Synthesis after a couple PINTs: Investigating the role of pause-internal phonetic particles in speech synthesis and perception

Mikey Elmers[1], Johannah O'Mahony[2], Éva Székely[3]

[1]*Saarland University (Germany),* [2]*University of Edinburgh (United Kingdom),* [3]*KTH Royal Institute of Technology (Sweden)*

elmers@lst.uni-saarland.de, johannah.o'mahony@ed.ac.uk, szekely@kth.se

Background

- Pause-internal phonetic particles (PINTs) include silences, inhalation and exhalation noises, filler particles "uh" and "um", and tongue clicks
- PINTs provide listener benefits for text-to-speech (TTS)[1-3]

Goals: 1) Model PINTs for synthesized speech (technological contribution)

2) Incorporate synthesized PINTs into perceptual experiment (methodological contribution)

Method

- English-language lectures from Open Yale Courses [4]
- Training data segmented into breath groups [5-6]
- ControlledPINT: transcribed PINTs
- AutoPINT: removed all PINTs transcriptions
- PyTorch implementation of Tacotron2 [7]
- Pre-trained on LJSpeech [8]
- Neural Vocoder HiFi-GAN [9]
- Participants listened to audio samples and evaluated how "certain" the speaker sounded of their opinion
- Likert: 1 (completely uncertain) to 7 (completely certain)
- Listeners heard 40 stimuli (10 sentences x 4 conditions)
- 50 native English participants from UK

condition	PINTs dur	total dur	prop
natural	15.96	41.57	38.39
ControlledPINT	13.82	40.82	33.86
AutoPINT	7.95	36.11	22.00

Fig. 1: Compared duration for five sentence excluded from training data. PINTs/total dur measured in seconds.

conditon	mean	median	mode	sd
PINTsless	5.90	6	7	1.10
long silence	4.31	4	4	1.30
filler particle	3.72	4	4	1.24
combinatory	3.51	3	4	1.27

Fig. 2: Descriptive statistics for certainty ratings by condition.

Results

- ControlledPINT performs like natural speech (**Fig. 1**)
- PINTsless condition most certain (**Fig. 2**)
- Best model:
  - *clmm(certain ~ condition + (1|id) + (1|stimuli))*
- Post-hoc pairwise comparison using Tukey method for conditions
- All comparisons significant except between filler particle and combinatory condition

Summary

- Modeled pause particles from spontaneous speech
- First system to produce discourse clicks
- Able to generate pause materials with and without labels
- PINTsless version was highest rated
- Combinatory condition was the lowest rated
- Tongue clicks can serve many functions
- Tongue clicks generated by TTS engine might behave differently than our intended function

References

[1] Dall, R., Tomalin, M., & Wester, M. 2016. Synthesising filled pauses: Representation and datamixing. In Proc. 9th ISCA Workshop on Speech Synthesis (SSW 9). pp. 7–13. **[2]** Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. 2021. Evaluating the effect of pauses on number recollection in synthesized speech. In Proc. 32nd Conference Elektronische Sprachsignalverarbeitung (ESSV '21). pp. 289–295. **[3]** Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. 2021. Take a breath: Respiratory sounds improve recollection in synthetic speech. In Proc. Interspeech 2021. pp. 3196–3200. **[4]** Hammer (2007). Open Yale courses. https://oyc.yale.edu/. License: Creative Commons BY-NC-SA. **[5]** Székely, É., Henter, G. E., & Gustafson, J. 2019. Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector. In ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6925–6929. **[6]** Székely, É., Henter, G. E., Beskow, J., & Gustafson, J. 2020. Breathing and speech planning in spontaneous speech synthesis. In ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7649–7653. **[7]** Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In ICASSP 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4779–4783. **[8]** Ito, K., & Johnson, L. 2017. The LJ speech dataset. **[9]** Kong, J., Kim, J., & Bae, J. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33 , pp. 17022–17033.