BME FACULTY OF ELECTRICAL ENGINEERING AND INFORMATICS

MŰEGYETEM 1782

TMiT

SmartLab
Intelligent Interactions

Creaky Voice via Speaker Adaptation within End-to-End Text to Speech Synthesis

Ali Raheem Mandeel, Mohammed Salah Al-Radhi, and Tamás Gábor Csapó
Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary

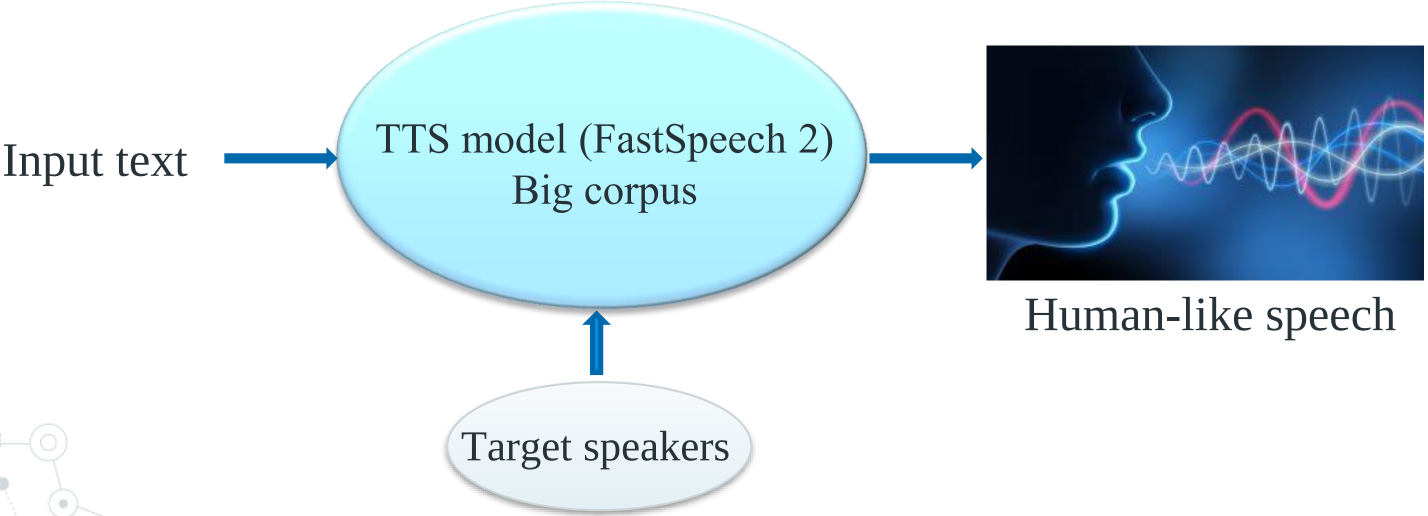
aliraheem.mandeel@edu.bme.hu

Outlines

- Introduction (speaker adaptation, creaky voice),
- Research objectives,
- Methods,
- Results,
- Conclusions.

Introduction \ Speaker Adaptation:

- It synthesizes speech of any individual,
- Used on a **few** speaker’s data / **low** computational resources.



Input text

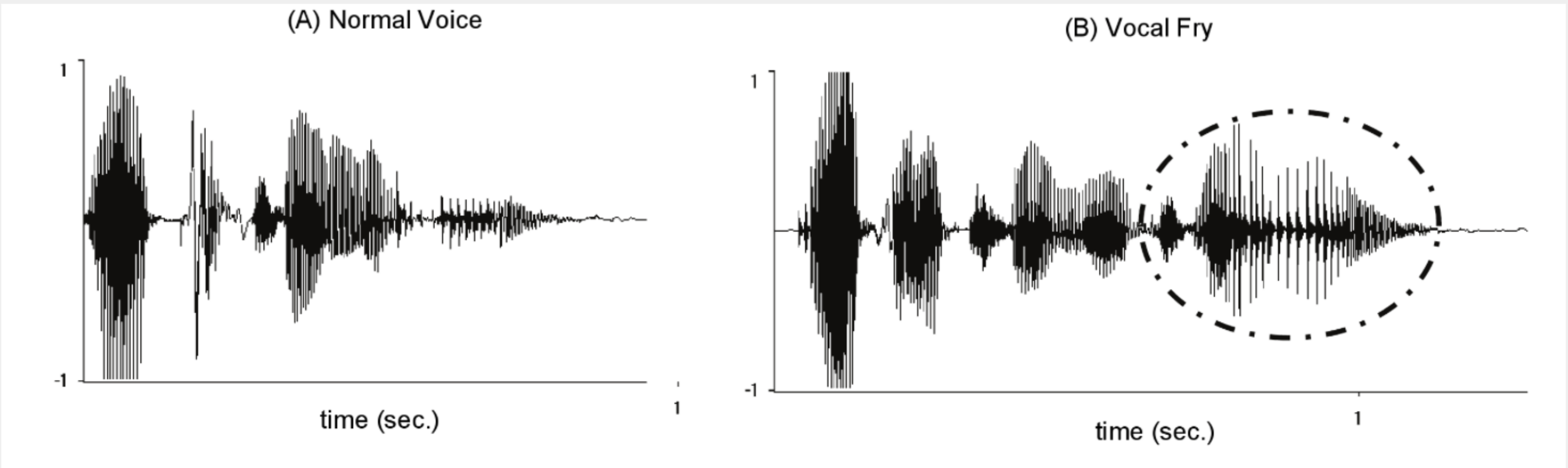
TTS model (FastSpeech 2)
Big corpus

Target speakers

Human-like speech

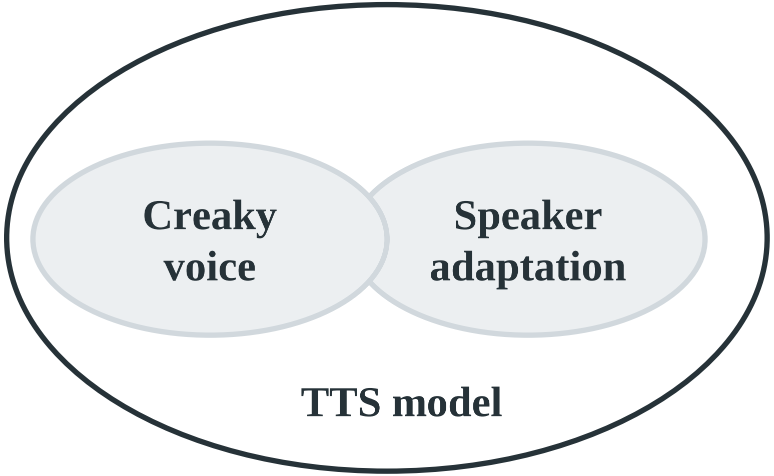
Introduction \ Creaky voice (vocal fry or glottalization):

- A speech pattern where the vocal cords are tightly constricted, causing a low-pitched, creaky sound.
- It is common in many languages / especially among young women.



Research objectives

- Investigate creaky voice into a TTS model using a limited dataset (speaker adaptation),



Methods

- End-to-end pretrained FastSpeech 2 model on LJSpeech (**English female speaker**),
- HiFi-GAN neural vocoder,
- 4 target speakers (two females and two males),
- We used 3 dataset types (Table 1), each one of only **100 English** sentences.

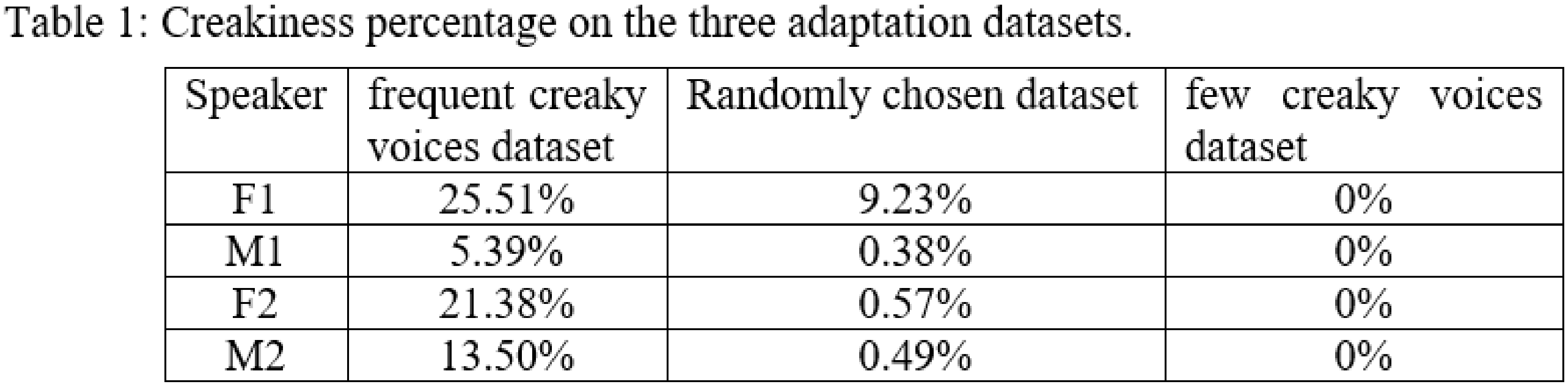


Table 1: Creakiness percentage on the three adaptation datasets.

Speaker	frequent creaky voices dataset	Randomly chosen dataset	few creaky voices dataset
F1	25.51%	9.23%	0%
M1	5.39%	0.38%	0%
F2	21.38%	0.57%	0%
M2	13.50%	0.49%	0%

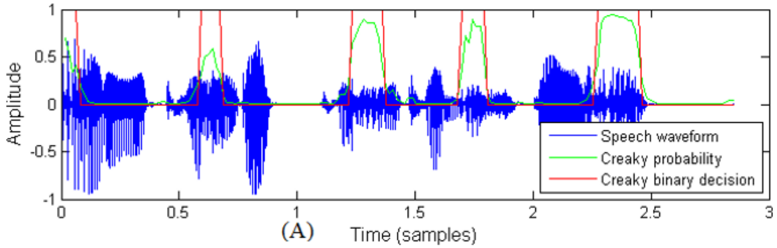
Results / Samples

- “To be modified by himself according to ever-changing circumstances.”

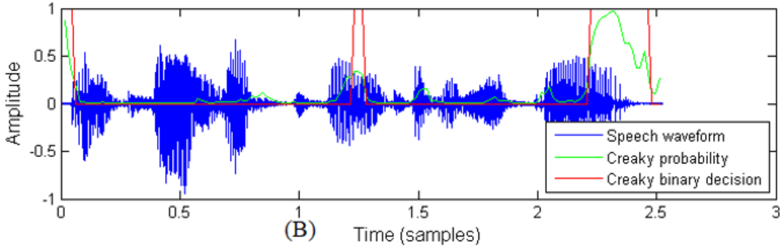
Reference (natural)	Synthesized sentence using the creaky dataset	Synthesized sentence using few creaky voices dataset	Synthesized sentence using the random dataset
🔊	🔊	🔊	🔊

Results / objective evaluation

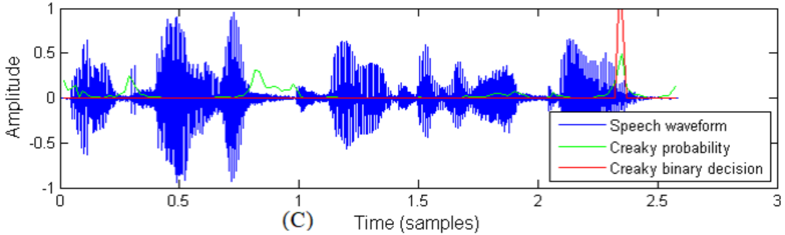
- The creakiness percentage,
- Jitter,
- Shimmer,
- Mean F0,
- Harmonic to Noise ratio (HNR).



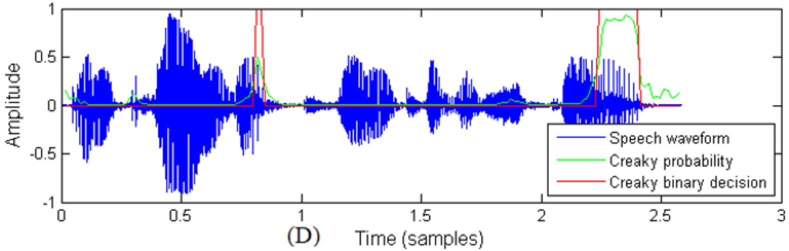
A) natural,



B) synthesized using the creaky dataset



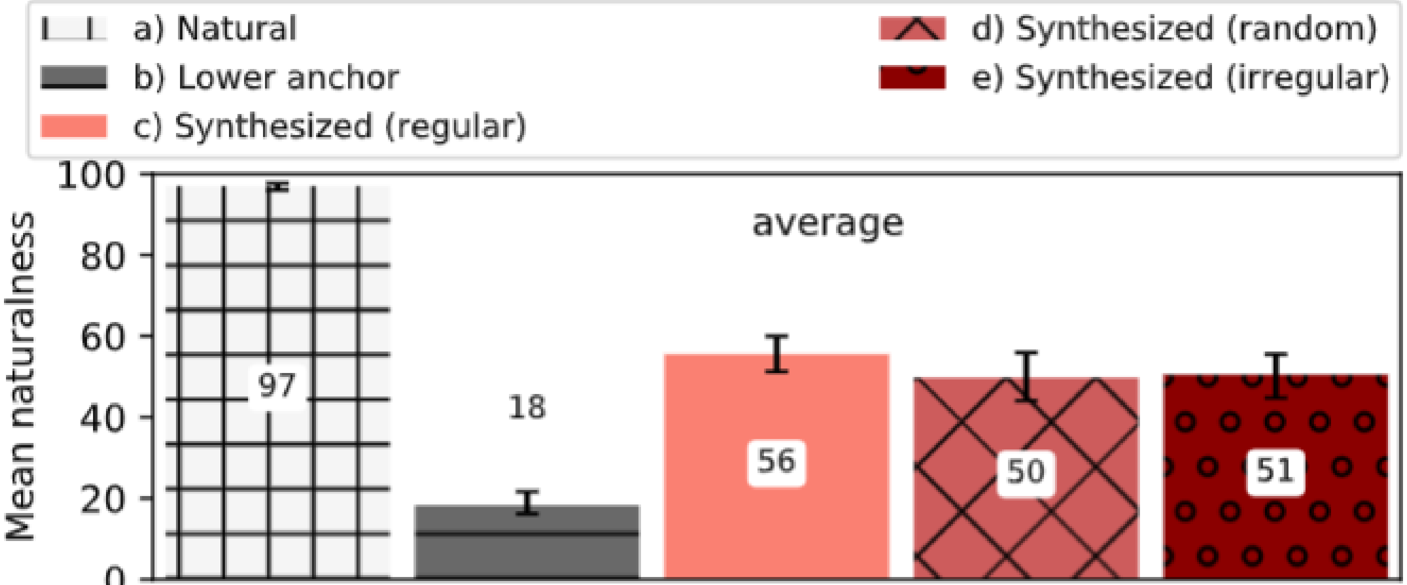
C) synthesized using few creaky voices



D) synthesized using random dataset

Results / subjective evaluation

- online MUSHRA- like test, 16 non-native individuals,
- The synthesized creaky voices received lower ratings.



Average similarity ratings of the four target speakers’ speech.

Conclusions

- Objective metrics: modeling a creaky voice is successful,
- Subjective results: Creaky speech has less rating than regular speech,
- Our findings help develop expressive, natural, and customized speech synthesis,
- More investigation of the acoustic features of the produced irregular voice samples.

References

[1] Alexandra Markó, Andrea Deme, Márton Bartók, Tekla Etelka Gráczi, Tamás Gábor Csapó, "Word-Initial Irregular Phonation as a Function of Speech Rate and Vowel Quality in Hungarian", International Seminar on Speech Production, ISSP 2017: Studies on Speech Production, Tianjin, Kína, pp. 134-145, 2018.

[2] Alexandra Markó, Andrea Deme, Márton Bartók, Tekla Etelka Gráczi, Tamás Gábor Csapó, "Speech Rate and Vowel Quality Effects on Vowel-related Word-initial Irregular Phonation in Hungarian", CHALLENGES IN ANALYSIS AND PROCESSING OF SPONTANEOUS SPEECH, MTA Nyelvtudományi Intézet, Budapest, pp. 49-74, 2018.

[3] Markó, Alexandra. 2012. “Az Irreguláris Zönge Szerepe a Magánhangzók Határának Jelölésében V(#)V Kapcsolatokban.” Beszédkutatás 2012 [Speech Research 2012], 5–29.

[4] Markó, Alexandra. 2011. “A Glottalizáció Határjelző Szerepe a Felolvasásban.” Beszédkutatás 2011 [Speech Research 2011], 31–45.

[5] Bőhm, Tamás Mihály. "Analysis and modeling of speech produced with irregular phonation." (2009).

[6] J. Laver (1980). ‘The Phonetic Description of Voice Quality’. Cambridge Univ. Press.

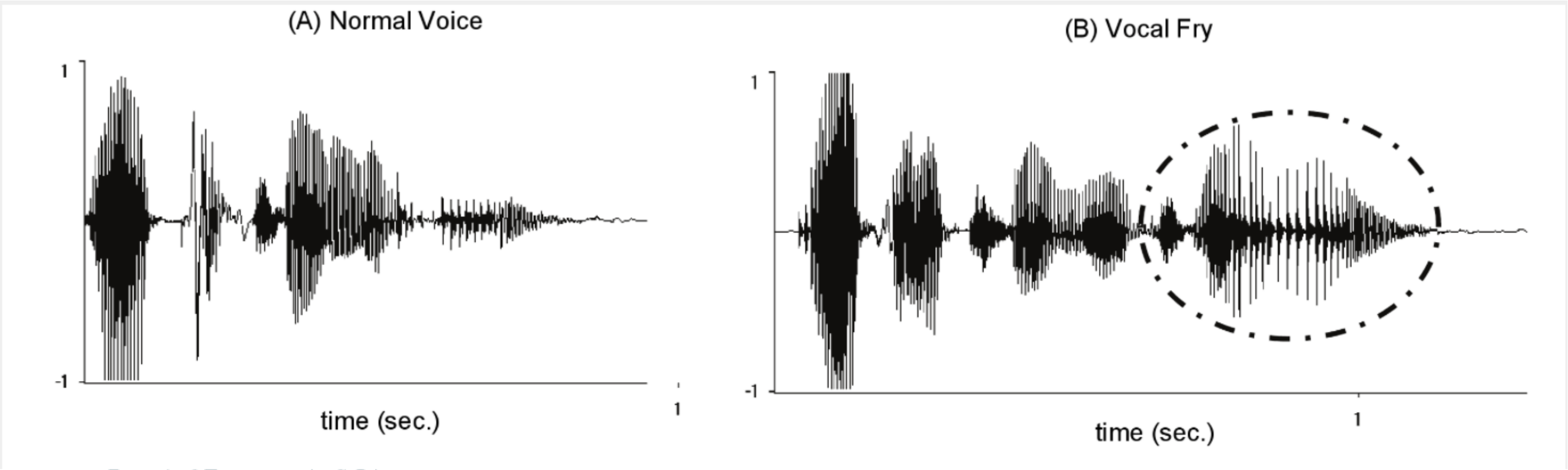
[7] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin (2012). " Habitual Use of Vocal Fry in Young Adult Female Speakers". Journal of Voice, vol. 26, no. 3, pp. e111–e116.

[8] N. P. Narendra and K. Sreenivasa Rao (2017). " Generation of creaky voice for improving the quality of HMM-based speech synthesis". Computer Speech & Language; vol. 42, pp. 38-58.

[9] J. Kane, T. Drugman, and C. Gobl (2013). ‘Improved automatic detection of creak’. Comp. Speech & Lang., vol. 27, no. 4, pp. 1028-1047.

Thank you for your listening!

Ali Raheem Mandeel
aliraheem.mandeel@edu.bme.hu


(A) Normal Voice
time (sec.)
(B) Vocal Fry
time (sec.)