Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics (BME TMIT), Budapest, Hungary

Automatic transformation of irregular to regular voice by residual analysis and synthesis

Tamás Gábor Csapó, Géza Németh, *{csapot,nemeth}@tmit.bme.hu*

1. Introduction

- **creaky voice**, laryngealization, vocal fry, glottalization
  - irregular vibration of vocal folds
  - irregular F0 and/or amplitudes (see Fig. 1)
- occurrence: up to 15% of vowels of natural speech
  - phrase boundaries
  - sentence endings
  - vowel-vowel transitions
- differences compared to regular speech (see Fig. 1)
  - 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
  - increased first formant bandwidth
  - more abrupt closure of the vocal folds
- irregular voice in speech technology
  - analysis of glottalization
  - automatic detection
  - transforming regular to irregular
  - voice conversion
  - synthesis of glottalization
- **previously no automatic irregular-to-regular transformation method was available**
  - our earlier work: semi-automatic [1], but manual correction of F0 curve necessary
- goal of this paper
  - **automatic irregular-to-regular conversion**
  - using a continuous F0 model [2]
  - applying an analysis-synthesis vocoder [3]

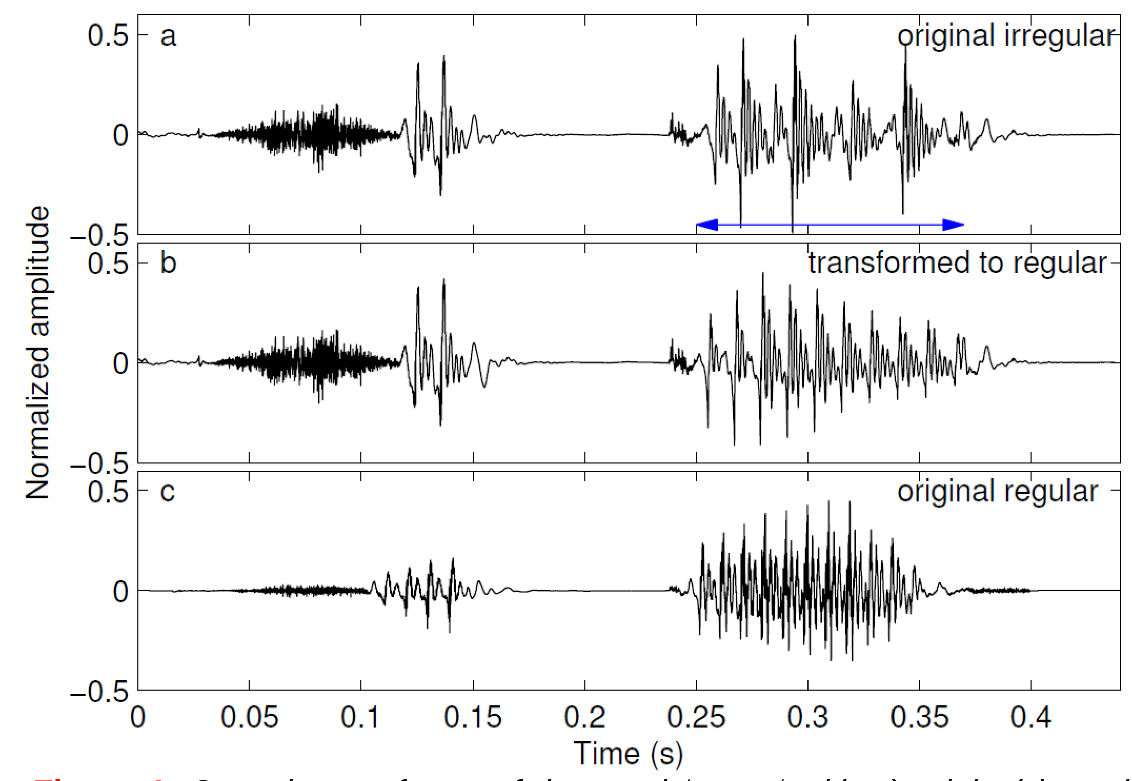

Figure 1. *Speech waveforms of the word /tsipɒː/ with a) original irregular ending (horizontal arrow: irregular voice), b) its transformed version to regular, c) another realization of the same word with original regular ending.*

2. Transformation method

- effect of irregular voice on pitch estimation
  - causes errors in standard methods (see Fig. 2, b)
  - interpolation necessary (see Fig. 2, c)
- **continuous pitch tracking** (CONT_F0, [2])
  - standard autocorrelation
  - no voiced/unvoiced decision
  - Kalman smoothing-based interpolation
- irregular-to-regular transformation (see Fig. 4)
  - based on our analysis-synthesis vocoder [3]
  - codebook of **pitch synchronous residuals**
  - parameters: F0,
    - gain: frame by frame energy
    - rt0: prominent values in the frame (see Fig. 5)
    - HNR: Harmonics-to-Noise Ratio
    - MGC: Mel-Generalized Cepstrum
  - automatic irregular voice detection using [5]
  - result: quasi-periodic speech signal (see Fig. 1, 3)

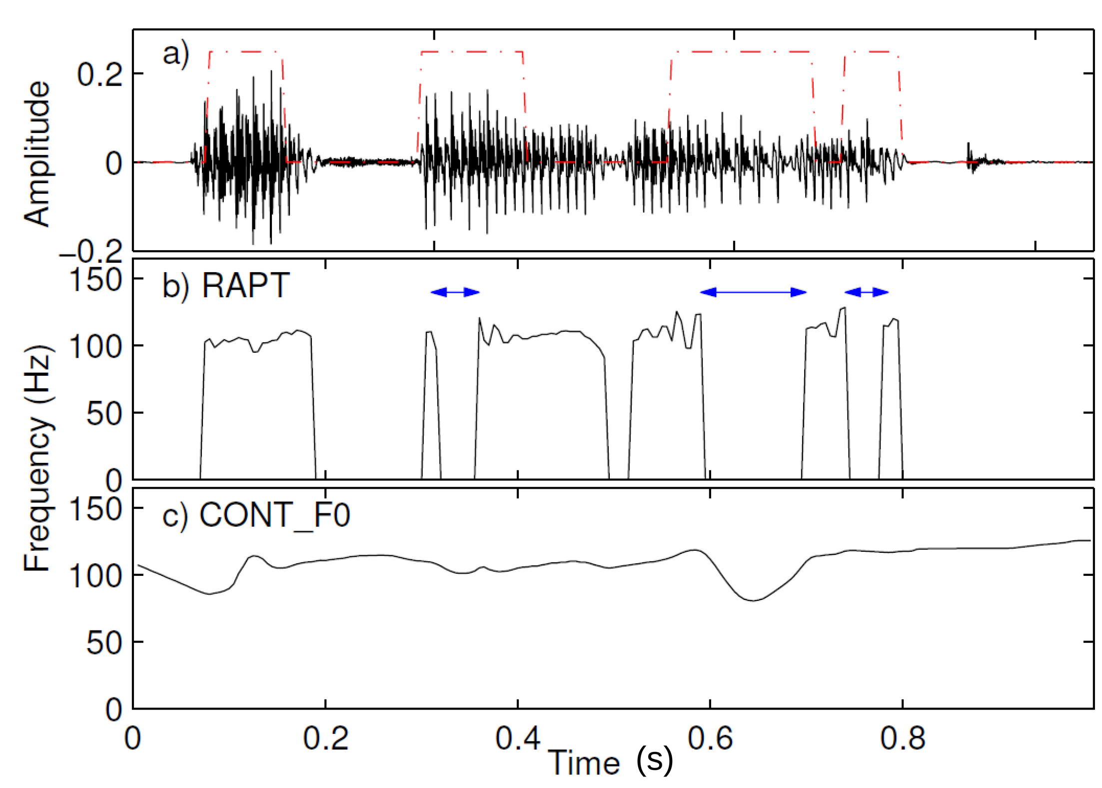

Figure 2. *Effect of creaky voice on pitch estimation. a) a sample speech waveform and regions of irregular voice (red dashed line); b) pitch estimated by RAPT [4] (blue arrows indicate inaccurate pitch estimation); and c) pitch estimated by the CONT_F0 pitch tracker [2].*

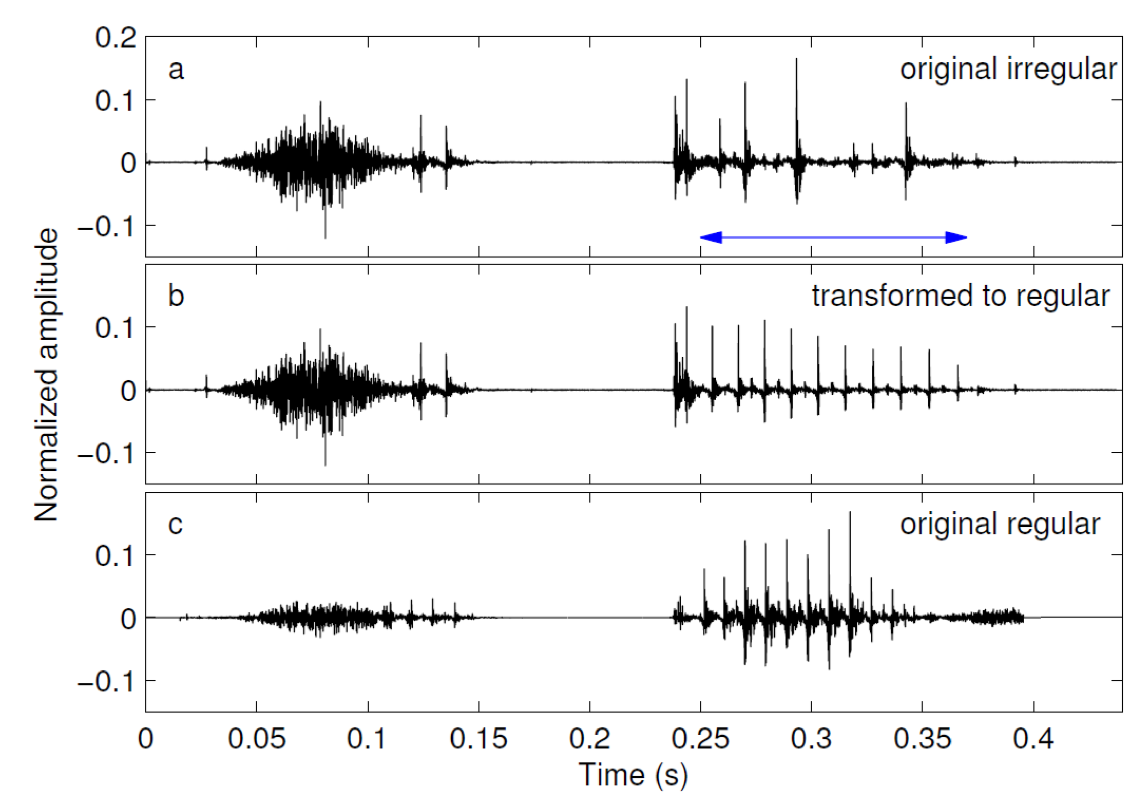

Figure 3. *Residuals of speech recordings of Figure 1: a) residual with original irregular ending (horizontal arrow: irregular voice), b) its transformed version to regular, c) is the residual of another realization of the same word with original regular ending.*

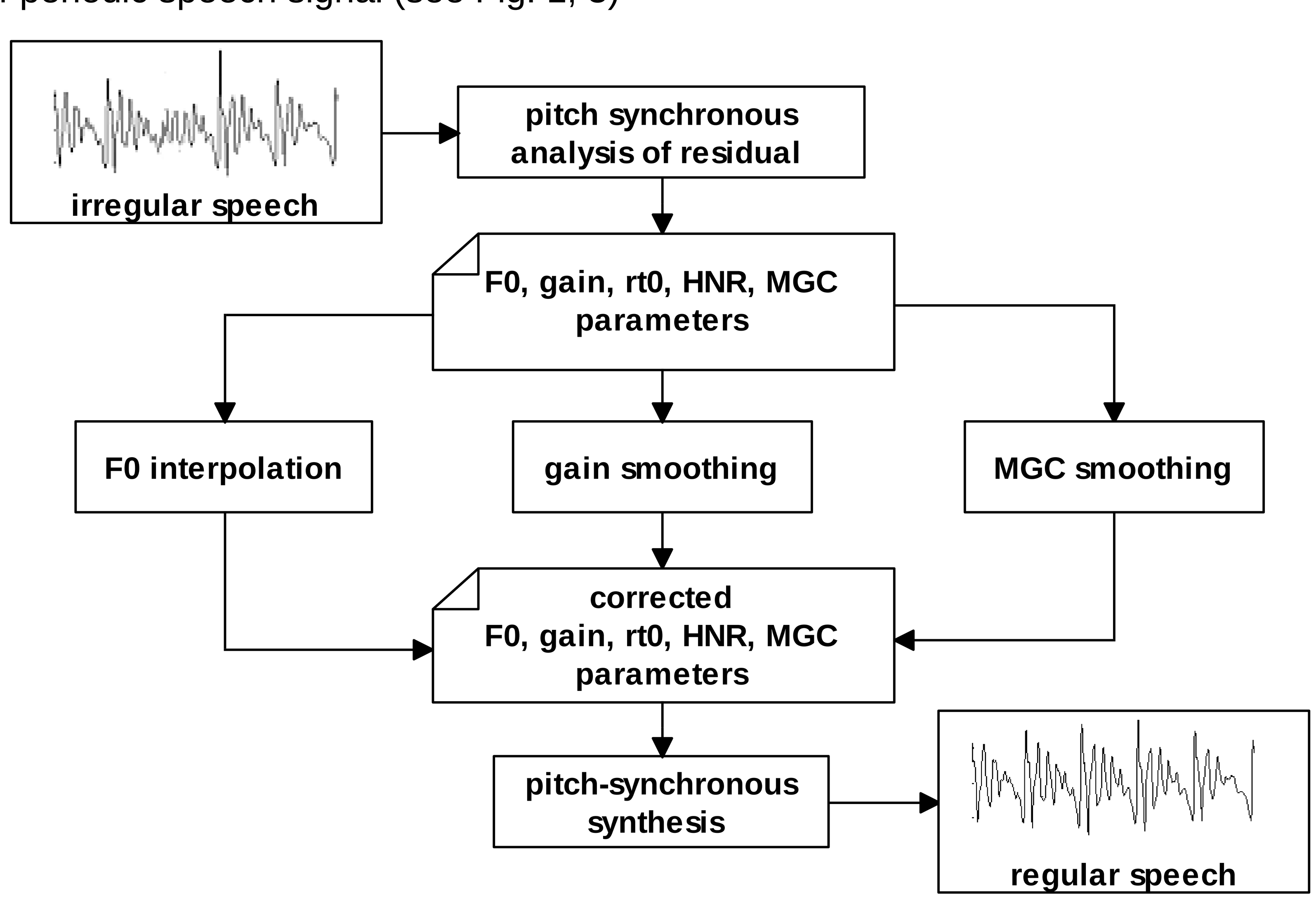

Figure 4. *Simplified block diagram of the irregular-to-regular transformation method.*

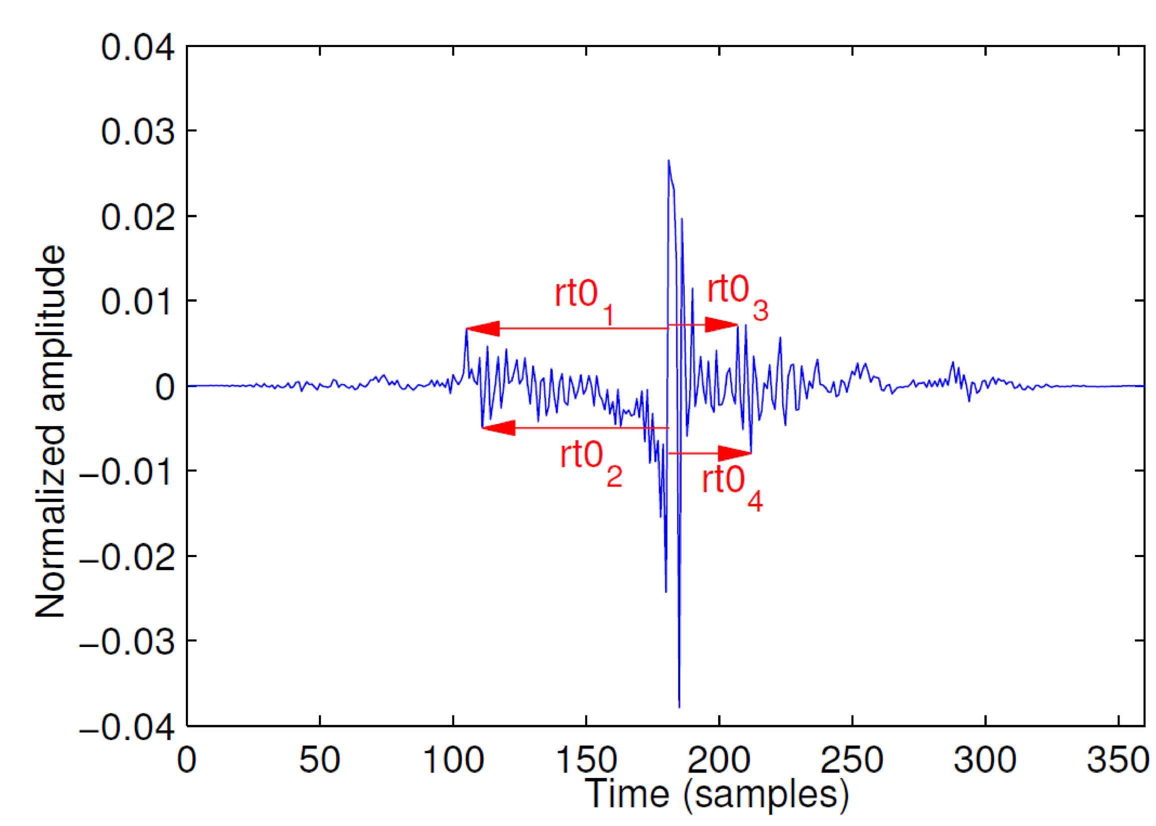

Figure 5. *Calculation of the rt0 parameter for a windowed residual segment. $rt0_i$ is the distance of prominent peaks from the main impulse, in samples.*

3. Perceptual evaluation

- stimuli for the perceptual evaluation
  - four Hungarian speakers (three males: FF1, FF3, FF4 and one female: NO3)
  - five sentences from each speaker, containing irregular voice in at least 15%
  - utterance versions having irregular sections were transformed to modal voice by the proposed method
- web-based listening test: „roughness" and „naturalness" of samples
  - both versions of each sentence (original irregular and transformed to regular), altogether 40 utterances (4 speakers * 5 sentences * 2 versions)
  - two 5 point MOS-like questions:
    - roughness ('1 – very rough' … '5 – not rough at all')
    - naturalness ('1 – very unnatural' … '5 – very natural')
  - 13 listeners (11 males, 2 females), native speakers of Hungarian, university students or speech technology experts; on average 8 minutes to complete
- results of the listening test (see Table 1)
  - for speakers FF1, FF3 and FF4, **the method was able to decrease the perceived roughness** (only significant for speaker FF4)
  - for speaker NO3, the transformation slightly increased the roughness
  - naturalness slightly decreased (significant for FF1, FF3 and NO3)

	roughness		naturalness	
speaker	original	transf.	original	transf.
FF1	2.77 (0.93)	2.92 (1.24)	3.71 (1.04)	2.49 (1.03)
FF3	2.80 (1.28)	2.89 (1.24)	3.94 (1.03)	3.02 (1.08)
FF4	2.89 (1.05)	3.26 (1.03)	3.94 (0.88)	3.26 (1.11)
NO3	3.71 (1.13)	3.69 (1.18)	3.88 (0.93)	2.80 (1.20)

Table 1. *Speaker by speaker Means and standard deviations (in parenthesis) for the roughness and naturalness questions.*

- potential reasons for unnaturalness
- interpolation of CONT_F0 sometimes inaccurate (contradicts to natural F0 contour)
- vocoder might cause 'buzzy' voiced quality
- 34th order of the MGC analysis may be too high for the female speaker

4. Discussion and Conclusions

- **fully automatic method to transform irregular voice to regular voice**
  - codebook-based residual analysis-synthesis
  - original irregular sections replaced by overlap-added frames from codebook
  - more suitable than direct waveform manipulation like PSOLA with hand-crafted weights, as the residual can be corrected automatically
- Kalman **smoothing of CONT_F0**: more suitable than the simple linear F0 interpolation we used in [1]
  - in some cases it causes high F0 at the end of the sentence (unnatural)
  - solution might be: combine F0 interpolation with rule-based intonation model
- it is known that **several types of glottalization** can be differentiated
  - our method was suitable for transforming the type used by speaker FF4
- applications of the model may include
  - correction of voices where unwanted irregular phonation occurs frequently (e.g. those of radio announcers or voice actors)
  - transform glottalized parts of large speech databases (help further **automatic speech processing**; voice conversion)

Key references

[1] T. G. Csapó and G. Németh, "Irreguláris beszéd regulárissá alakítása beszédkódoláson alapuló módszerrel [Transforming irregular speech to regular speech based on voice coding] (in Hungarian)," *Beszédkutatás 2014 [Speech Research 2014]*, pp. 193–204, 2014.
[2] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, 2013.
[3] T. G. Csapó and G. Németh, "A novel codebook-based excitation model for use in speech synthesis," in *IEEE CogInfoCom*, 2012, pp. 661–665.
[4] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.
[5] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech and Language*, vol. 28, no. 5, pp. 1233–1253, 2014.

Acknowledgements

This research is partially supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech prosody (SNSF n° IZ73Z0_152495-1).