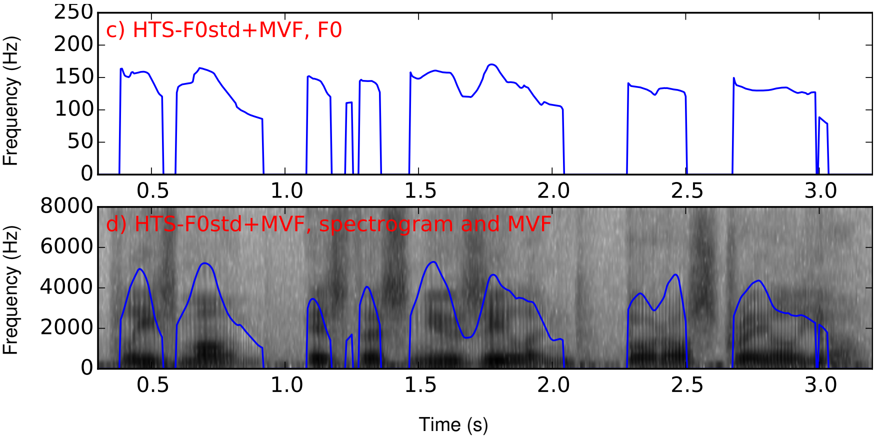
- standard pitch tracking
- voiced / unvoiced boundary fixed at 6 kHz




*'Please Mom, is this New Zealand, or Australia?'* (sample)

# System B: HTS-F0std+MVF

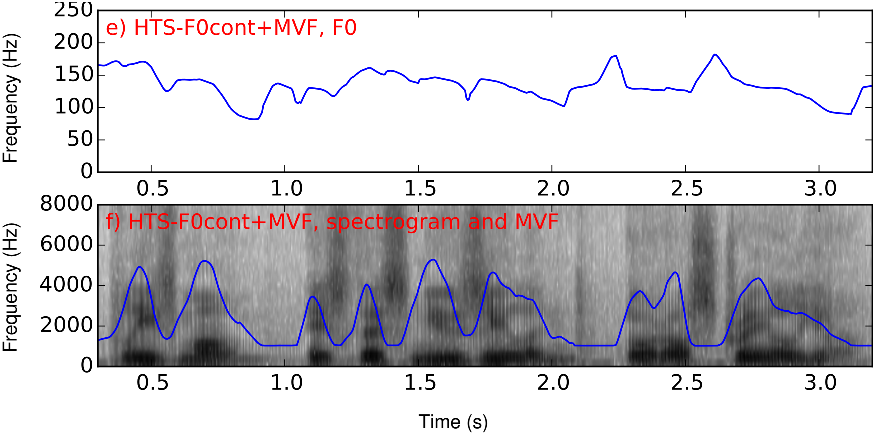
- standard pitch tracking
- voiced / unvoiced boundary according to MVF parameter




*'Please Mom, is this New Zealand, or Australia?'* (sample)

# System C: HTS-F0cont+MVF

- continuous pitch tracking
- voiced / unvoiced boundary according to MVF parameter




*'Please Mom, is this New Zealand, or Australia?'* (sample) [ A, B, C ]

# Listening test

- Web-based paired comparison test with one CMOS-like question
- 3 systems, 10 sentences, 2 speakers
- Which of the sentences is more natural?
  - 1: first much more natural
  - 2: first more natural
  - 3: equal
  - 4: second more natural
  - 5: second is much more natural
- 8 listeners, not native speakers of English
- `http://leszped.tmit.bme.hu/slsp2015_en/`

# Results of the listening test

- Speaker SLT (female)
  - System **A** < System **B** < System **C**
  - (sample A), (sample B), (sample C)
  - Proposed excitation model preferred
- Speaker AWB (male)
  - System **C** < System **B** = System **A**
  - Probably because high background noise
  - Vocoding caused audible artifacts

# Summary and conclusions

# Summary and conclusions

- Novel residual-based excitation model
  - PCA-based residual
  - Continuous F0 modeling
  - Maximum Voiced Frequency
- Evaluation
  - Improvement in perceived naturalness (for female)
  - Effect of creaky voice eliminated
  - Disturbing artifacts caused by unwanted voicing
- Possible application
  - TTS on smart devices (e.g. Android smartphones)
  - Personalized systems

# Future directions

- Improved modeling of the unvoiced sounds
    - Rule-based voiced/unvoiced decision
    - New parameter for voicing (e.g. Harmonics-To-Noise)
- Vocoding
    - Application in low bitrate speech coding

# Thank you for your attention!


- Tamás Gábor Csapó, Géza Németh, Milos Cernak, „Residual-based Excitation with Continuous F0 Modeling in HMM-based Speech Synthesis”
- `csapot@tmit.bme.hu`



- Supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech prosody (SNSF no IZ73Z0 152495-1) and by the EITKIC project (EITKIC 12-1-2012-001).

# References I


Drugman, T. and Stylianou, Y. (2014).
Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra.
*IEEE Signal Processing Letters*, 21(10):1230–1234.

Fant, G. (1960).
*Acoustic theory of speech production*.
Mouton, The Hague.

Garner, P. N., Cernak, M., and Motlicek, P. (2013).
A simple continuous pitch estimation algorithm.
*IEEE Signal Processing Letters*, 20(1):102–105.

Hu, Q., Richmond, K., Yamagishi, J., and Latorre, J. (2013).
An experimental comparison of multiple vocoder types.
In *Proc. ISCA SSW8*, pages 155–160.

Kominek, J. and Black, A. W. (2003).
CMU ARCTIC databases for speech synthesis.
Technical report, Language Technologies Institute.

Tokuda, K., Mausko, T., Miyazaki, N., and Kobayashi, T. (2002).
Multi-space probability distribution HMM.
*IEICE Transactions on Information and Systems*, E85-D(3):455–464.

# References II


Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., and Black, A. (2007).
The HMM-based speech synthesis system version 2.0.
In *Proc. ISCA SSW6*, pages 294–299, Bonn, Germany.

Zen, H., Tokuda, K., and Black, A. W. (2009).
Statistical parametric speech synthesis.
*Speech Communication*, 51(11):1039–1064.