Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents

Positive impacts of language technology: TISLID 22

Sven Albrecht
sven.albrecht@phil.tu-chemnitz.de

TU Chemnitz

28.05.2022

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?

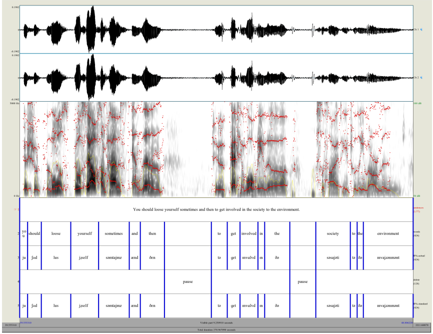
Pilot Study

Sociolinguistic interview with Chinese PhD student at TUC

- reading passages
- wordlist
- interview questions based on Tagliamonte (2006, Appendix B) supplemented with target group specific questions

Transcription of data

- orthographic (sentences, words)
- IPA (phonemes)



1	You should loose yourself sometimes and then to get involved in the society to the environment.																	sentences (1/77)
2	you	should	loose	yourself	sometimes	and	then		to	get	involved	in	the		society	to	the	environment	words (424)
3	ju	ʃʊd	lus	jɹsɛlf	səmtajmz	ænd	ðen		tə	get	ɪnvɑlvd	ɪn	ðə		səsajəti	tə	ðə	ɪnvajɹənmənt	IPA-actual (424)
4								pause						pause					delete (128)
5	ju	ʃʊd	lus	jɹsɛlf	səmtajmz	ænd	ðen		tə	get	ɪnvɑlvd	ɪn	ðə		səsajəti	tə	ðə	ɪnvajɹənmənt	IPA-standard (424)

39.555310 · Visible part 9.250910 seconds · 48.806220 · 222.160870
Total duration 270.967090 seconds

Figure 1: Transcription Screenshot

TTS System

Goal

A TTS synthesis system that can synthesize English text in different Chinese accents.

In the synthesized speech we want to control the following features:

- morphosyntactic cues e.g. syntax, grammar
- phonetic cues e.g. pronunciation of phonemes
- prosodic cues e.g. stress, intonation

Our TTS system is based on two different models and uses transfer learning.

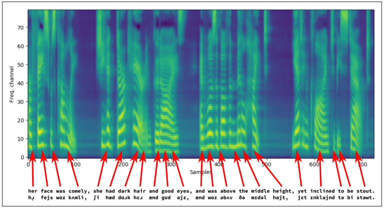


Figure 2: Example output mel-spectrogram

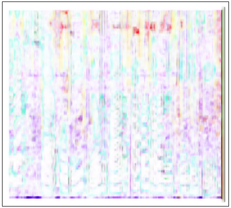


Figure 3: Difference original and synthesized audio

TTS System

Currently we are able to control:

- morphosyntactic cues with a rule based approach
- phonetic cues with a phone-based TTS (based on Tacotron2 by Shen et al., 2018)

We developed some helpful tools for speech synthesis:

- for recordings: e.g. resampling, automatic detection of silence
- for text: e.g. G2P conversion, symbol mapping

Audio Examples

https://stefantaubert.github.io/tacotron2/

Preparation of Data Collection

Challenge

Corpus linguistic literature (e.g. Love et al., 2017) and computer science literature (e.g. Bozkurt et al., 2003) propose different corpus creation criteria.

Sociolinguistic methodology has to be adapted to TTS application:

- previous TTS systems trained on reading passages
- selection of reading passages by selecting sentences according to phone and diphone coverage using greedy selection (Taubert et al., fc.)
- supplementary word list (i.e. missing phonemes/diphones)
- interview questions
    - many questions from Tagliamonte (2006) not relevant
    - add questions about high school and university life

Data Collection and Covid-19

Challenge

International travel impossible, Chinese visa application suspended indefinitely, yet data collection through field work is crucial for the project's success.

Alternative forms of data collection:

- send recording equipment to Chinese partner university and conduct online/hybrid interviews
- equipment stuck in customs for months
- include other varieties, in particular Nigerian English
- leverage social network of visiting scholars from Nigeria (in-group interviews)

Collecting data for TTS is relatively easy, collecting data for TTS and linguistic analysis is challenging.

Measuring Vowel Spaces

Quantification workflow (WIP)

- Forced alignment using the Montreal Forced Aligner (McAuliffe et al., 2017)
- Automated vowel formant measurements in Praat
- Vowel plots generated in R
- Hampel filtering of outliers (Hampel, 1974)
- speaker intrinsic, vowel extrinsic, formant intrinsic normalization (Lobanov, 1971)

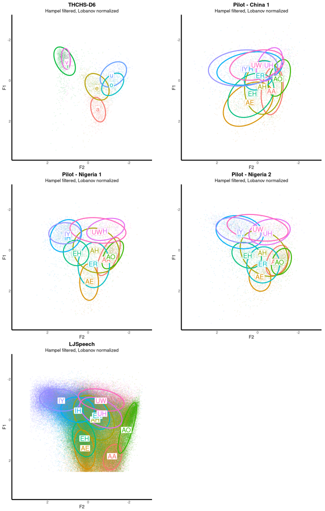


Figure 4: Vowel space plot of pilot test subjects

Measuring Vowel Spaces

Quantification workflow (WIP)

- Forced alignment using the Montreal Forced Aligner (McAuliffe et al., 2017)
- Automated vowel formant measurements in Praat
- Vowel plots generated in R
- Hampel filtering of outliers (Hampel, 1974)
- speaker intrinsic, vowel extrinsic, formant intrinsic normalization (Lobanov, 1971)

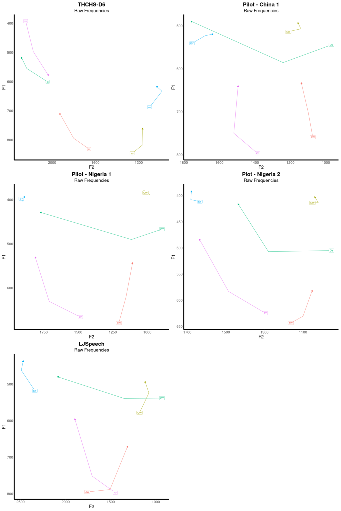


Figure 5: Vowel space plot of pilot test subjects

Vowel Formant TTS Quality Metric

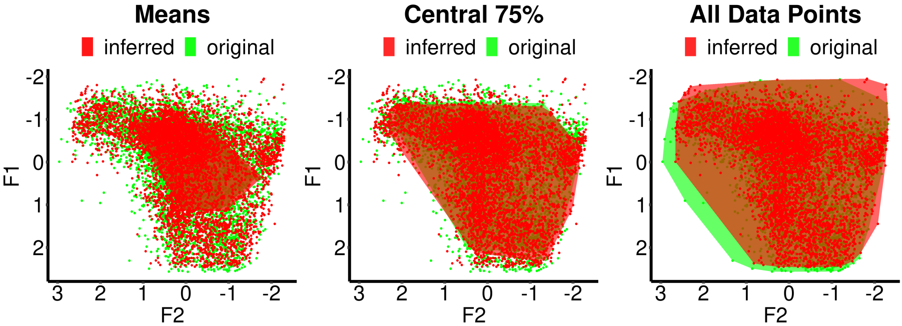


Figure 6: LJ Speech Vowel Space, plotted as F1 - F2 space (Lobanov Normalized, Hampel Filtered)

Vowel Formant TTS Quality Metric

Table 1: Vowel Space Overlap

Dataset	Overlap
Phoneme Averages	93.20%
Central 75%	97.70%
All Data Points	91.50%

Vowel Formant TTS Quality Metric

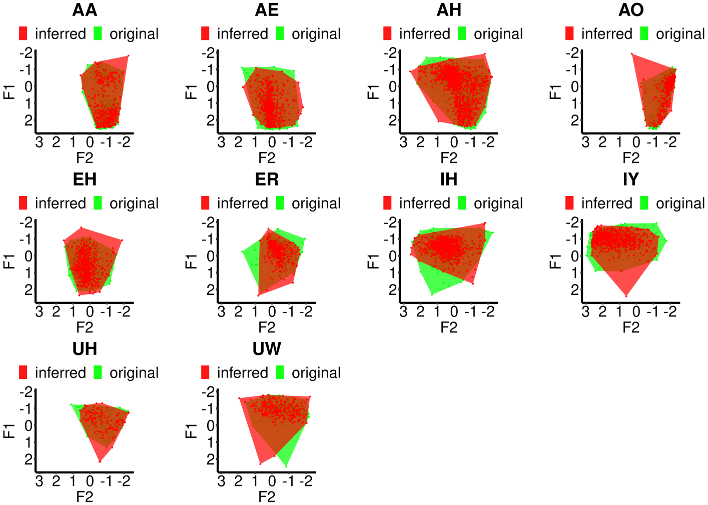


Figure 7: LJ Speech Phoneme Space All Data Points (Lobanov Normalized, Hampel Filtered)

Vowel Formant TTS Quality Metric

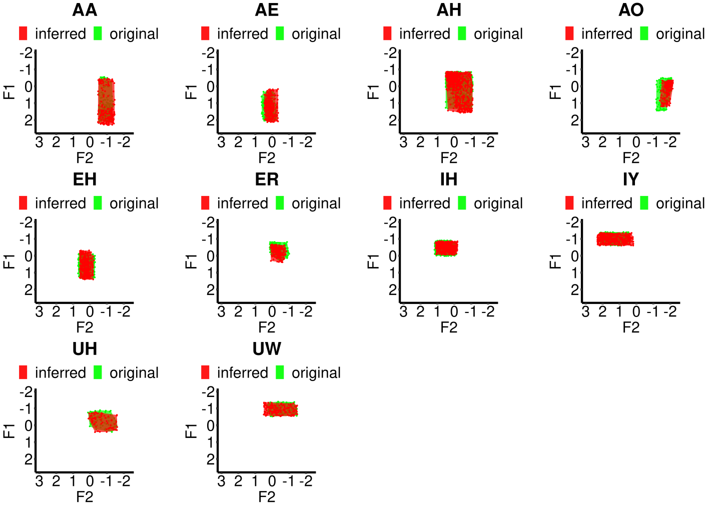


Figure 8: LJ Speech Phoneme Space Central 75% of All Data Points (Lobanov Normalized, Hampel Filtered)

Vowel Formant TTS Quality Metric

Table 2: Phoneme Space Overlap

Phoneme	Central 75%	All Data Points
AA	98.22%	95.84%
AE	81.07%	86.69%
AH	95.22%	93.24%
AO	56.54%	98.03%
EH	83.24%	92.04%
ER	67.60%	72.20%
IH	84.93%	74.91%
IY	93.61%	81.24%
UH	86.52%	91.41%
UW	92.19%	87.44%

Measuring Linguistic Credibility

Assumptions:

- some linguistic features contribute more to perceived credibility than others
- the NN TTS systems allows fine grained control of linguistic features
- higher linguistic credibility leads to better learning outcomes (c.f. Rey & Steib, 2013)

Learning outcomes will be measured in participant studies (post Covid):

- prior knowledge, interest (could be confounding variables)
- retention, i.e. remembering facts
- transfer, i.e. applying knowledge to new tasks

Note

The stimuli will be based on factual knowledge about a subject, no language learning.

Measuring Linguistic Credibility

Online Questionnaire

- brief listening examples
- rating scales (0-100)
- variation of features
- variation human-robot
- Cronbach’s alpha = 0.98

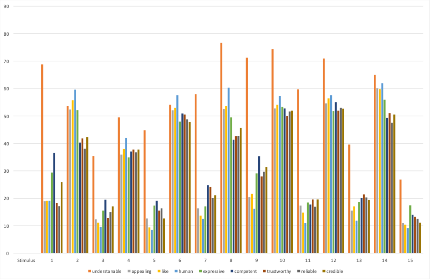


Figure 9: Average Scores of Participants in Pilot Study

Next Steps

Completing work packages:

- running TTS system with very limited input data
- adapting our linguistic TTS quality metric to include diphthongs
- data collection in Chemnitz and China
- testing the significance of intercultural factors influencing credibility and learning
- refining our credibility and learning performance measurement instruments

Planned publications:

- Linguistic Credibility as a Key Issue in the Communication of Humans and Humanoids in a Hybrid Society
- Methodological Considerations - Comparison of online/distance vs. fieldwork/interviews
- The Chinese English Vowel Space - A Big Data Approach

References

Bozkurt, B., Ozturk, O., & Dutoit, T. (2003). Text design for TTS speech corpus building using a modified greedy selection. Eighth European Conference on Speech Communication and Technology.

Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association, 69(346), 383–393. https://doi.org/10.1080/01621459.1974.10482962

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. The Journal of the Acoustical Society of America, 49(2B), 606–608. https://doi.org/10.1121/1.1912396

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi.. Interspeech, 2017, 498–502.

Rey, G. D., & Steib, N. (2013). The personalization effect in multimedia learning: The influence of dialect. Computers in Human Behavior, 29(5), 2022–2028.

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. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Tagliamonte, S. A. (2006). Analysing sociolinguistic variation. Cambridge University Press.

Taubert, S., Sternkopf, J., Kahl, S., & Eibl, M. (fc.). A Comparison of Text Selection Algorithms for Sequence-to-Sequence Neural TTS. NeurIPS 2021.

Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents

Positive impacts of language technology: TISLID 22

Sven Albrecht
sven.albrecht@phil.tu-chemnitz.de

TU Chemnitz

28.05.2022

HYBRID SOCIETIES

Funded by DFG Deutsche Forschungsgemeinschaft
German Research Foundation