Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder

Tamás Gábor Csapó[1], Géza Németh[1], Milos Cernak[2], Philip N. Garner[2]

csapot@tmit.bme.hu

[1]Budapest University of Technology and Economics, Hungary

[2]Idiap Research Institute, Switzerland

EUSIPCO 2016
Budapest

Aug 31, 2016

Speech synthesis

Text-To-Speech synthesis (TTS)

- Convert written text to human-like speech
- Important in human-computer interaction
- Applications for vision impaired and blind people
- Research goal:
  develop natural and flexible solutions

Statistical parametric synthesis I

- State-of-the-art TTS synthesis technique

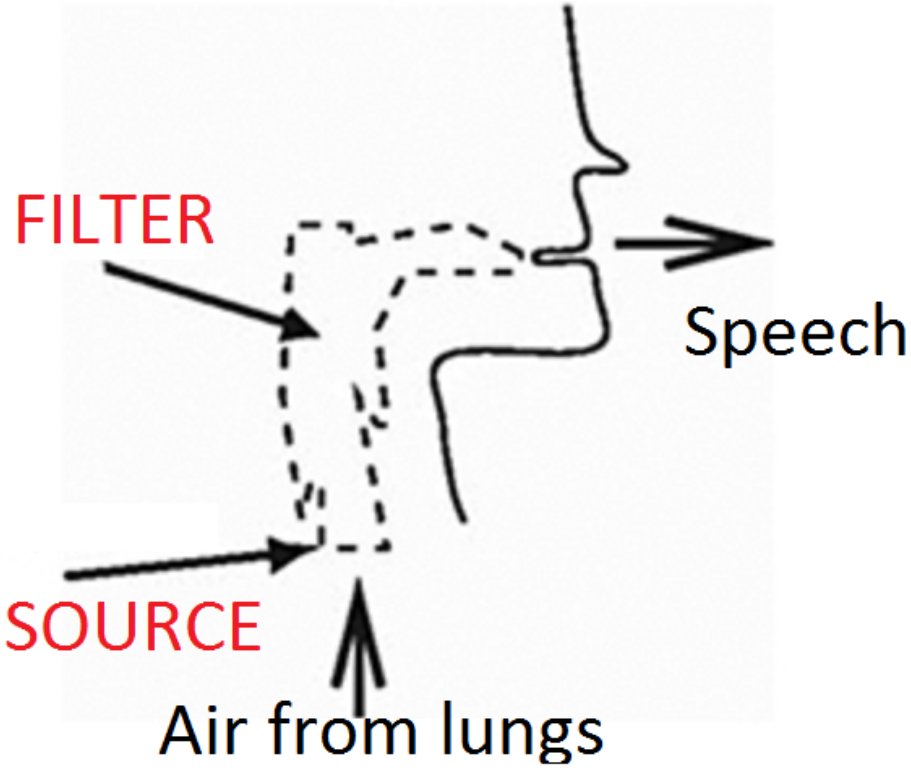
  [Zen et al., 2009]
- Statistical
  - Hidden Markov-models (HMM)
    - Generative models with maximum likelihood criterion
  - Deep Neural Networks (DNN)
- Parametric
  - Excitation and spectral modeling
  - Speech signal is encoded to parameters and decoded to speech, using parametric vocoders

Statistical parametric synthesis II

- Advantages
  - Flexibility, smoothness and small footprint
  - Speaker and style adaptation
- Main limiting factors
  - Over-simplified vocoder techniques (excitation models)
  - Acoustic modeling accuracy
  - Over-smoothing during parameter generation

Vocoding 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

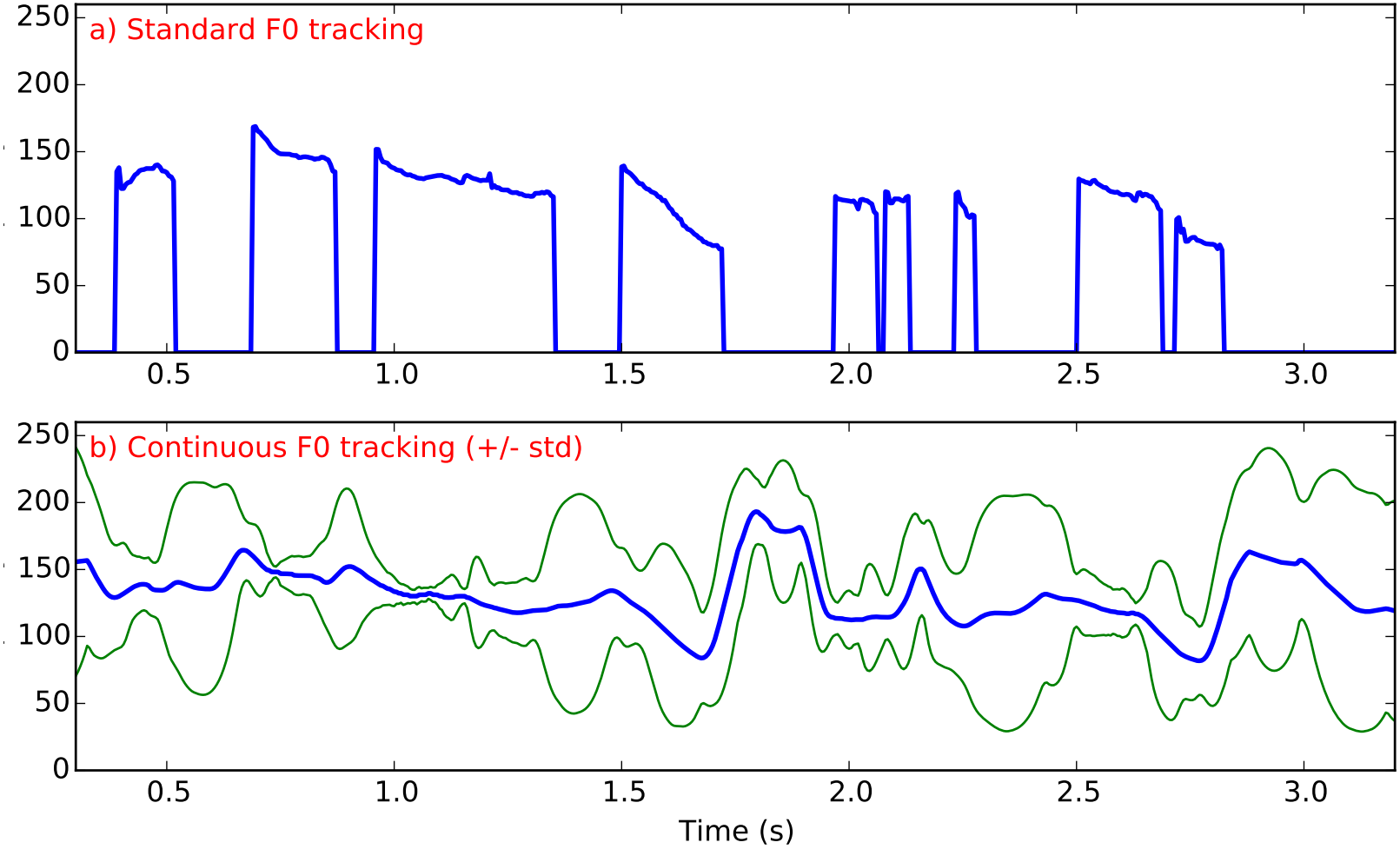


Proposed continuous vocoder

A: Baseline vocoder

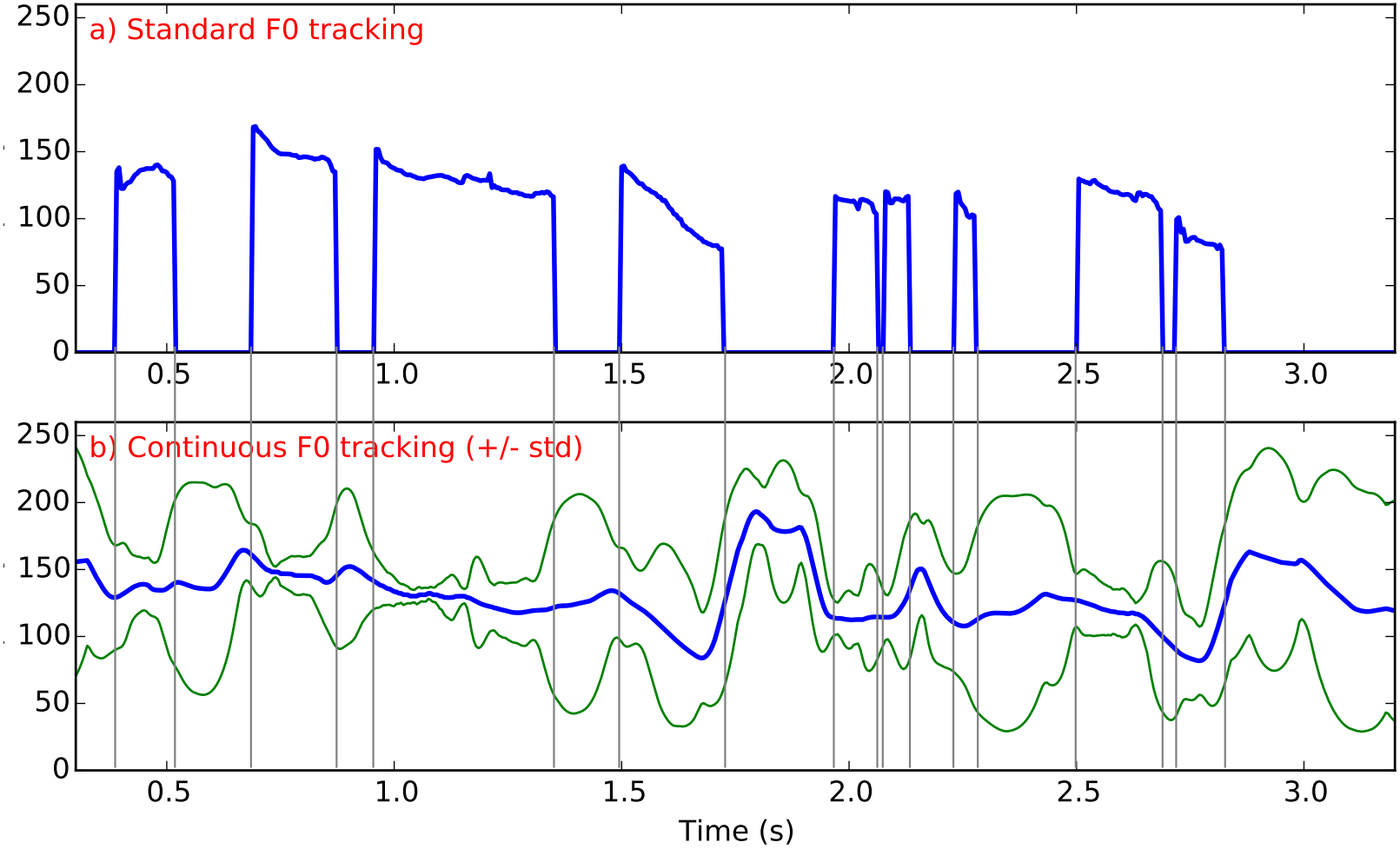
- Analysis
  - Linear Prediction residual-based excitation [Csapó et al., 2015]
  - Continuous pitch (F0) modeling [Garner et al., 2013]
  - Divide spectrum to two frequency bands, Maximum Voiced Frequency (MVF) [Drugman and Stylianou, 2014]
- Statistical training of HMMs
  - Decision tree-based context clustering [Zen et al., 2007]
  - Independent decision trees for all the parameters, maximum likelihood criterion
- Synthesis
  - Voiced and unvoiced excitation component added together according to MVF

A: continuous F0



'I saw it all myself, and it was splendid.'

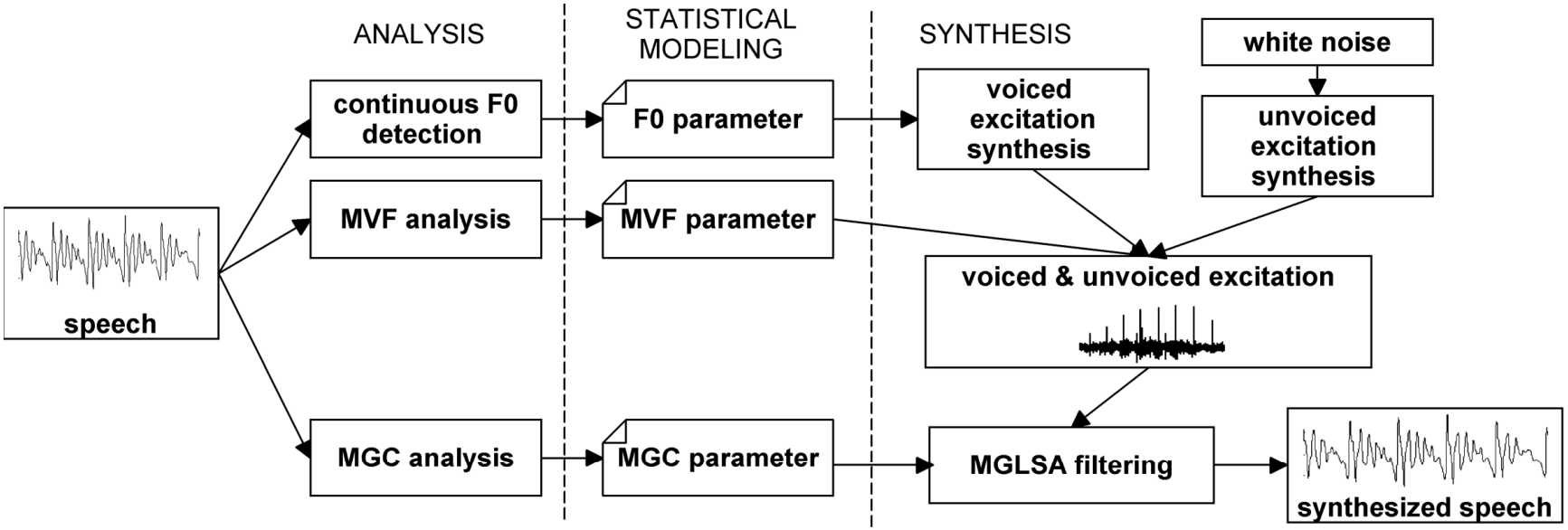
A: continuous F0



'I saw it all myself, and it was splendid.'

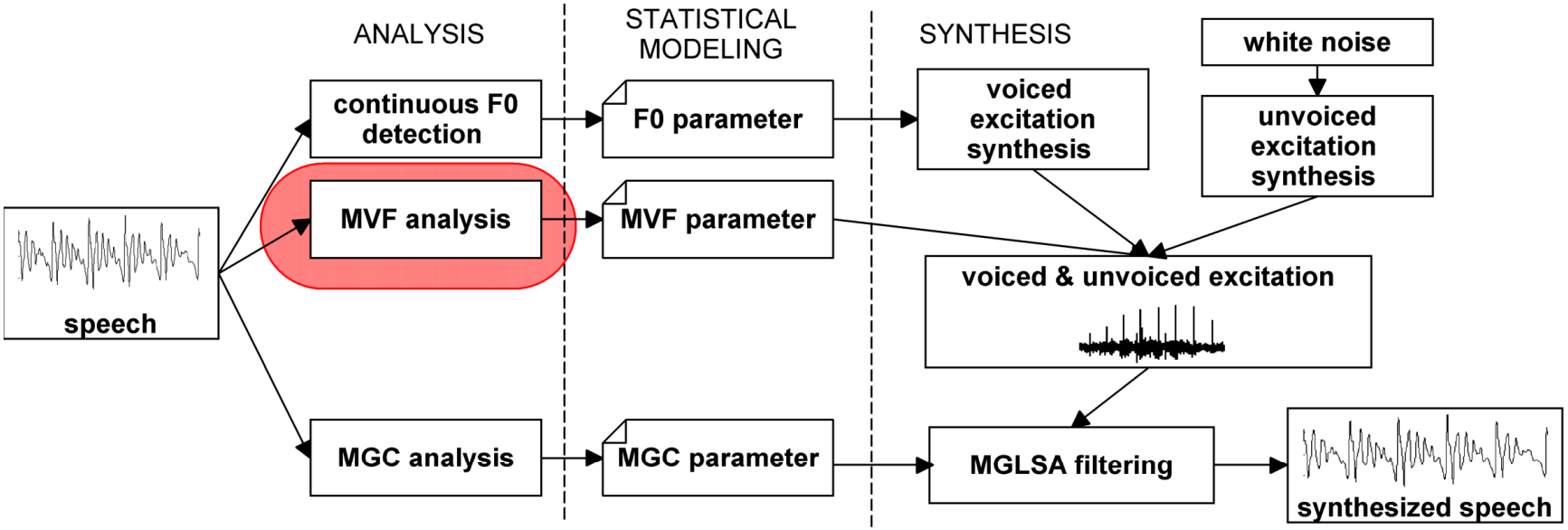
A: Baseline vocoder

- Problem: estimated MVF too high
- Causes buzzy speech in unvoiced segments (sample)



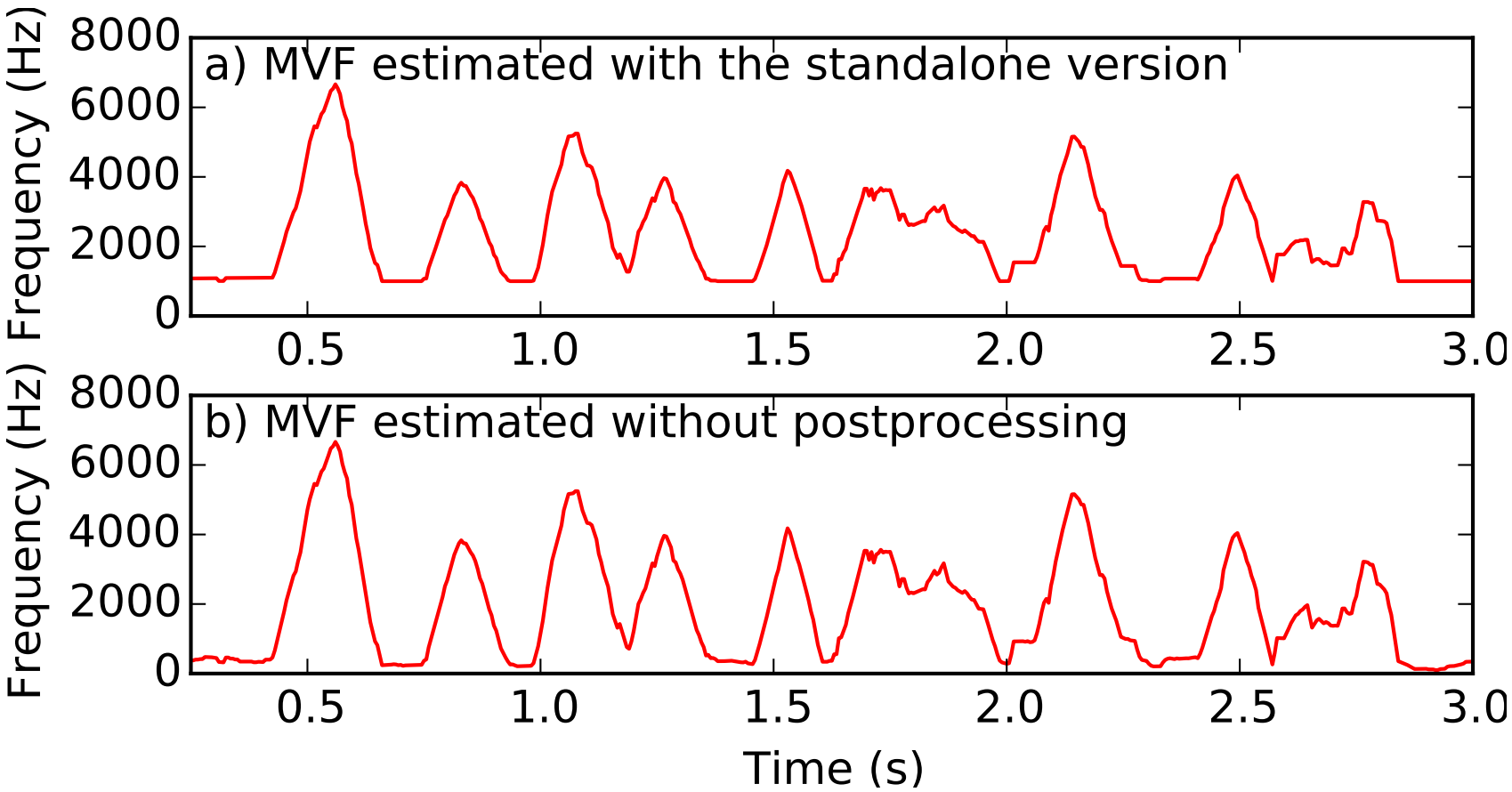
A: Baseline vocoder

- Problem: estimated MVF too high
- Causes buzzy speech in unvoiced segments (sample)



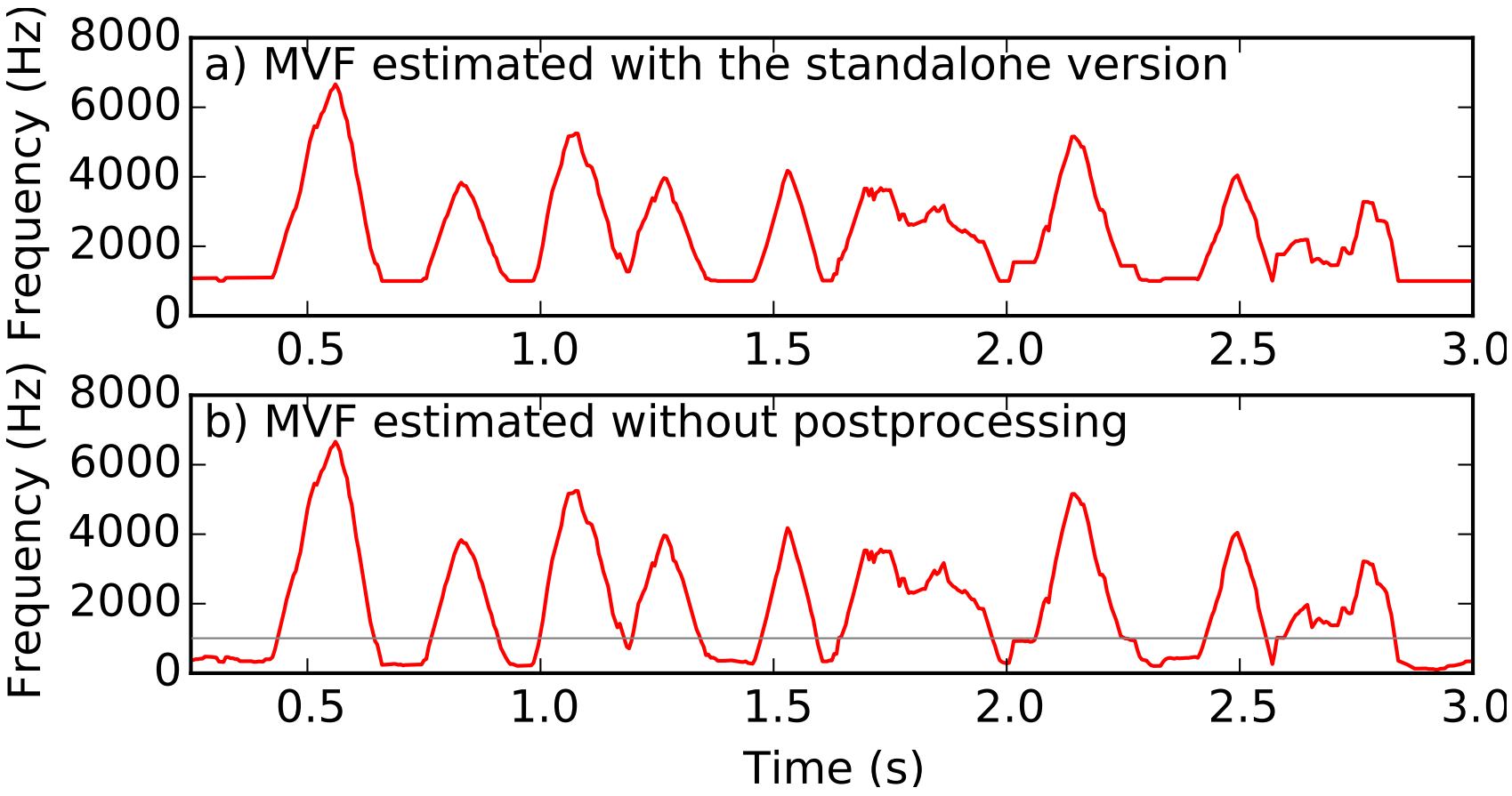
B: Proposed vocoder: data-driven

- No post-processing during MVF estimation



B: Proposed vocoder: data-driven

- No post-processing during MVF estimation



C: Proposed vocoder: rule-based

- Textual labels of speech sounds are known at synthesis time
- Decide about voicing based on textual labels
- Set MVF to 0 at unvoiced sounds /ch, f, k, p, s, sh, t, th/
- Unvoiced sounds are synthesized with fully unvoiced excitation

Hypotheses

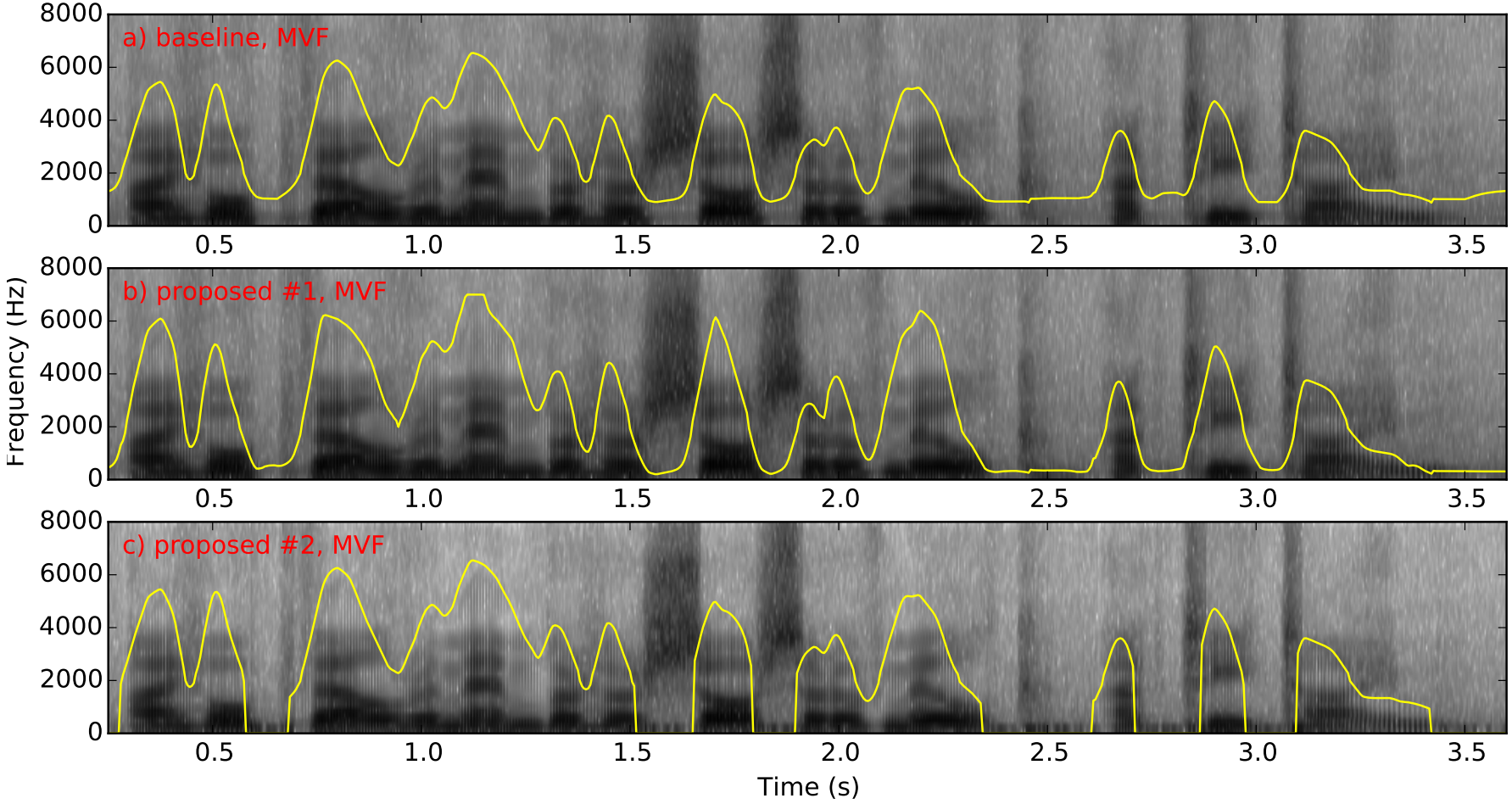
1. Both the data-driven (Vocoder B) and rule-based (Vocoder C) MVF modeling strategies will be superior to the baseline (Vocoder A)
2. Lower MVF values of Vocoder B will lead to less buzziness
3. The abrupt voiced-unvoiced transitions in Vocoder C are not audible / disturbing

Evaluation

Data for the evaluation

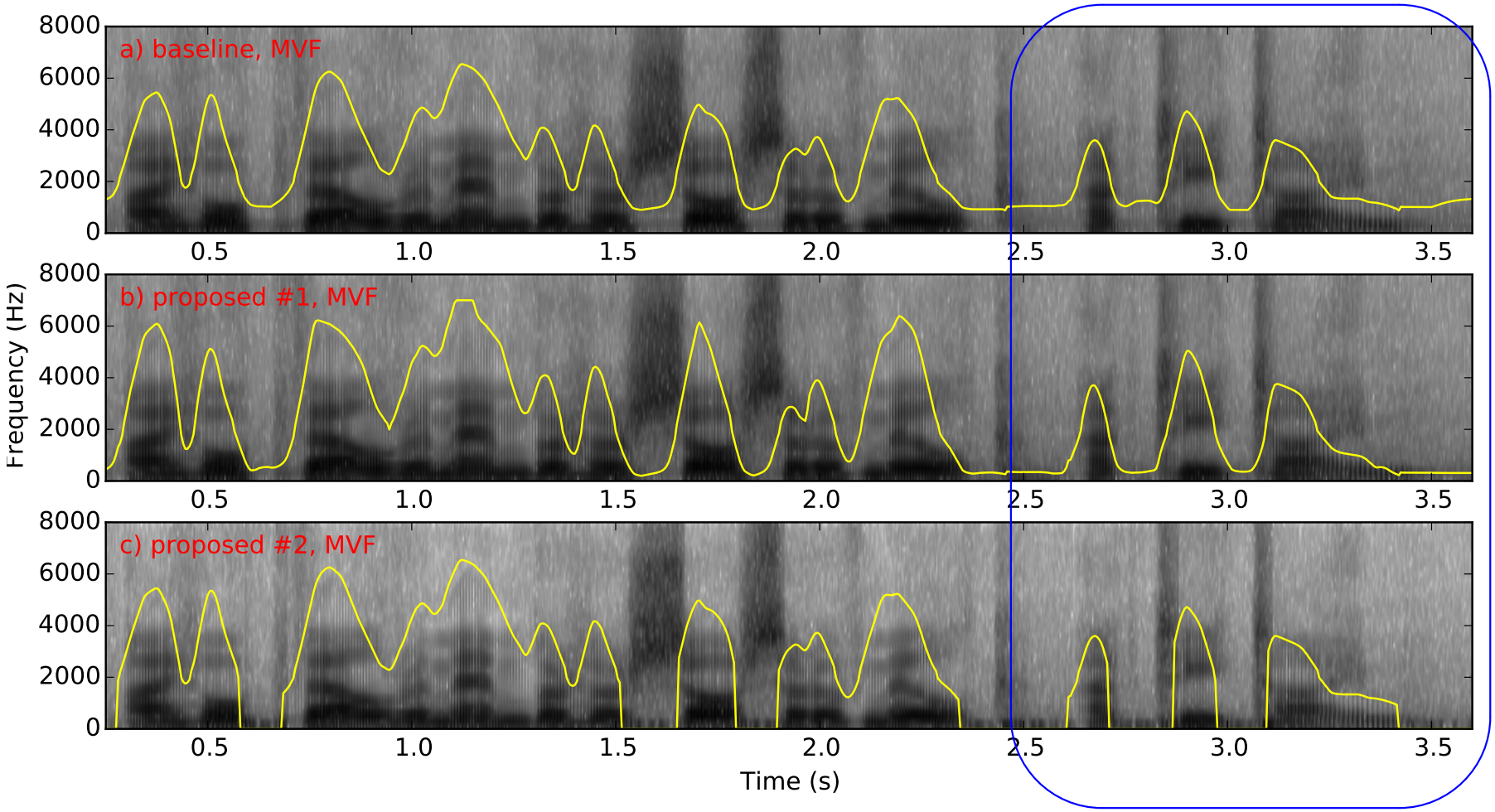
- Two English speakers from CMU-ARCTIC database [Kominek and Black, 2003]
  - AWB (Scottish English, male)
  - SLT (American English, female)
- 16 kHz sampling rate
- 90% of the sentences (~1000) used for single speaker training of HMM-TTS

Samples for system A, B, C



'I hope they'll remember her saucer of milk at tea-time.' (A, B, C)

Samples for system A, B, C



'I hope they'll remember her saucer of milk at tea-time.' (A, B, C)

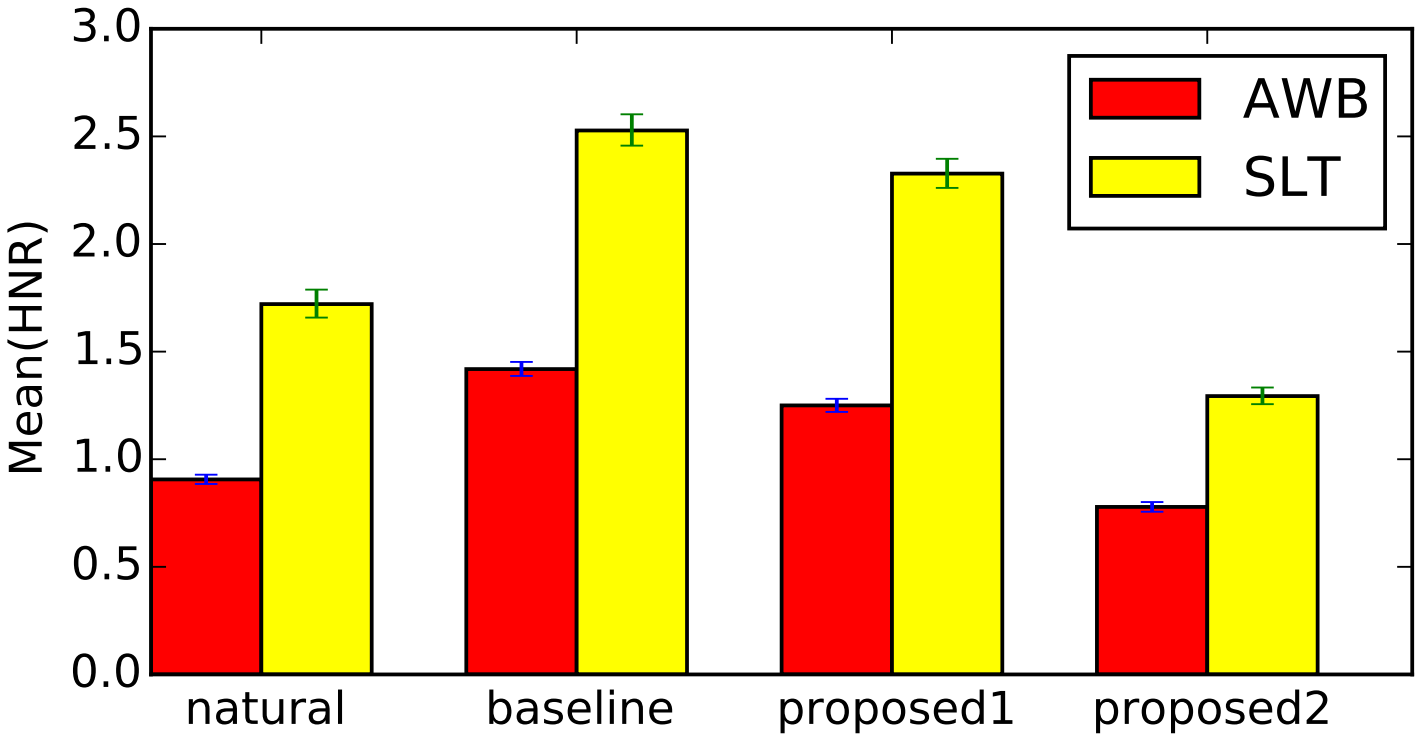
Objective experiment (methods)

- Unvoiced sounds selected from natural and synthesized samples
  /ch, f, k, p, s, sh, t, th/
- Harmonics-to-Noise (HNR) ratio measured at 5 ms frame shift
- Lower HNR = less voiced

- Goal: compare unvoiced components

Objective experiment (results)

- $HNR_{baseline(A)} > HNR_{proposed(B)} > HNR_{proposed(C)}$
- Proposed B, C closer to natural than baseline

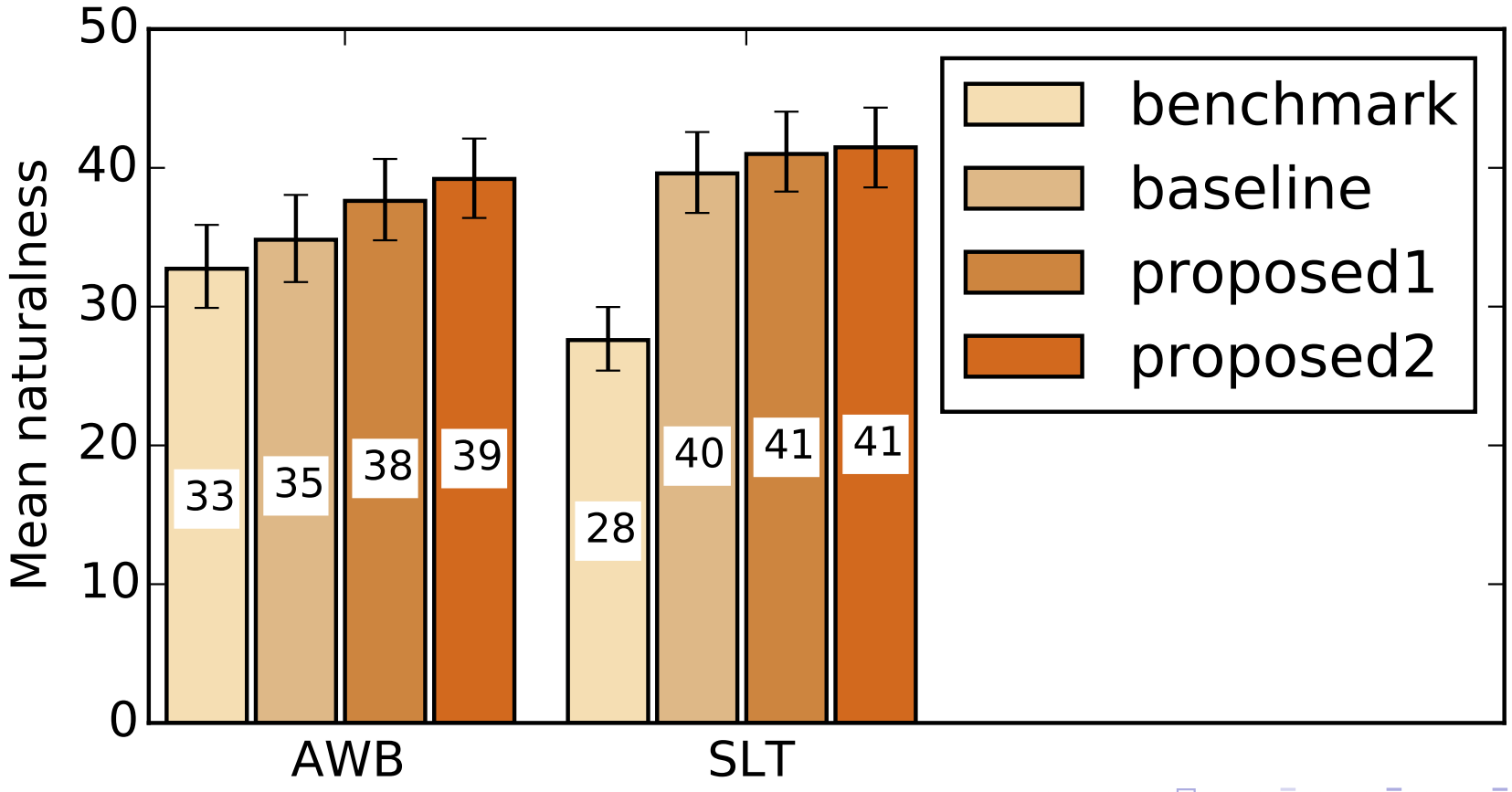


Subjective experiment (methods)

- MUSHRA listening test
  - benchmark: impulse-noise excitation
  - Vocoder A (baseline)
  - Vocoder B (proposed, data-driven)
  - Vocoder C (proposed, rule-based)
  - natural sentences
- 5 types, 10 sentences, 2 speakers
- Goal: evaluate overall naturalness
- 10 listeners (one native speaker of English)
- `http://leszped.tmit.bme.hu/eusipco2016_en/`

Subjective experiment (results)

- baseline (A) < proposed (B) < proposed (C)
- significant for speaker AWB



Summary and conclusions

Summary and conclusions

- Improved unvoiced sounds in novel vocoder
  - Proposed B: no postprocessing during MVF estimation
  - Proposed C: rule-based voicing decision
- Evaluation
  - Proposed vocoders preferred over baseline
  - Proposed B is less buzzy than baseline
  - Proposed C unvoiced-voiced transitions are OK
  - Improvement in perceived naturalness (for male)
- Possible application
  - TTS on smart devices (e.g. Android smartphones)
  - Personalized systems

Future directions

- Uniform modeling of voice qualities (e.g. creaky and breathy voice)
- New parameter for voicing (e.g. Harmonics-To-Noise)
- Application in low bitrate speech coding

- Journal manuscript: „Parametric Vocoder with Continuous F0 Modeling and Residual-based Excitation for Speech Synthesis", submitted

Thank you for your attention!

- T. G. Csapó, G. Németh, M. Cernak, P. N. Garner, „Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder”
- `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).

References I

Csapó, T. G., Németh, G., and Cernak, M. (2015).
Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis.
In Dediu, A.-H., Martín-Vide, C., and Vicsi, K., editors, *Lecture Notes in Artificial Intelligence*, volume 9449, pages 27–38. Springer International Publishing, Budapest, Hungary.

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.

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.