2865. Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles

Tamás Bőhm[1], Nicolas Audibert[2], Stefanie Shattuck-Hufnagel[3], Géza Németh[1], Véronique Aubergé[2]

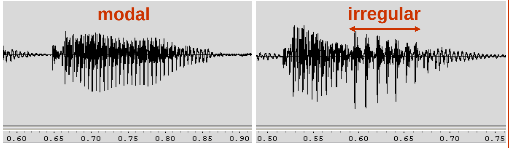
[1]Department of Telecommunications and Media Informatics, BME, Budapest, Hungary
[2]GIPSA-lab, Speech & Cognition Dept (ICP), CNRS UMR 5216, Grenoble, France
[3]Research Laboratory of Electronics, MIT, Cambridge, MA USA

MŰEGYETEM 1782

Massachusetts Institute of Technology

Introduction

- Irregular phonation (vs. modal): voiced speech with irregular time or amplitude over adjacent pitch periods or with unusually low F_0 (Slifka 2007)
  - Also known as: creaky voice, vocal fry, glottalization, laryngealization
  - Perceived as 'rough voice'
  - Focus on utterance-final intermittent irregular phonation

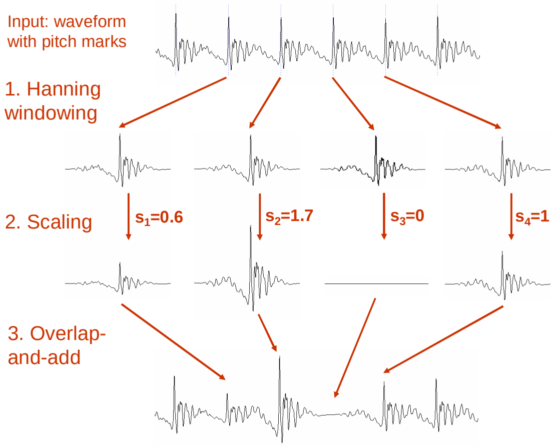


- Cue to stop consonants, prosodic boundaries, affective states, and speaker identity

→ Transforming between modal and irregular voice may contribute to speech synthesis technologies (naturalness, expressivity, personalization)

- Earlier work:
  - Formant synthesis: either unnatural (automatic) or laborious (manual) (Edgington 1997)
  - Increasing jitter and shimmer: does not predict perceived voice quality (Verma, Kumar 2005)
- Presented here: a simple, semi-automatic transformation method to introduce irregular pitch periods into modal speech
  - Approaches natural irregular phonation both perceptually and acoutically

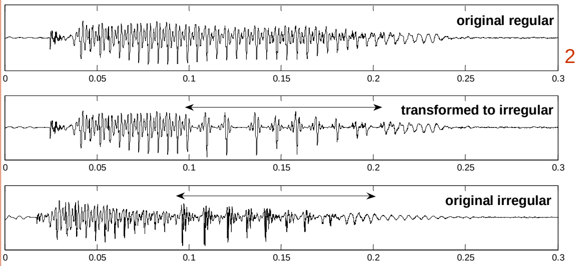
Transformation method


Input: waveform with pitch marks

1. Hanning windowing

2. Scaling

3. Overlap-and-add



- In contrast to PSOLA, the cycles are not moved in time – we need abrupt, substantial changes in glottal pulse spacings
- Background noise (e.g. from the end of the recording), Hanning-windowed and scaled with $max(1-s_i,0)$, can be added
- Scaling factors can be modeled after a region of natural speech with irregular phonation in two ways:

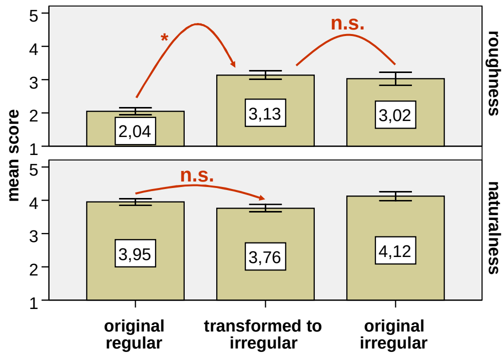
1. Setting the scaling factors manually:
   - If a cycle is 2x or 3x longer in the irregular than in the regular waveform: remove 1 or 2 cycles (s_i=0)
   - Reproduce the relative amplitudes of the irregular pulses
   - Iterative fine-tuning until perceptually acceptable

OR

2. Pattern copying: scaling factors ('stylized' pulse pattern) extracted from model waveform
   - If a cycle is 2x or 3x longer than the reference cycle length: appropriate number of 0's are inserted
   - Cycle amplitudes (relative to the reference cycle amplitude) are extracted
   - Reference cycle length and amplitude: mean of 5 values before the irregular region

Perceptual evaluation

- Recordings:
  - 4 American English speakers: 2 habitually used irregular phonation, 2 seldom used it
  - 4 words recorded with both regular and irregular utterance-final phonation
- Regular recordings were transformed to irregular
- Stimulus set:
  - original regular (16 stimuli),
  - original irregular (8 stimuli),
  - transformed to irregular (16 stimuli)
- 12 listeners rated the roughness and naturalness of each stimulus separately on a 5-point scale
- Before naturalness ratings: all the stimuli played
- Before roughness ratings: examples played
- One-way ANOVA for both rating tasks with Tukey's post hoc tests (95% level)
- Transformed speech sounds natural and as rough as naturally-occurring irregular phonation



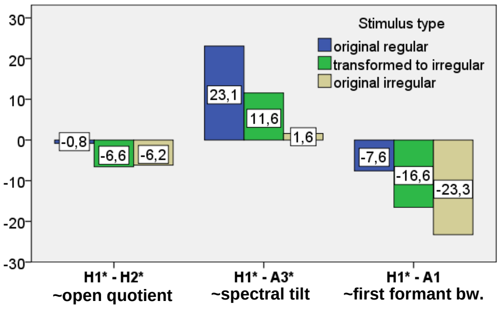
Acoustic evaluation

- Perceptual structure of irregular phonation: unclear but it has a number of acoustic correlates
- If transformed utterances match most of these, it may explain their perceptual salience

1. Low F0 and/or high jitter ✓
2. Low intensity ✓
3. Low open quotient ?
4. High first formant bandwidth ?
5. Low spectral tilt (abrupt vocal fold closure) ?

} tested by acoustic measurements

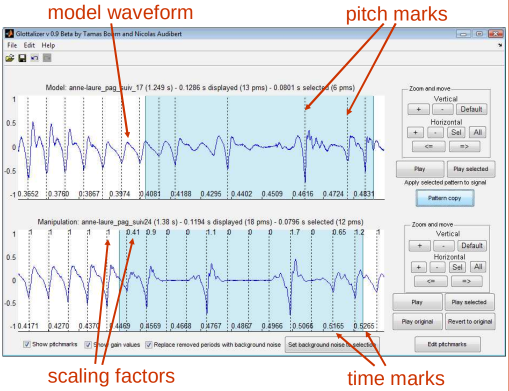
- Materials: stimuli used in perceptual evaluation
- Spectral correlates of 3-5 were measured (Holmberg et al. 1995) and corrected for formants (Hanson 1997)
- One-way ANOVAs for all three parameters with Tukey's post hoc tests (95% level)
- The transformation reproduces many of the acoustic correlates of irregular phonation (or, at least, approaches them)



Graphical program

- Allows fast and convenient application of the transformation method
- Freely available for non-commercial use:

www.bohm.hu/glottalizer.html



Model: anne-laure_pag_suiv_17 (1.249 s) - 0.1286 s displayed (13 pms) - 0.0801 s selected (6 pms)

Manipulation: anne-laure_pag_suiv24 (1.38 s) - 0.1194 s displayed (18 pms) - 0.0796 s selected (12 pms)

☑ Show pitchmarks ☑ Show gain values ☑ Replace removed periods with background noise

- Pitch mark file is needed to open a waveform
- Model waveform: can guide the transformation
- Pattern copy:
  - Select a region in the model waveform (to extract the 'stylized' pulse pattern from)
  - Select a region in the bottom panel (where the pattern is to be applied)
  - Have enough cycles before the selections (to calculate reference values)
- Usual sound displaying and playing functionalities

Edgington (1997): Investigating the limitations of concatenative synthesis, Eurospeech97, 593-596

Hanson (1997): Glottal characteristics of female speakers: acoustic correlates, JASA 101, 466-481

Holmberg, Hillman, Perkell, Guiod, Goldman (1995): Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice, JSHR 38, 1212-1223

Slifka (2007): Irregular phonation and its preferred role as cue to silence in phonological systems, ICPhS 2007, 229-232

Verma, Kumar (2005): Introducing rougness in individuality transformation through jitter modeling and modification, ICASSP2005, 5-8