A novel irregular voice model for HMM-based speech synthesis

Tamás Gábor Csapó, Géza Németh

Budapest University of Technology and Economics, Hungary
Dept. of Telecommunications and Media Informatics

8th Speech Synthesis Workshop
2013 September 2
Barcelona, Spain

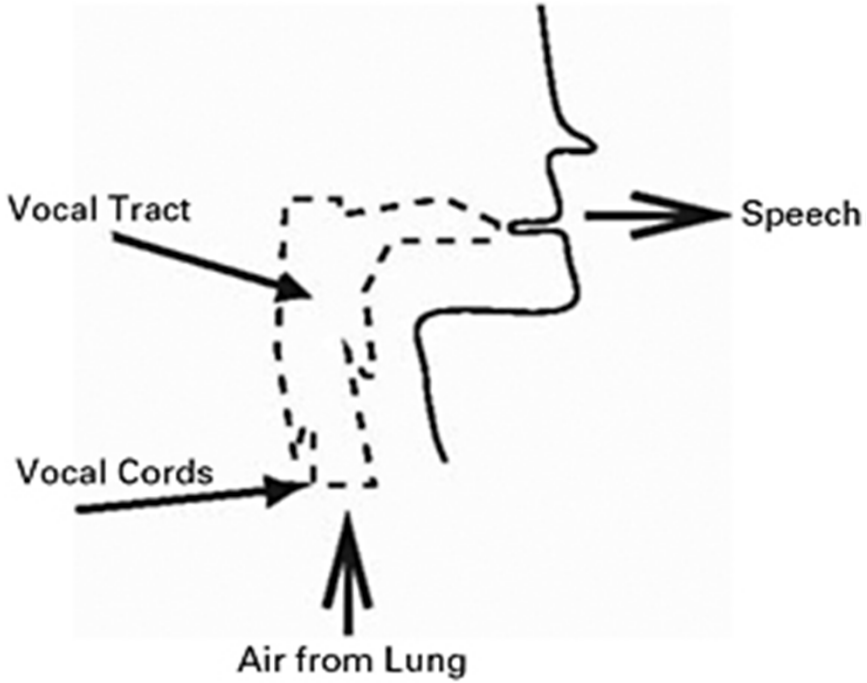
Contents

- Excitation models in HMM-TTS
- Irregular voice and its models
- Novel irregular voice model
- Perceptual & acoustic evaluation

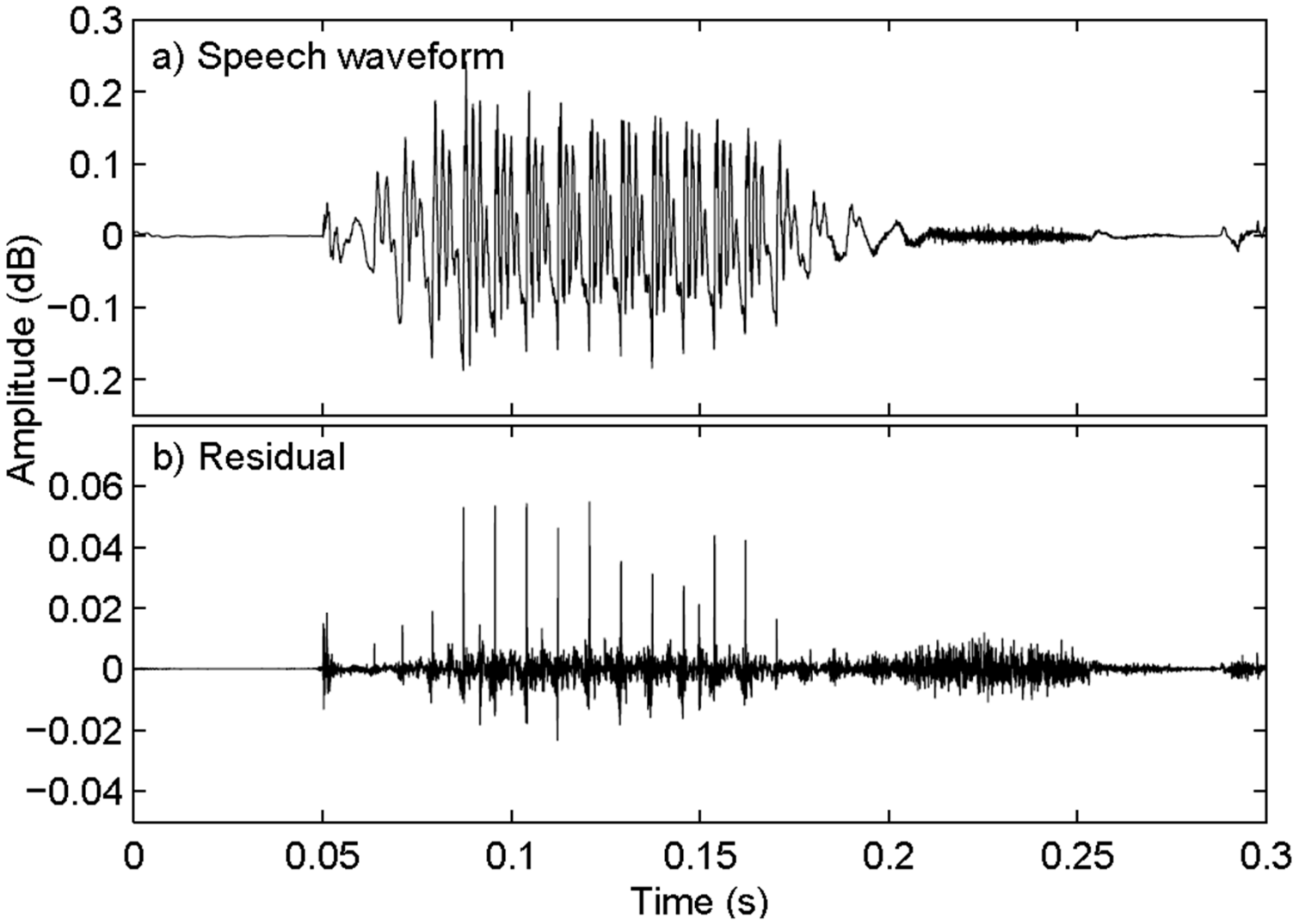
INTRODUCTION

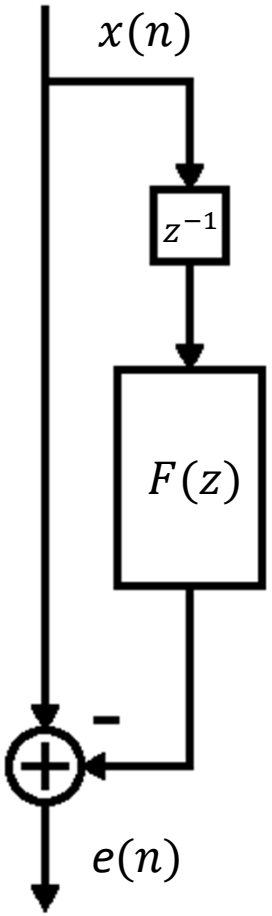
Speech excitation models in HMM-TTS

- Goal: model human speech production
- Source-filter separation [Fant'60]
- Types [Hu;'13] SSW8
  - Impulse-noise
  - Mixed excitation
  - Glottal source
  - Harmonic plus noise
  - Residual based



Linear Prediction residual of speech

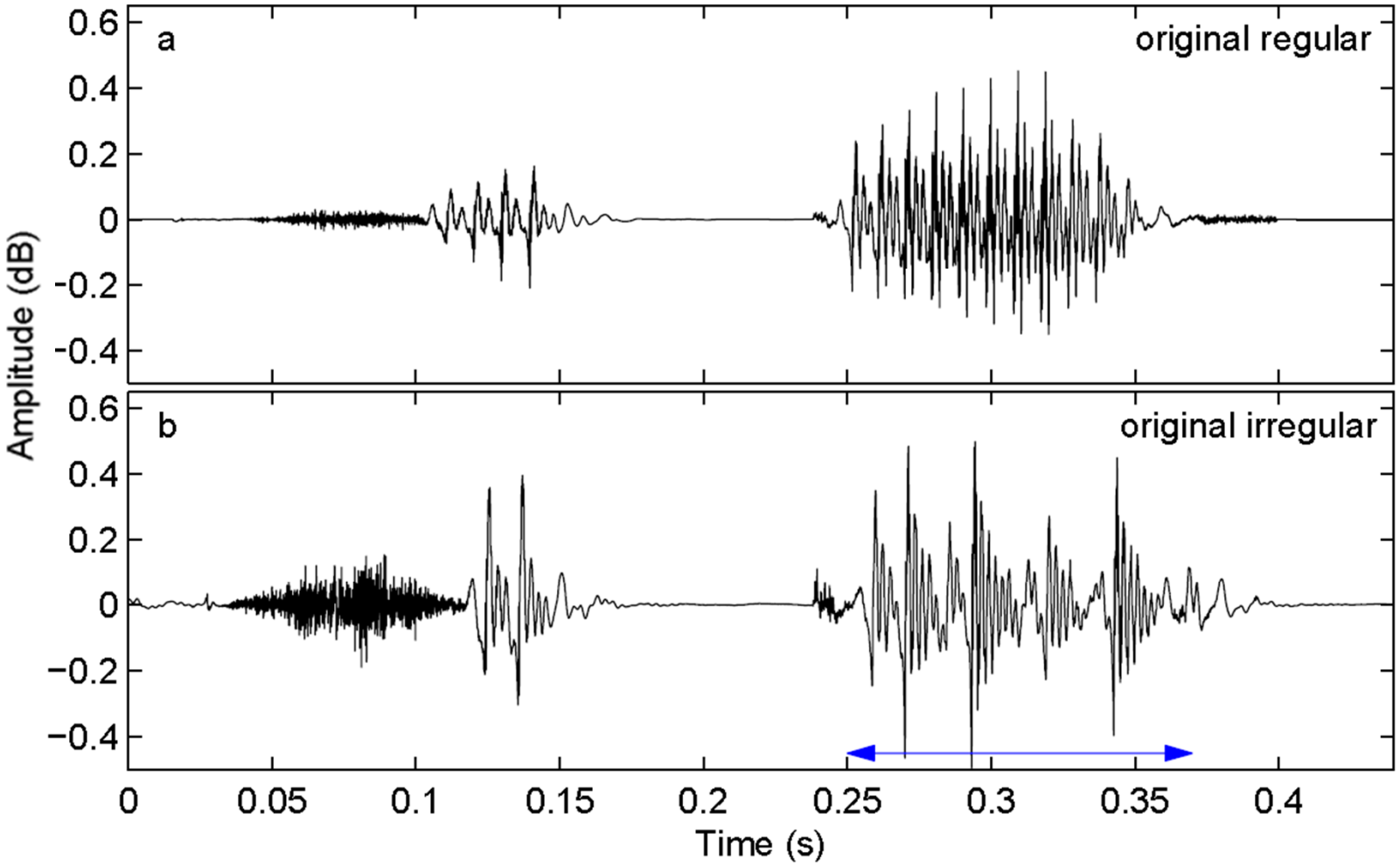




Irregular voice: occurrance

- Irregular vibration of vocal folds [Blomgren;'98] [Gobl&Chasaide'03]
  - Irregular F0 and/or amplitudes
- Creaky voice, laryngealization, vocal fry, glottalization
- Up to 15% of vowels of natural speech [Bőhm;'09]
- Location [Dilley;'96]
  - Phrase boundaries
  - Sentence endings
  - Vowel-vowel transitions

Irregular voice: example



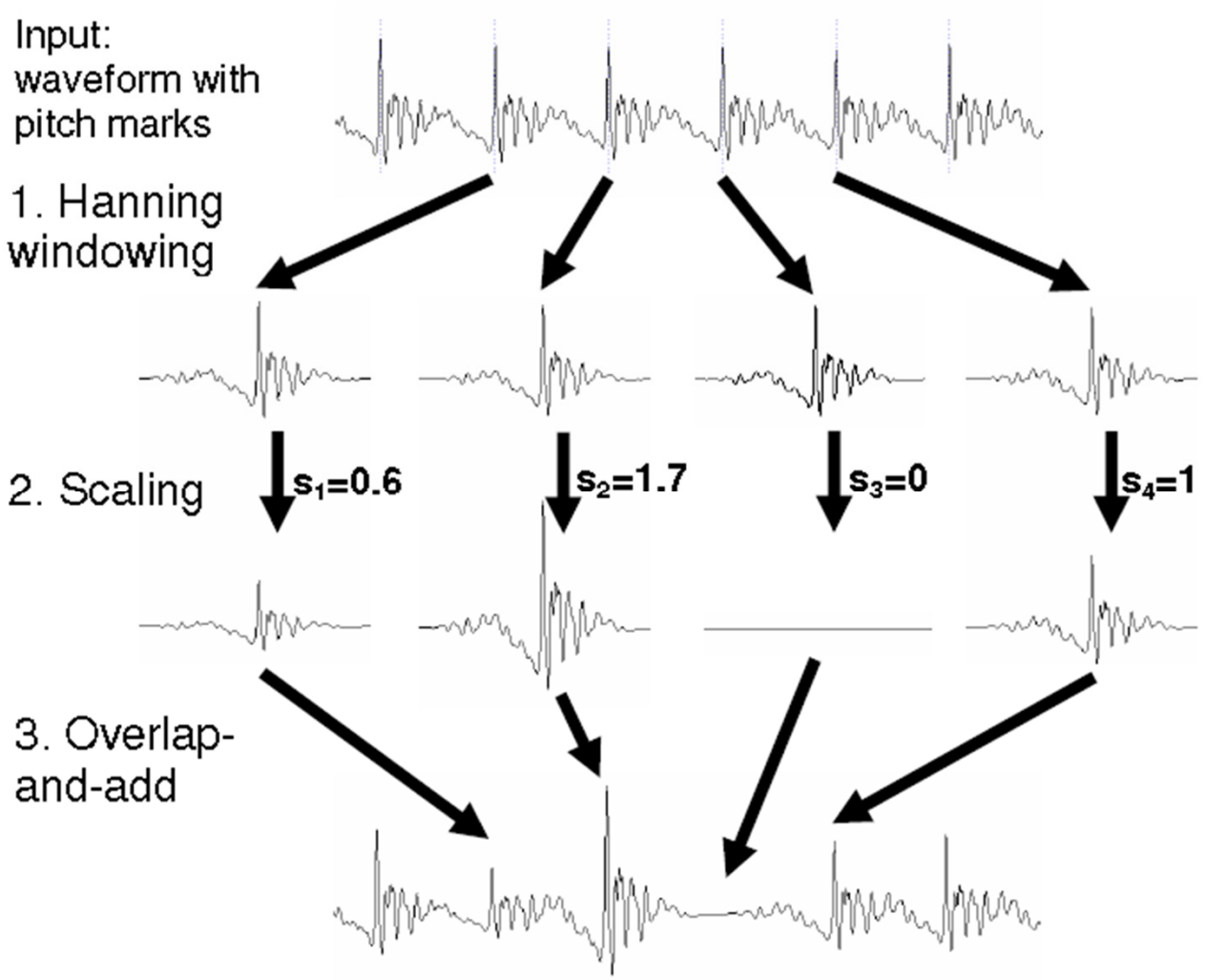
Irregular voice: acoustic properties

- Differences compared to regular speech [Klatt&Klatt'90] [Bőhm;'09]
  - time between successive glottal pulses longer and more irregular
  - lower F0 and higher jitter
  - abrupt changes in the amplitude of the periods
  - lowered open quotient (proportion of the glottal cycle where the glottis is open)
  - increased first formant bandwidth because of more acoustic losses at the glottis
  - more abrupt closure of the vocal folds

Irregular voice: models in HMM-TTS

- [Silén;'09] Interspeech
  - Robust F0 measure and two-band voicing
  - Not focusing on characteristics of irregular voice
- [Drugman;'12] Interspeech
  - Extension of DSM model: secondary pulses in the residual excitation
- [Drugman;'13] ICASSP
  - Prediction of creaky voice position
- [Raitio;'13] Interspeech
  - Creaky voice integrated into HTS
- Proposed method
  - Uses another excitation model
  - Improvement of previous regular-to-irregular transformation
  - 3 heuristics model irregular voice

[Bőhm;'09] regular-to-irregular transformation



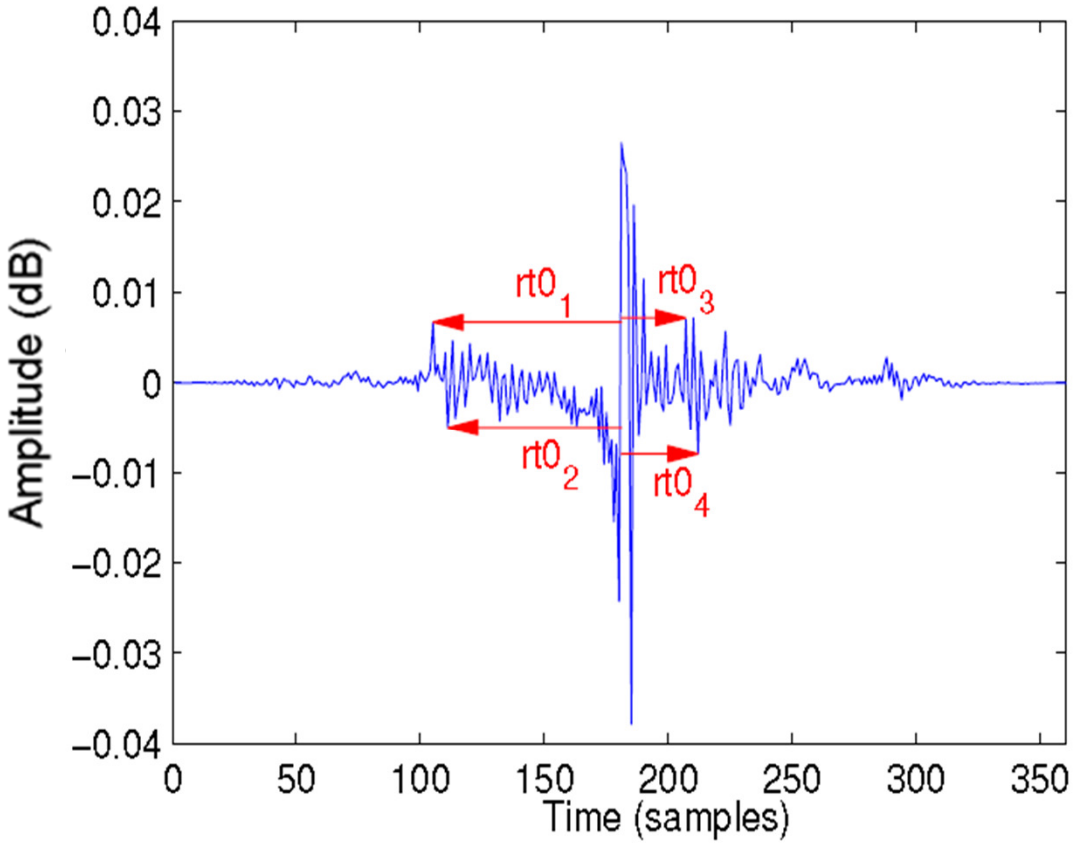
OUR METHODS

Baseline: HTS-CDBK excitation model

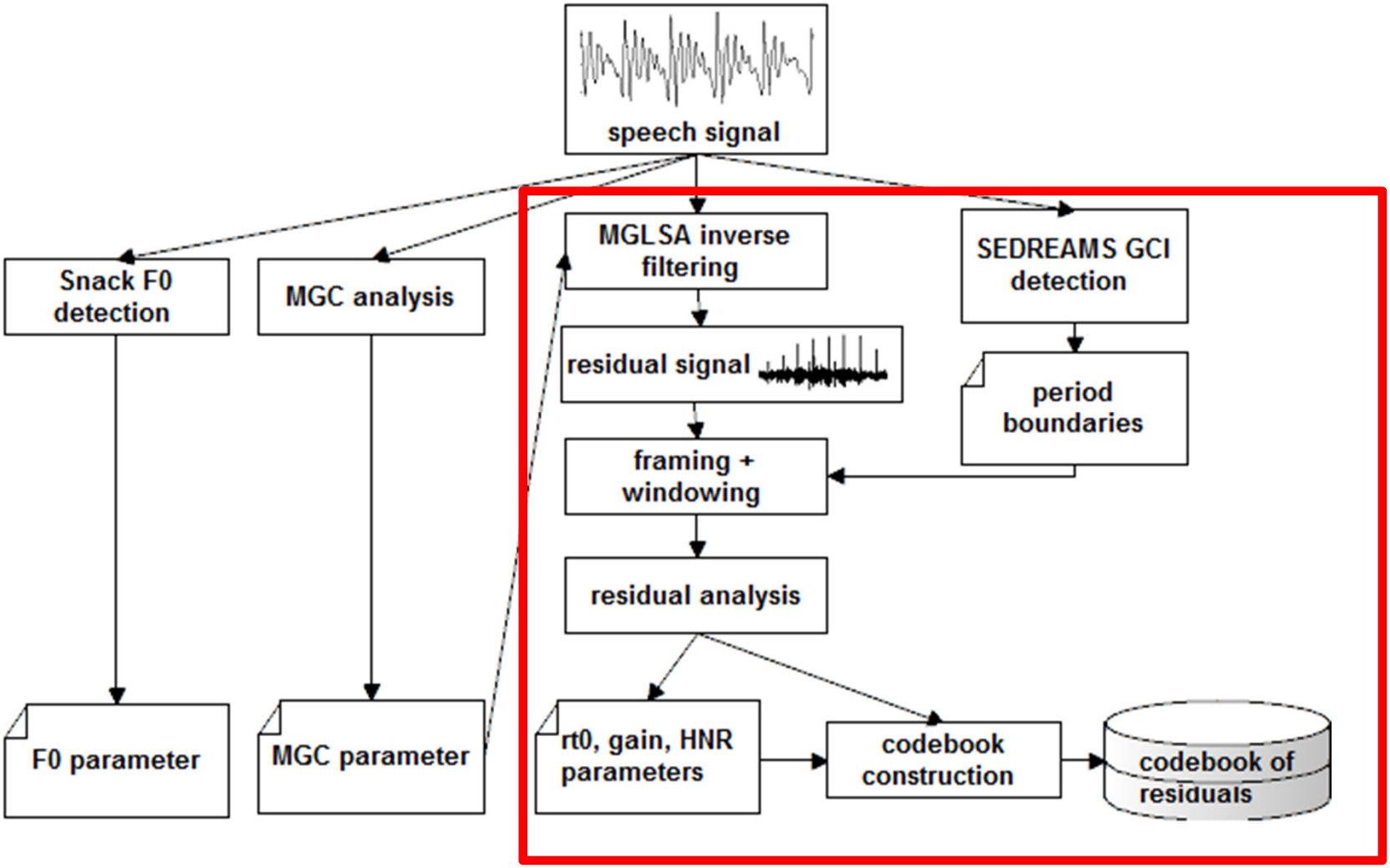
- HTS-CDBK [Csapó&Németh'12]
  - Residual based
  - MGC analysis
  - Codebook of pitch-synchronous residuals
  - White noise above 6 kHz
- Parameters
  - MGC: Mel-Generalized Cepstrum
  - F0: of the frame
  - gain: RMS energy of the windowed frame
  - rt0 peak indices: the locations of peaks in the frame
  - HNR: Harmonics-To-Noise ratio of the frame [de Krom'93]

Baseline: HTS-CDBK rt0 parameter

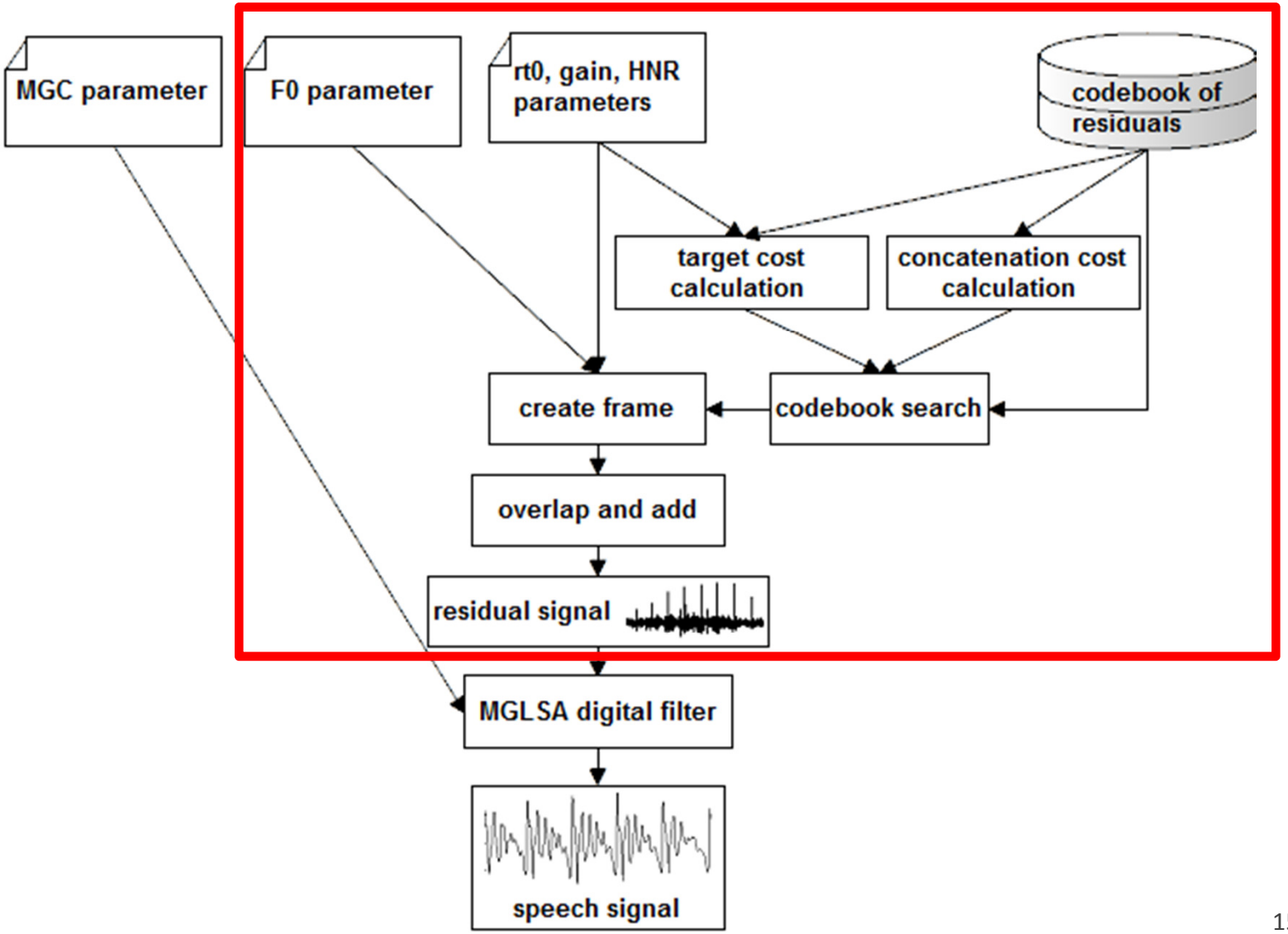
- position of peaks (distance)
- simple peak picking
- suitable for machine learning



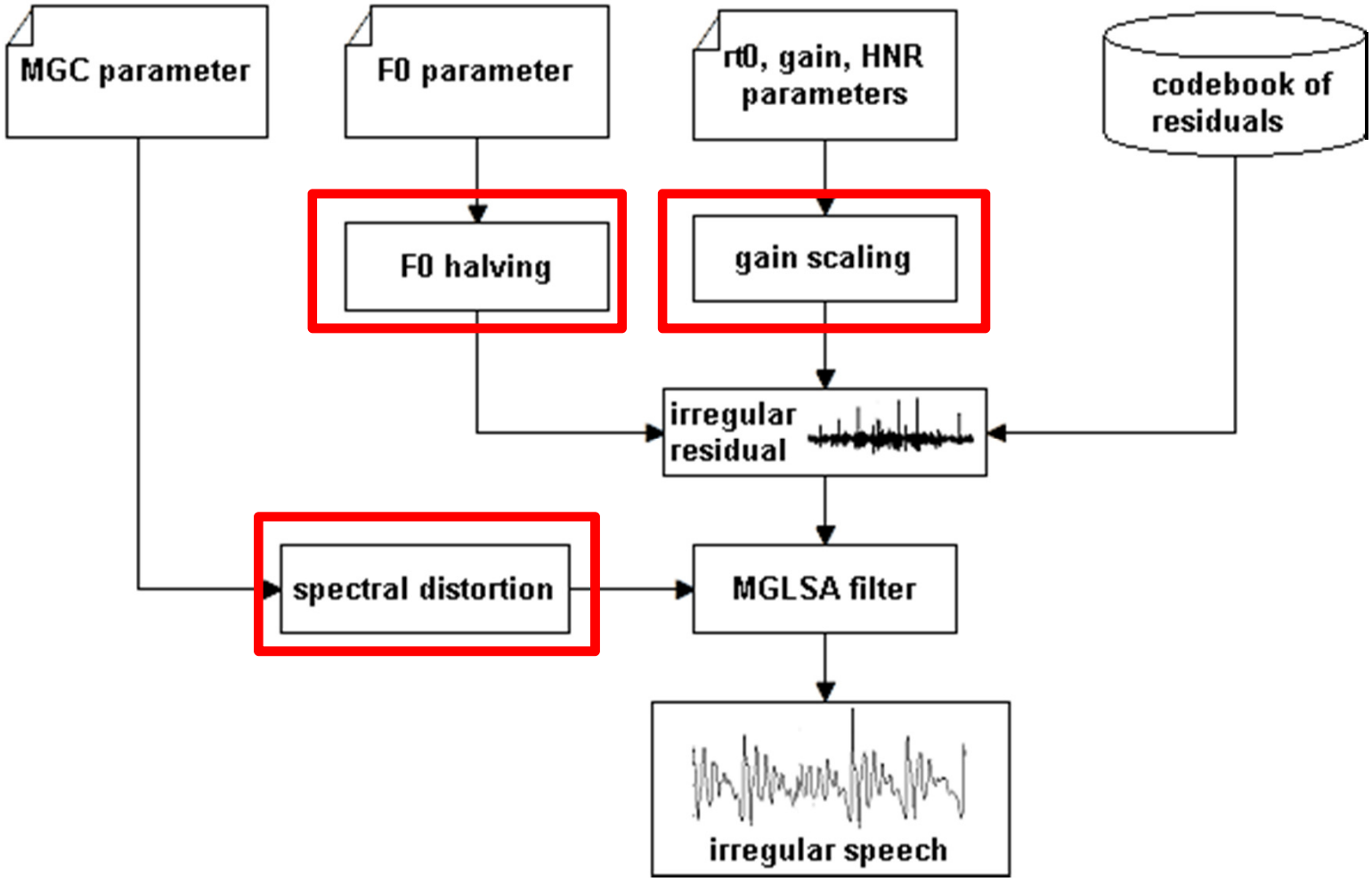
Baseline: HTS-CDBK analysis



Baseline: HTS-CDBK synthesis



Novel: HTS-CDBK+Irreg-Rule synthesis



Heuristic #1: F0 halving

- Irregular speech: often significantly lower F0 than regular speech

- Synthesis: half of the F0 of the generated parameter sequence is used
  - Residual frames are zero padded
  - Similar effect as removing every 2nd pitch cycle
  - Results in decreased open quotient

Heuristic #2: gain scaling

- Irregular speech: often strong amplitude attenuations during the consecutive cycles

- Synthesis: residual frames are multiplied by random scaling factors in the range of {0..1}
  - do not boost any of the periods, only attenuate or leave them unchanged

Heuristic #3: Spectral distortion

- Irregular speech: frame-by-frame MGC parameters are less smooth than those of regular speech

- Synthesis: distort MGC parameters
  - parameter values are multiplied by random numbers between {0.995…1.005}
  - yields less smooth parameter sequence

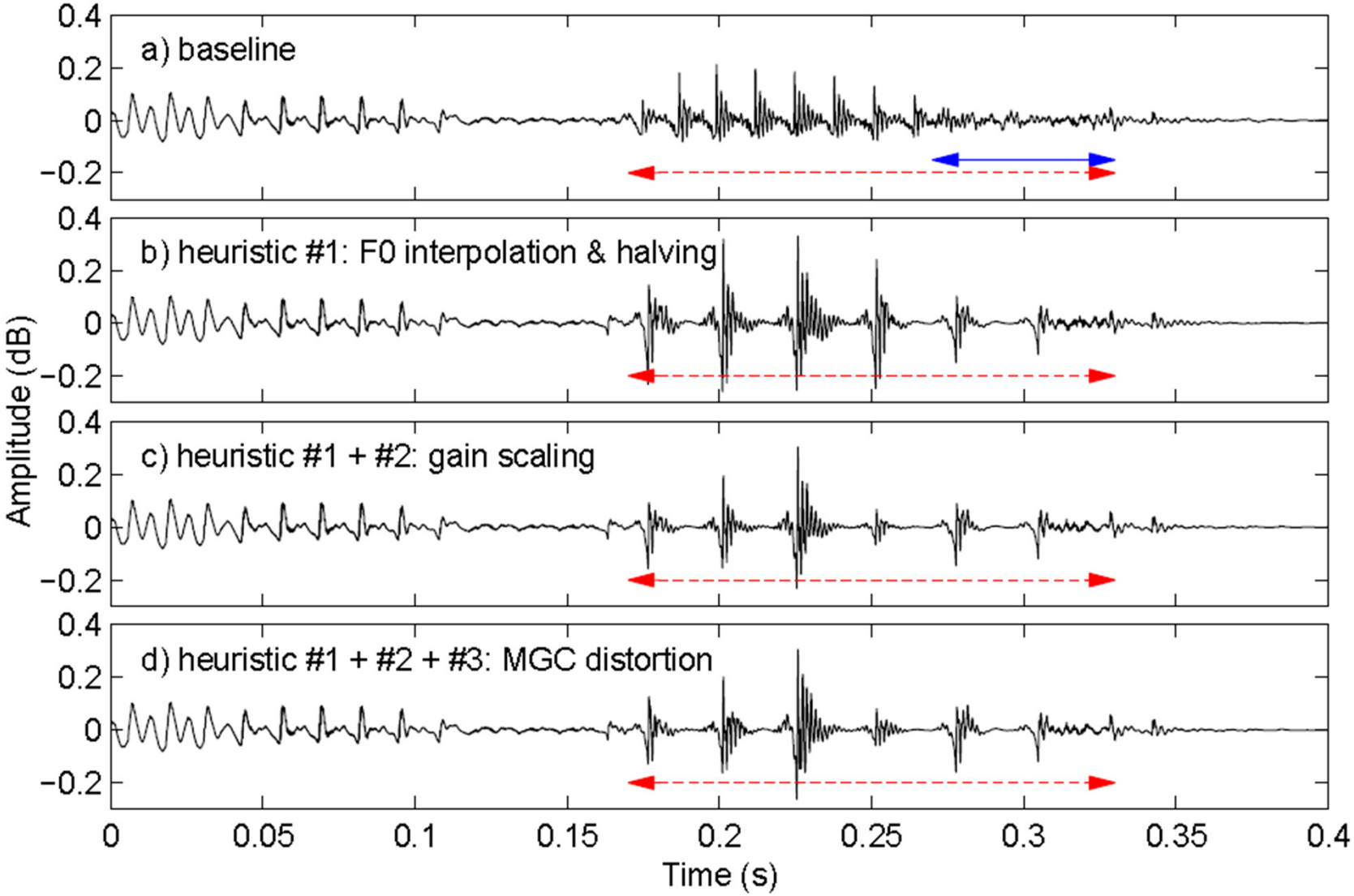
Position of irregular speech

- Irregular speech: often causes F0 detection errors in sentence-final vowels (F0=0)

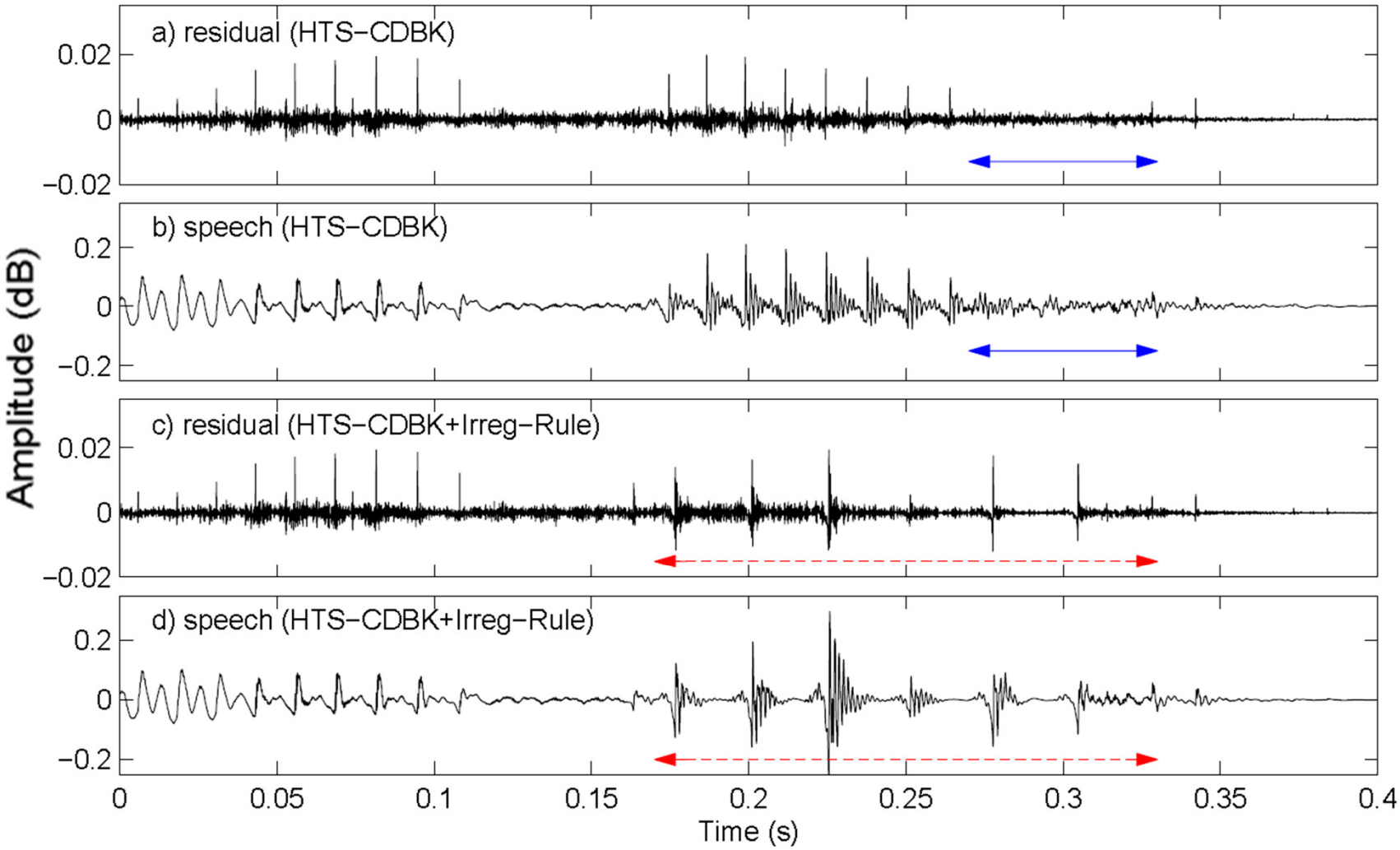
- Synthesis: F0=0 pattern of sentence-final vowels is modeled by machine learning
  - Irregular voice applied if 5 consecutive frames have F0=0
  - Indirect method for position of creaky voice
  - F0 interpolation between voiced parts

RESULTS

Waveforms: 3 heuristics



Residuals + speech: baseline vs. novel



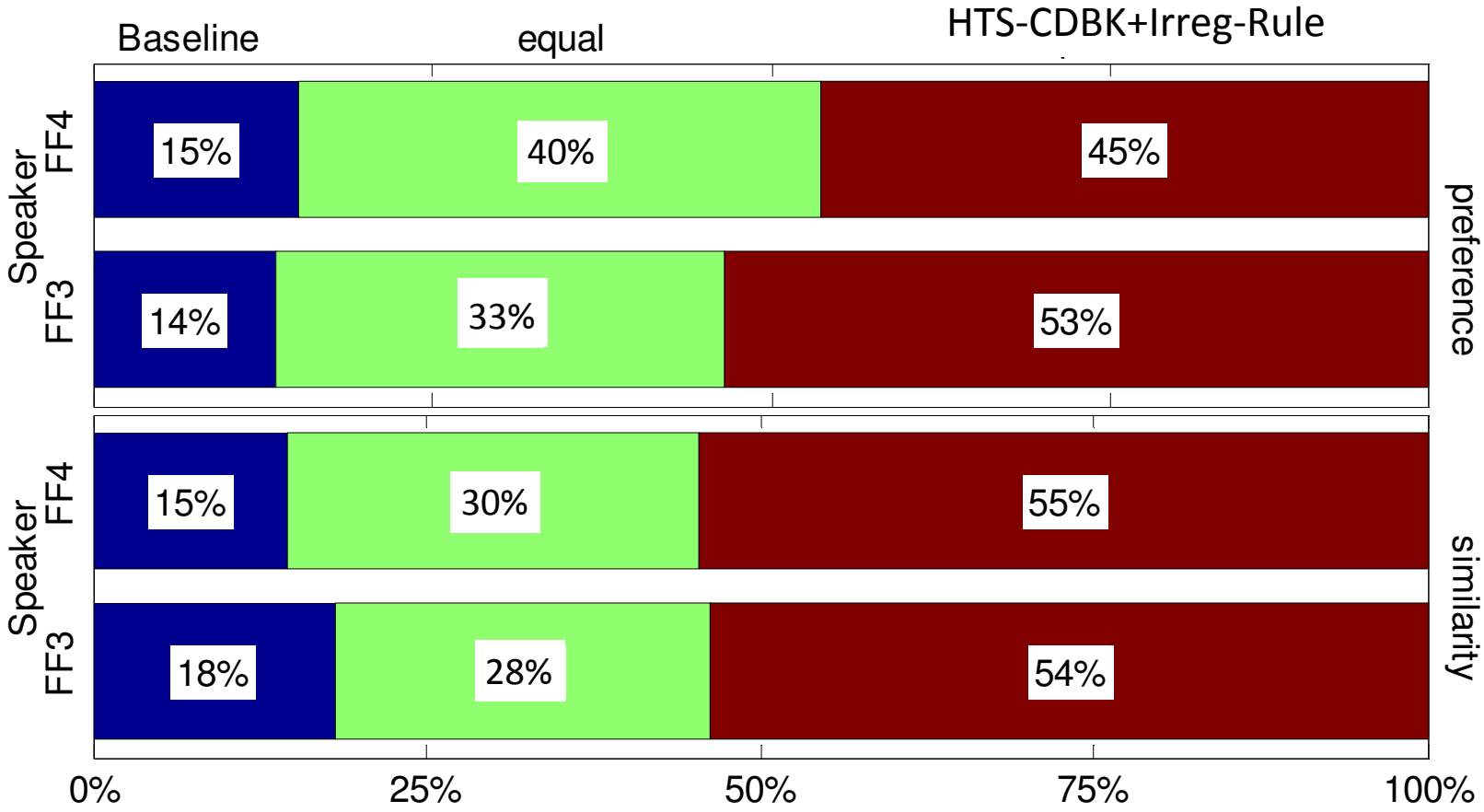
Perceptual evaluation: speech data

- 2 Hungarian male speakers with frequent irregular voice
  - About 2 hours of speech (1940 sentences)
  - 16 kHz, 16 bit waveforms + labels
  - Single speaker training with HTS-CDBK and HTS-CDBK+Irreg-Rule
  - 10-10 synthesized samples from baseline and novel systems
    - words from sentence endings with irregular voice

Perceptual evaluation: methods

- Internet-based test
  - Paired comparison
- Questions: Comparative MOS (CMOS)
  - 1: preference (‘Which version do you think is more pleasant?’)
  - 2: similarity to the original speaker (‘Which version is more similar to the original speaker?’)
- Listeners
  - 11 students and professionals

Perceptual evaluation: results

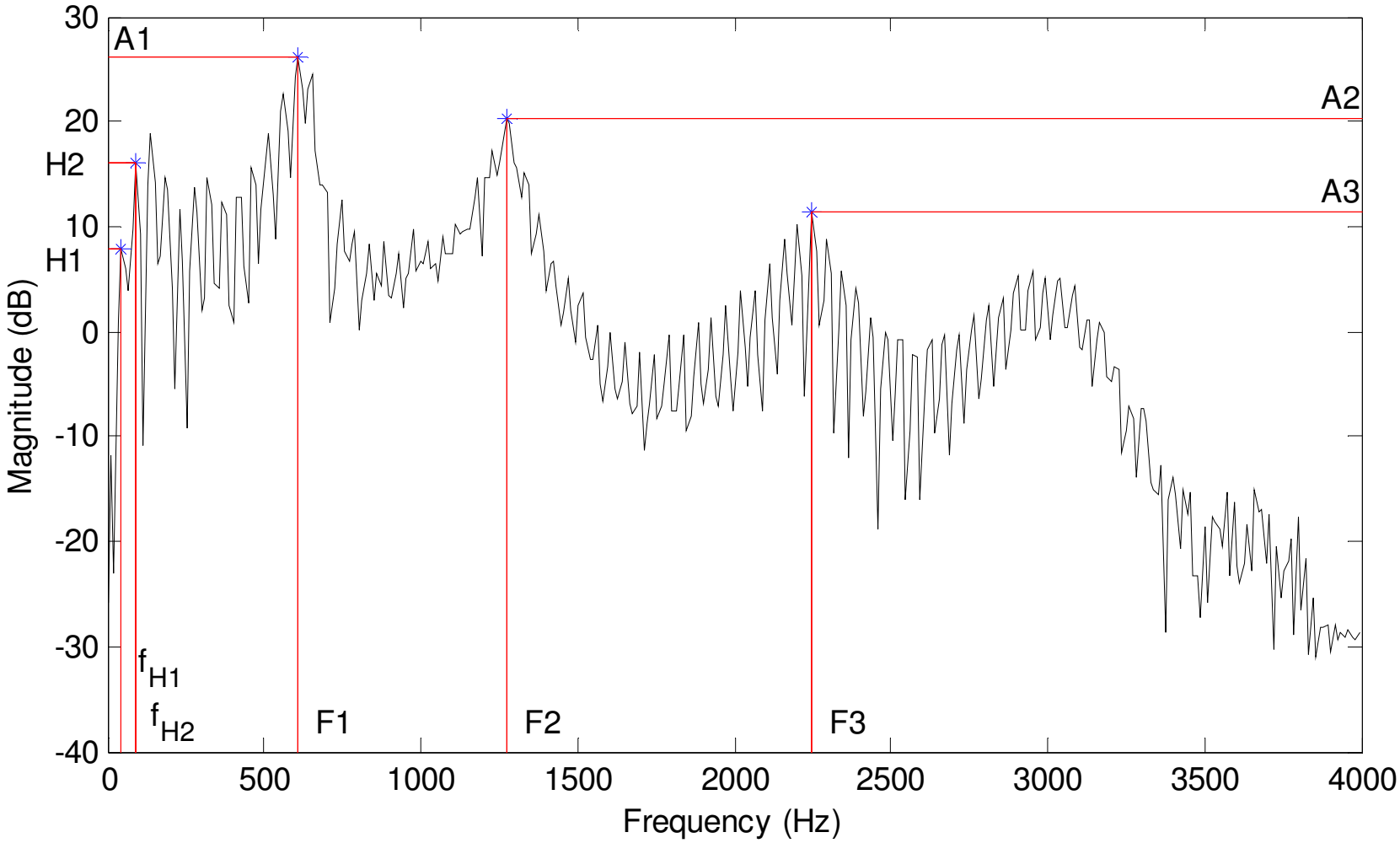


- Significant differences ($p<0.0005$) for proposed model

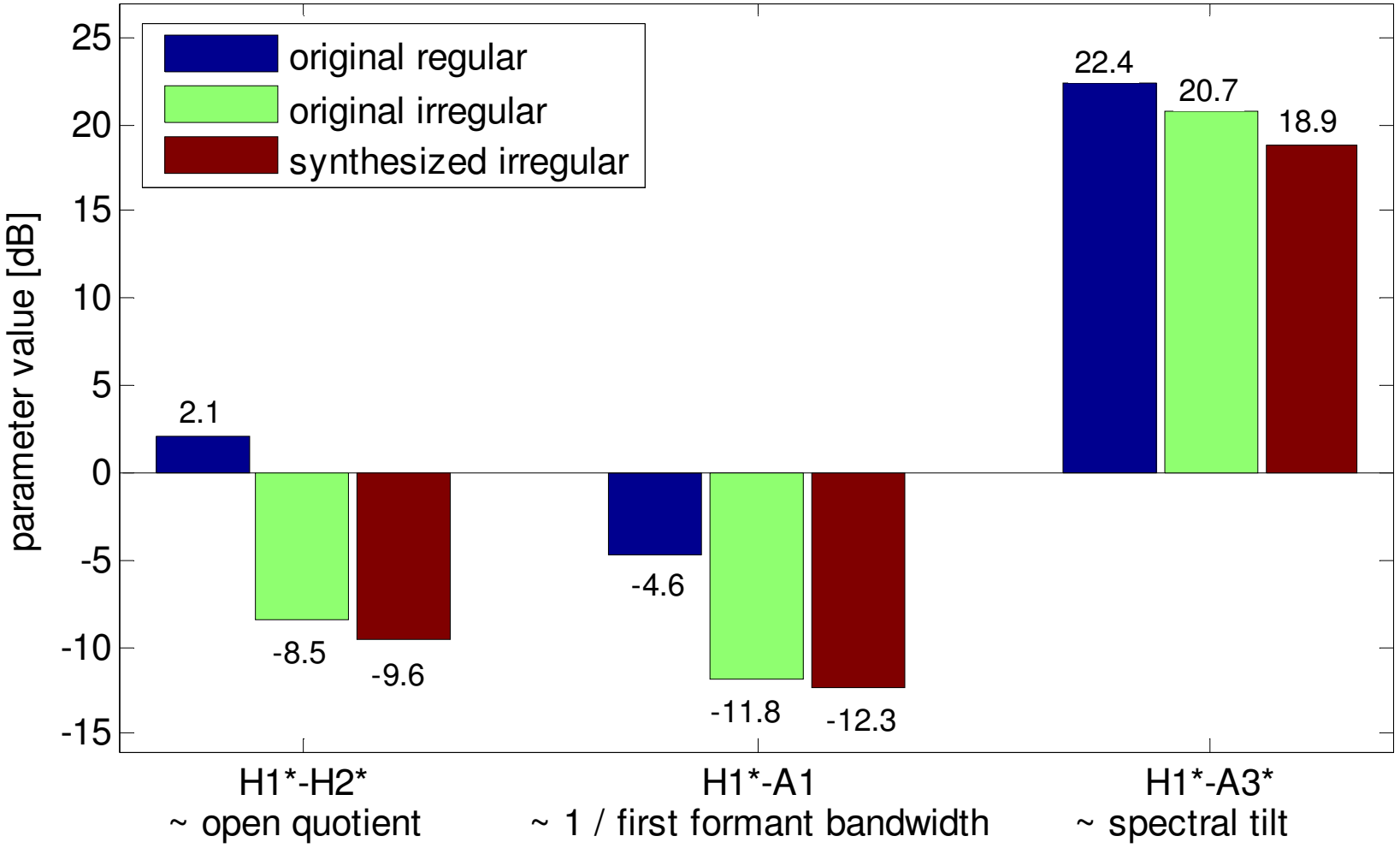
Acoustic evaluation: methods

- Acoustic cues: irregular vs. regular speech
  [Klatt&Klatt'90] [Bőhm;'09]
  - lower open quotient (OQ)
  - increased first formant bandwidth (B1)
  - lower spectral tilt (TL)
- Measurement in the frequency domain
  - OQ ~ H1-H2 (the difference of the amplitudes of the first two harmonics)
  - 1/B1 ~ H1-A1 (H1 relative to the first formant amplitude)
  - TL ~ H1-A3 (H1 relative to the third formant amplitude)
  - compensation of the first three formants
- Samples
  - 10 original regular, 10 original irregular, 10 synthesized irregular

Acoustic evaluation: measurements



Acoustic evaluation: results



SUMMARY

Discussion and conclusions

- Irregular phonation: no strict definition
- 3 heuristics to model in synthesis
  - Extremely low F0
  - Amplitude attenuations
  - Perturbations in spectrum
- Perception & acoustic tests
  - More preferred and more similar to original speaker
  - Similar to original irregular samples
- Possible applications
  - Expressive speech synthesis (e.g. sad)
  - Personalized systems

Future directions

- Pre-defined stylized pulse patterns instead of random scaling [Bőhm;'09]
- Data-driven irregular voice model
  - Csapó & Németh ,,Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation'', IEEE Journal of Selected Topics in Signal Processing, Oct 2013
- Use parameters for irregular voice position [Drugman;'13]
- Compare with other models [Drugman;'12] [Raitio;'13]

Tamás Gábor Csapó, Géza Németh:
A novel irregular voice model for HMM-based speech synthesis

csapot@tmit.bme.hu

This research is partially supported by the following projects:

- Paelife (Grant No AAL-08-1-2011-0001)
- CESAR (Grant No 271022)
- EITKIC_12-1-2012-001
- Campus Hungary

MAGYARORSZÁG MEGÚJUL

A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

References

- Blomgren, M. et al., 1998. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *The Journal of the Acoustical Society of America*, 103(5), pp.2649–2658.
- Bőhm, T. et al., 2008. Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. In *Acoustics'08*. Paris, France, pp. 6141–6146.
- Csapó, T.G. & Németh, G., 2012. A novel codebook-based excitation model for use in speech synthesis. In *IEEE CogInfoCom*. Kosice, Slovakia: IEEE, pp. 661–665.
- Csapó, T.G. & Németh, G., 2013a. Statistical parametric speech synthesis with a novel codebook-based excitation model. *Intelligent Decision Technologies*.
- Csapó, T.G. & Németh, G., 2013b. Transformation of irregular voice to regular voice by residual analysis and synthesis. *IEEE Signal Processing Letters*.
- Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M., 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4), pp.423–444.
- Drugman, T. et al., 2013. Prediction of Creaky Voice from Contextual Factors. In *Proc. ICASSP*. Vancouver, Canada.
- Drugman, T., Kane, J. & Gobl, C., 2012. Modeling the Creaky Excitation for Parametric Speech Synthesis. In *Proc. Interspeech*. Portland, Oregon, USA, pp. 1424–1427.
- Drugman, T., Wilfart, G. & Dutoit, T., 2009. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. Interspeech*. Brighton, UK, pp. 1779–1782.
- Fant, G., Liljencrants, J. & Lin, Q., 1985. A four-parameter model of glottal flow. *STL-QPSR*, 4, pp.1–13.
- Gobl, C. & Chasaide, A.N., 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2), pp.189–212.
- Klatt, D.H. & Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2), pp.820–857.
- De Krom, G., 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, 36(2), pp.254–266.
- Raitio, T. et al., 2013. HMM-based synthesis of creaky voice. In *Proc. Interspeech*.
- Silén, H. et al., 2009. Parameterization of vocal fry in HMM-based speech synthesis. In *Proc. Interspeech*. Brighton, UK, pp. 1775–1778.
- Zen, H., Toda, T., et al., 2007. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Transactions on Information and Systems*, E90-D(1), pp.325–333.
- Zen, H., Nose, T., et al., 2007. The HMM-based speech synthesis system version 2.0. In *Proc. ISCA SSW6*. Bonn, Germany, pp. 294–299.

Samples

- FF3_HTS-CDBK  + Irreg-Rule 
- FF3_HTS-CDBK  + Irreg-Rule 
- FF4_HTS-CDBK  + Irreg-Rule 
- FF4_HTS-CDBK  + Irreg-Rule