Establishing Conversational Pedagogical Agents as Credible Knowledge Providers: the Case of Synthesized Italian English

Hybrid Societies Conference 2023

Sven Albrecht Rewa Tamboli Stefan Taubert Felicia Meusel Maximilian Eibl Günter Daniel Rey Josef Schmied

TU Chemnitz

15–17 March 2023

HYBRID SOCIETIES
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Objectives

Mission

In hybrid societies, humans and embodied digital technologies should interact as seamlessly as humans among each other.

RQ1 Which specific non-native linguistic cues of CPAs influence the learning performance of non-native human learners?

RQ2 Which specific non-native linguistic cues influence attributed credibility and acceptance of CPAs by non-native human learners?

RQ3 How much does a linguistically credible CPA influence the learning performance in non-native educational contexts?

TTS System

Our TTS pipeline is based on

- Tacotron 2 (Shen et al., 2018) and
- WaveGlow (Prenger, Valle, & Catanzaro, 2019)

We extended the pipeline to be able to learn

- phoneme stress
- speaker identity
- duration of pauses

We developed some tools for working with pronunciation dictionaries and TextGrids (Praat files) and published them to PyPI[1].

[1] `https://pypi.org/user/stefantaubert`

TTS System - Training

Tacotron

- pre-training: LJ Speech dataset
  - training set: 13000 utterances (~23.7h)
  - validation set: 100 utterances (~11min)
- transfer-learning: internal Italian English dataset
  - training set: 467 utterances (~25min)
  - validation set: 2 utterances (~27s)
- parameters:
  - 500 epochs
  - batchsize: 64
  - learning rate: 0.001
  - separate learning of: phonemes, stress, speaker

For WaveGlow we used a public available pre-trained model (Nvidia, 2019).

TTS System - Synthesis

- Denoising using Facebook denoiser (Défossez, Synnaeve, & Adi, 2020)
- Normalizing audio level using FFmpeg
- Transfer-test: texts about plate tectonics and tsunamis (2–4min)
- MOS test: 5 sentences (2–9s)
- Linguistic test: 15 sentences (2–4s)
  - used Griffin-Lim vocoder (Griffin & Lim, 1984; Vogel, 2017) to create robotic voices

Linguistic Stimuli

- stimuli from authentic example of an Italian politician speaking English
- features of Italian English added in IPA
- variation along the human-robot voice scale

(1) a. Everybody love Italy, but not Brexit.
b. ˈɛvribədi ləv ˈɪdəli, bət nɑt brˈɛgzət.
c. ˈɛ v r i b ə d i l ə v ˈɪ d ə l i , SIL1 b ə t n ɑ t b r ˈɛ g z ə t . SIL2

(6) a. But now the istory of Brexit is over, Britain is hout of the EU.
b. bət naʊ ðə ˈɪstəri əv brˈɛgzət iz ˈoʊvər, brˈɪtn iz haʊt əv ði ijˈu.
c. b ə t n a ʊ ð ə ˈɪ s t ə r i ə v b r ˈɛ g z ə t i z ˈo ʊ v ə r, SIL1 b r ˈɪ t n i z h a ʊ t ə v ð i i j ˈu . SIL2

Learning Survey

- survey tested credibility, knowledge recall and knowledge transfer using synthesized audios of two factual texts
- credibility instruments were two 5-point Likert scales measuring source credibility and message credibility respectively
- text order was randomized to eliminate order effects
- study comprised of 10 retention questions, 8 transfer questions and the credibility scale

Results: Perception

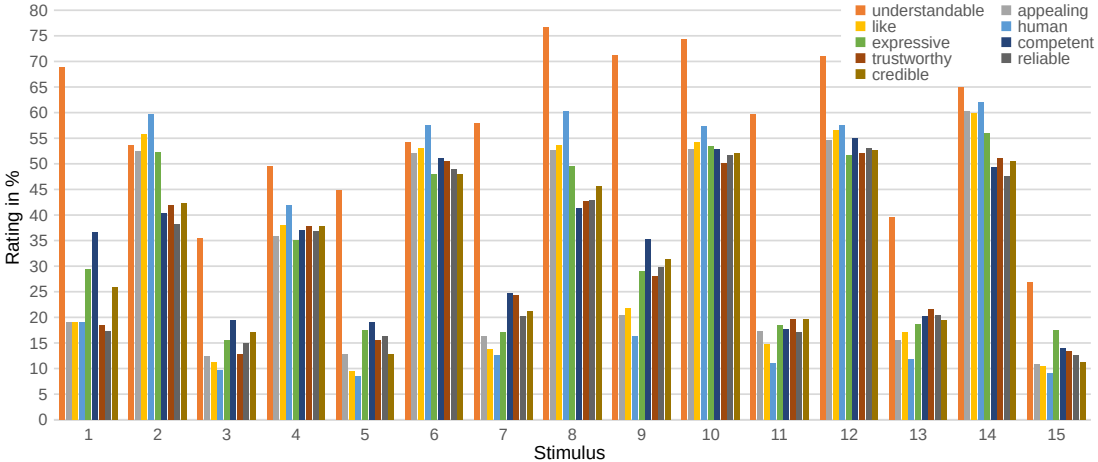

Figure 1: ratings of linguistic stimuli

Results: Learning

- participants rated the standard American variety as significantly more credible
- the American variety was also rated significantly higher in Trustworthiness, Expertise and Attractiveness
- the information of both learning texts was found to be credible, without significant difference in credibility ratings

	M_A	SD_A	M_I	SD_I	$t(8)$	p
Trustworthiness	3.82	0.61	3.27	0.67	2.83	0.02
Expertise	3.91	0.53	3.3	0.69	3.36	0.01
Attractiveness	3.59	0.66	3.04	0.72	2.86	0.02

Table 1: Results of statistical analysis for American English

Discussion

Limitations of our approach:

- audio playback conditions was less than ideal
- low number of participants
- too many variables in the linguistic survey

Conclusion

- TTS pipeline and measurement instruments are valid
- a standard variety is perceived as more credible than a non-native variety
- learning performance equally good for native and non-native variety
- more data needed to further validate psychological instruments (for other varieties)

References

Défossez, A., Synnaeve, G., & Adi, Y. (2020, October). Real Time Speech Enhancement in the Waveform Domain. In Interspeech 2020 (pp. 3291–3295). ISCA. Retrieved 2022-06-13, from https://www.isca-speech.org/archive/interspeech_2020/defossez20_interspeech.html doi: 10.21437/Interspeech.2020-2409

Griffin, D., & Lim, J. (1984, April). Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2), 236–243. doi: 10.1109/TASSP.1984.1164317

Nvidia. (2019). WaveGlow LJS 256 channels. https://catalog.ngc.nvidia.com/orgs/nvidia/models/waveglow_ljs_256channels.

Prenger, R., Valle, R., & Catanzaro, B. (2019, May). WaveGlow: A Flow-based Generative Network for Speech Synthesis. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3617–3621). Brighton, United Kingdom: IEEE. Retrieved 2022-06-13, from https://ieeexplore.ieee.org/document/8683143/ doi: 10.1109/ICASSP.2019.8683143

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., … Wu, Y. (2018, April). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779–4783). doi: 10.1109/ICASSP.2018.8461368

Vogel, B. K. (2017). griffin_lim. https://github.com/bkvogel/griffin_lim.

Establishing Conversational Pedagogical Agents as Credible Knowledge Providers: the Case of Synthesized Italian English

Hybrid Societies Conference 2023

Sven Albrecht Rewa Tamboli Stefan Taubert Felicia Meusel Maximilian Eibl Günter Daniel Rey Josef Schmied

TU Chemnitz

15–17 March 2023

HYBRID SOCIETIES
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation