Linguistic Perspectives on Building a Neural TTS System for Teaching and Learning Settings

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

TU Chemnitz

16.05.2023

HYBRID SOCIETIES
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

About me

- BA & MA in English and American Studies from TU Chemnitz
- worked as vocational school teacher in Germany
- worked as high school teacher in China
- worked on the Erasmus+ project **TEFL-ePAL**
- currently working at TUC as part of the DFG funded CRC **Hybrid Societies**

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

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.

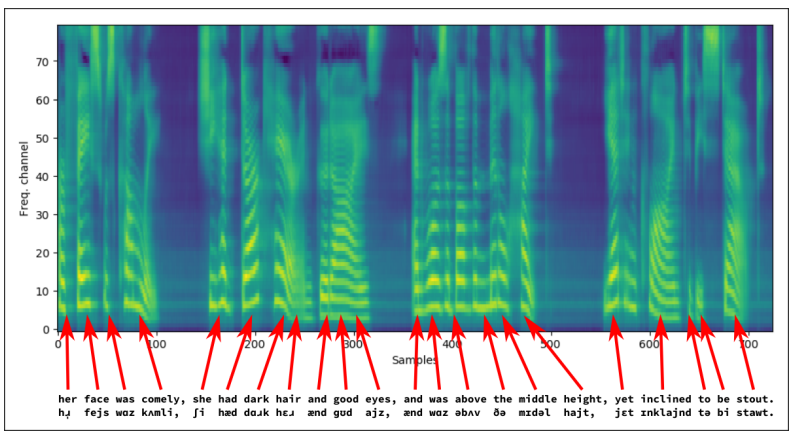

Figure 1: Example output mel-spectrogram

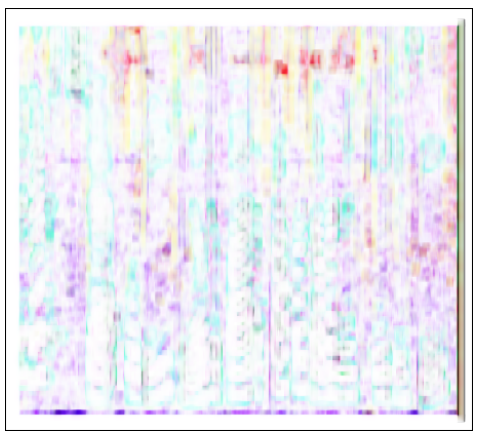

Figure 2: Difference original and synthesized audio

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.)
- interview questions
  - many questions from Tagliamonte (2006) not relevant
  - additional 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, broadening our approach:

- added Nigerian English, leveraging social network of visiting scholars from Nigeria (in-group interviews)
- added Italian English, leveraging personal connections
- added Czech English, via Marina's project
- recorded Chinese students in Chemnitz
- Chinese collegues provided all 30 planned recordings

Features of Chinese English

Definition

Chinese English is a developing variety of English, which is subject to ongoing codification and normalization processes. It is based largely on the two major varieties of English, namely British and American English. It is characterized by the transfer of Chinese linguistic and cultural norms in discourse, syntax, pragmatics, lexis, and phonology. (Albrecht, 2021)

Figure 3: 'China linguistic map' U.S. Central Intelligence Agency, marked as public domain, Wikimedia

Features of Chinese English

- discourse: e.g. ”ancestral hometown discourse” or discourse of ’face’ (Xu, 2020)
- syntax: three types of variation (Xu, 2008)
  - preference: use of imperatives
  - innovation: unmarked usage of nominalization
  - transference: null-subject/object utterances
- pragmatics: general-particular pattern (*theme* and *rheme*, see Halliday & Matthiessen, 2013) vs. problem-solution pattern (Wang & Li, 1993)
- lexis
  - Chinese loanwords in English: *dimsum, guanxi* (Xu, 2020)
  - Chinese nativized words: *propaganda, cadre, comrade, individualism* (Chang, 2001; Xu, 2020)
  - common English words used differently: *all-around* (Liang & Li, 2017)
- phonology: vowels and consonants

Measuring Vowel Spaces

Quantification workflow

- 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)

Number of data points:

- American English: 229,100
- Chinese English: 37,089

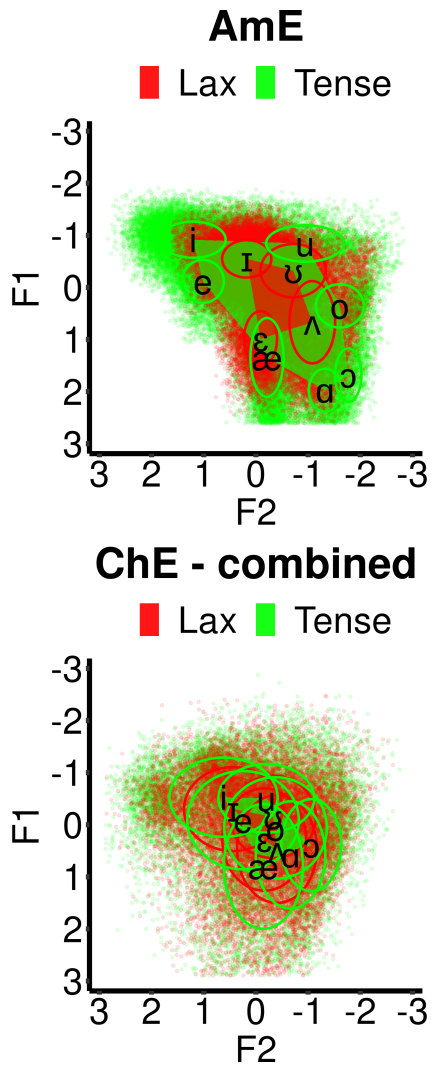

Figure 4: Vowel space plot of pilot test subjects

Measuring Vowel Spaces

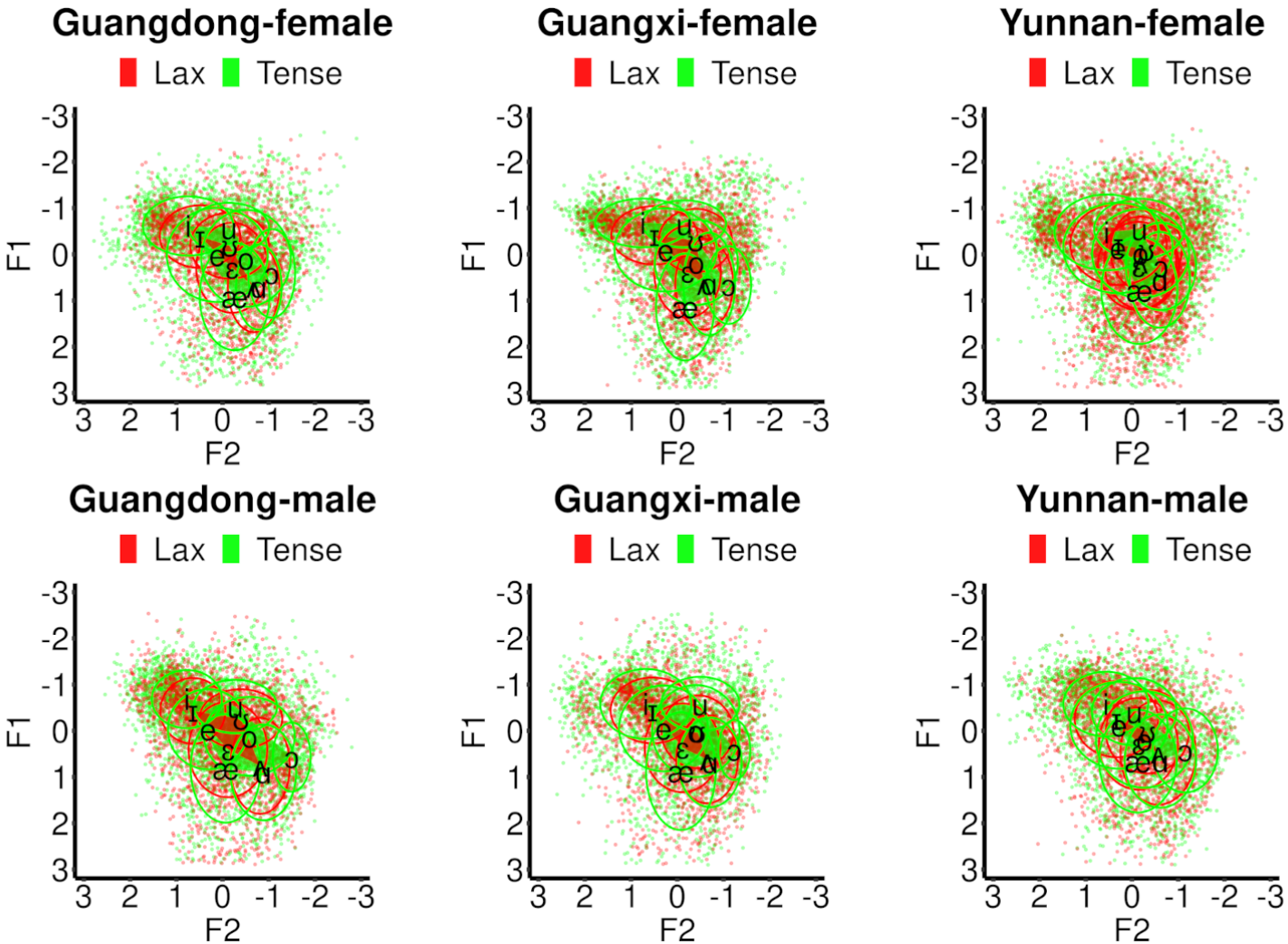

Figure 5: Vowel spaces of Chinese English subnational varieties

Measuring Vowel Spaces

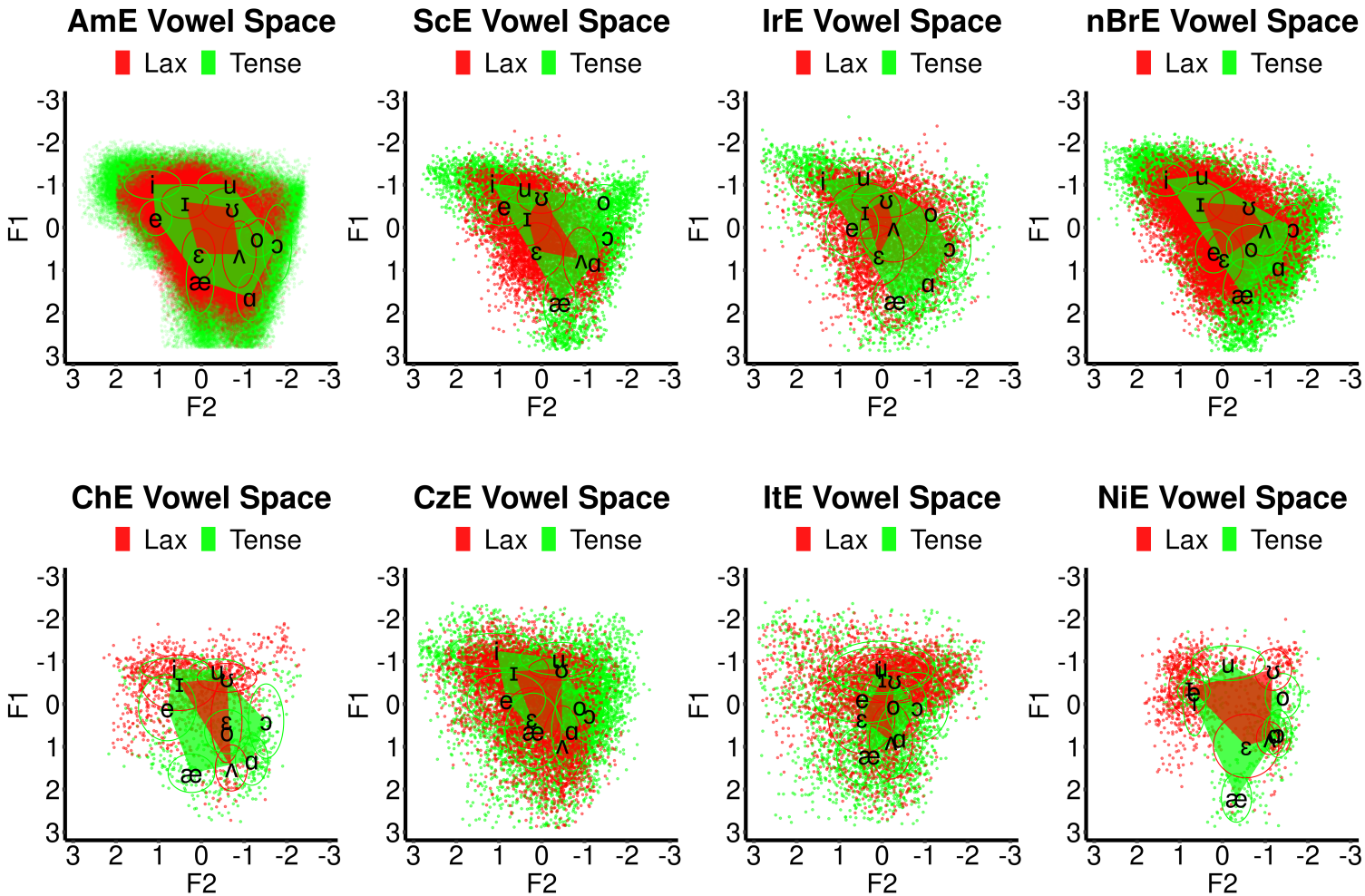

Figure 6: Vowel spaces of all varieties under investigation

Measuring Vowel Length

Method:

- best done in Praat (same script as formant measurements)
- analysis and plotting in R
- speech rate normalization requires a separate Praat script

Methodological Considerations:

- raw values
- speech rate normalization
- Lobanov normalization

Measuring Vowel Length

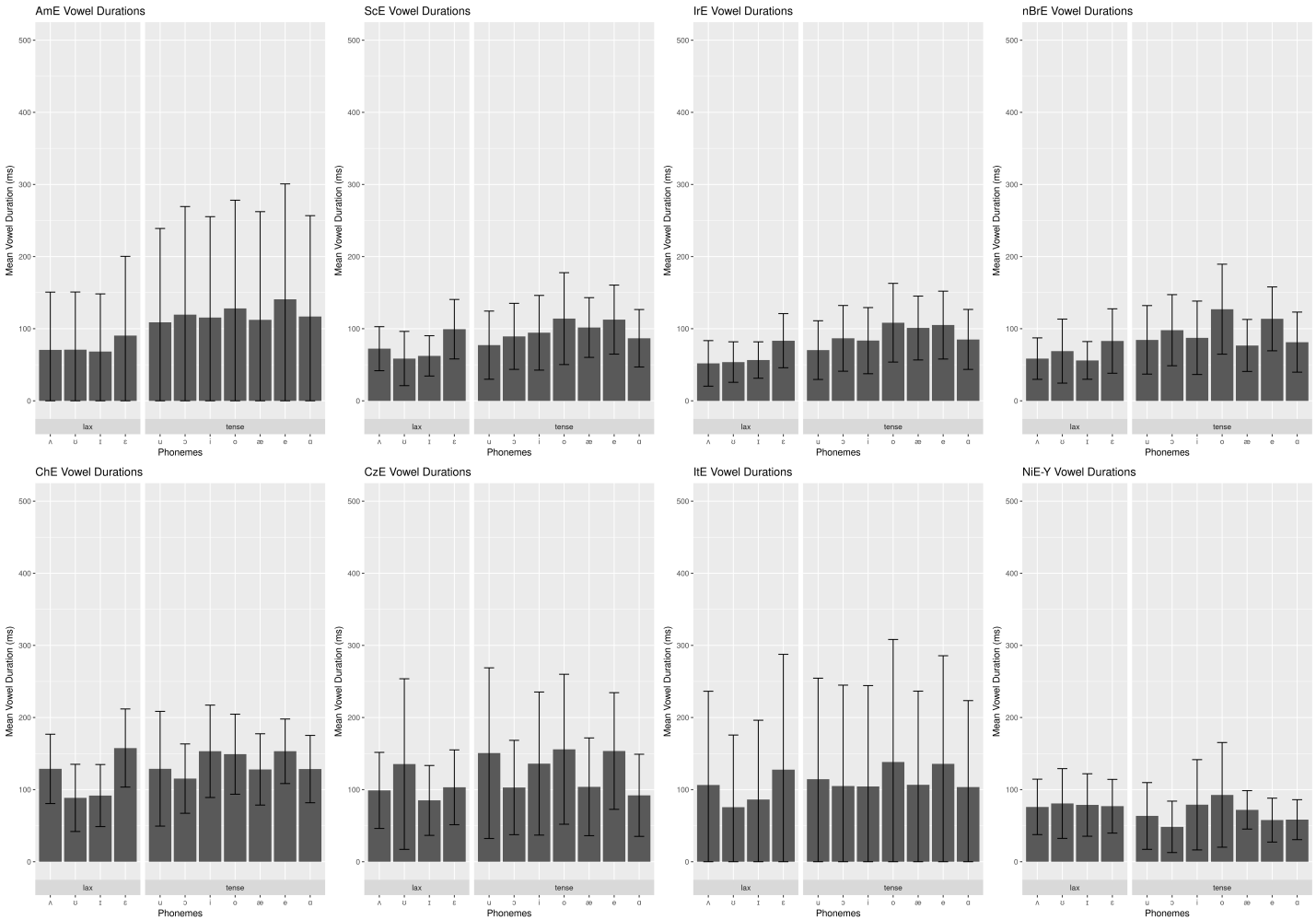

Figure 7: Raw vowel duration of all varieties under investigation

Measuring Vowel Length

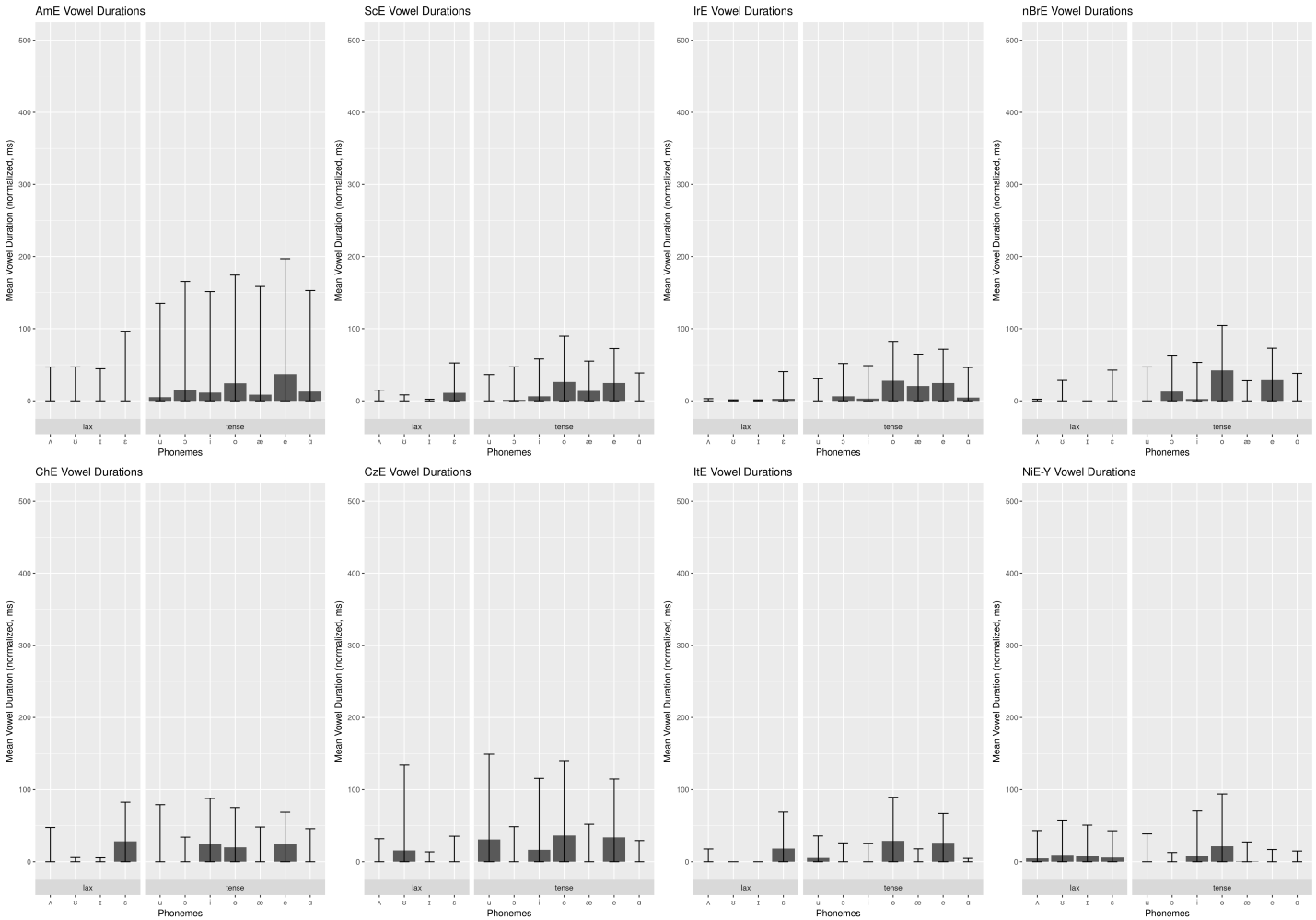

Figure 8: Speech rate normalized vowel duration of all varieties under investigation

Measuring Vowel Length

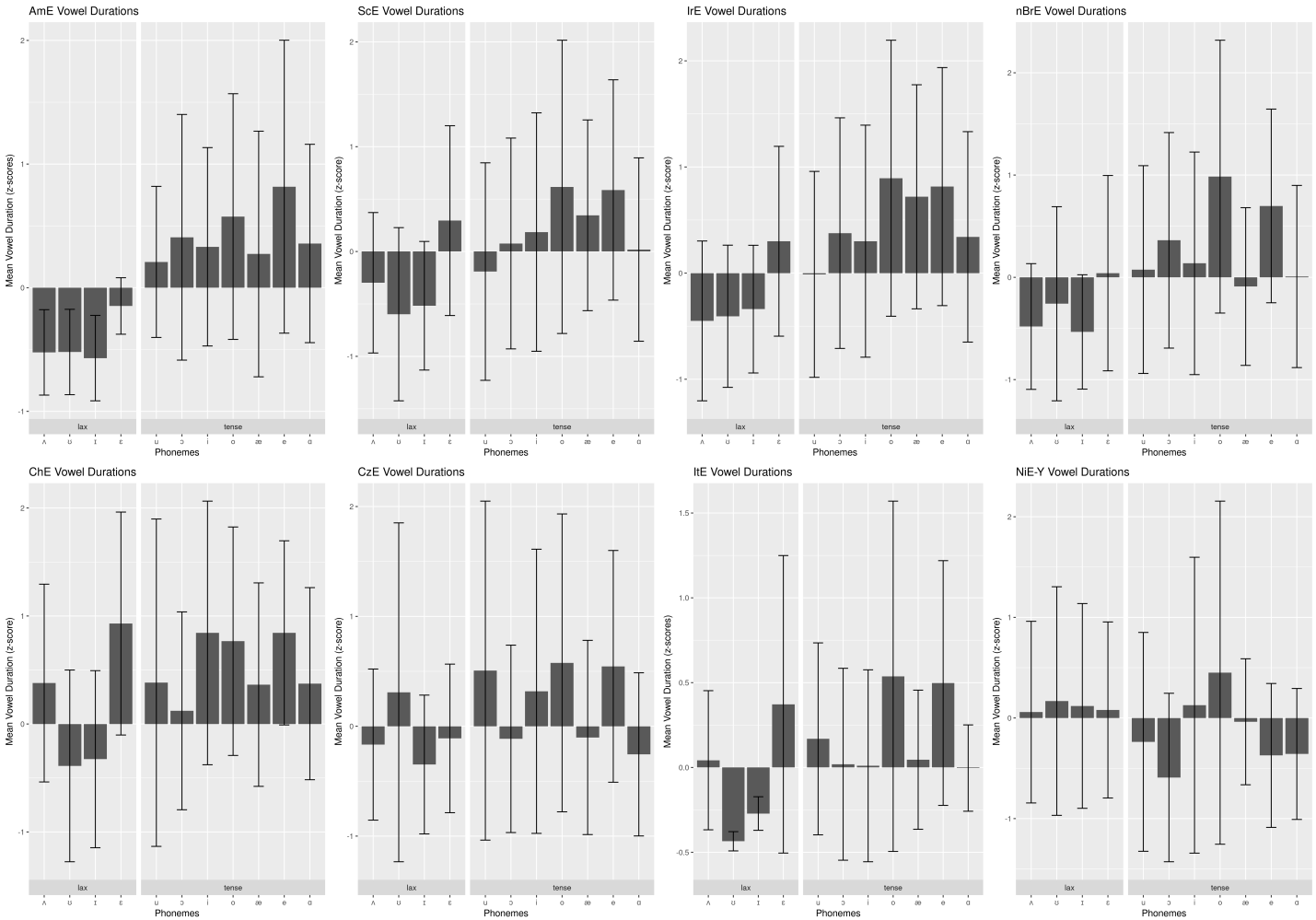

Figure 9: Lobanov normalized vowel duration of all varieties under investigation

Vowels Conclusion

- inner circle varieties have clear tense-lax contrast, but European varieties also have variation
- expanding circle varieties mostly have no clear tense-lax contrast
- tense-lax contrast not clearly visible in vowel length data across all three circles
- Chinese English data shows subnational variation
- automatic vowel formant and duration measurement and analysis methods are valid and reliable

Vowel Formant TTS Quality Metric

- most TTS systems have some form quality metric built in
- tacotron2: visual comparison of mel spectrograms

Idea

Can we use our vowel space measurement pipline to establish a linguistically motivated TTS quality metric?

- two data sets:
  - original: recordings used to train TTS system
  - inferred: speech generated by the TTS system
- plot vowel space for both
- draw a convex polygon around means, central 75%, and all data points
- calculate overlap of the polygons

Published at CIVEMSA 2022 (Albrecht et al., 2022) and received Best Paper Award 2nd Place

Vowel Formant TTS Quality Metric

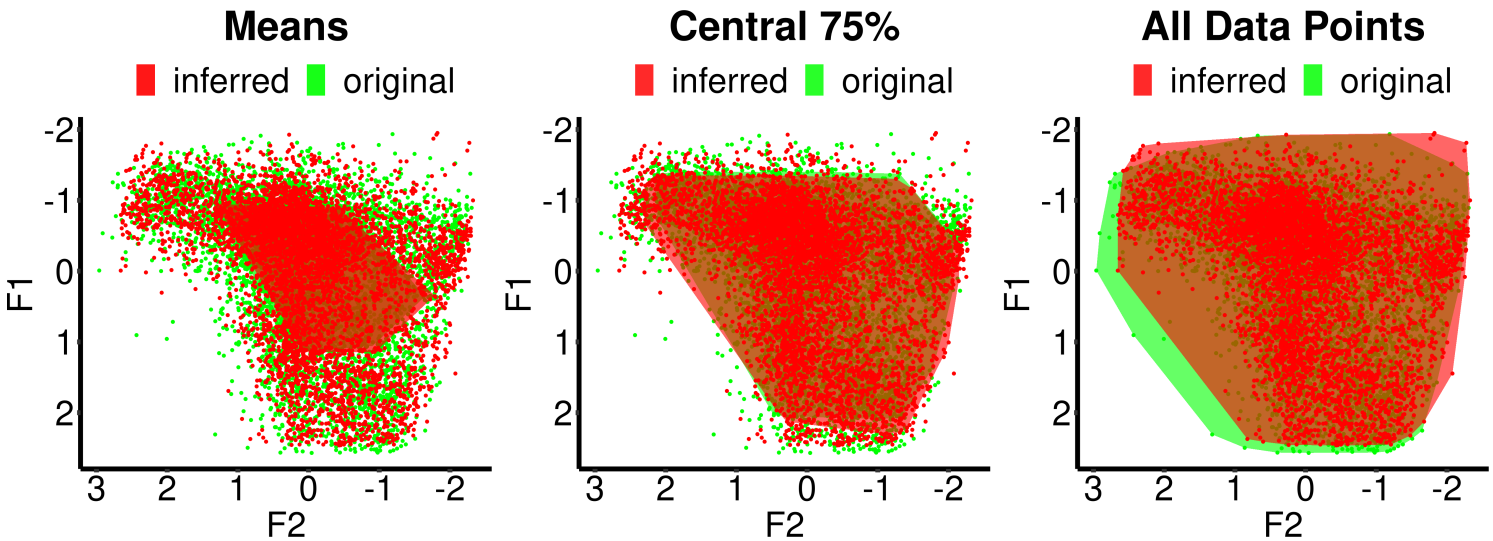

Figure 10: LJ Speech Vowel Space, plotted as F1 - F2 space (Lobanov Normalized, Hampel Filtered)

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

Table 1: Vowel Space Overlap

Vowel Formant TTS Quality Metric

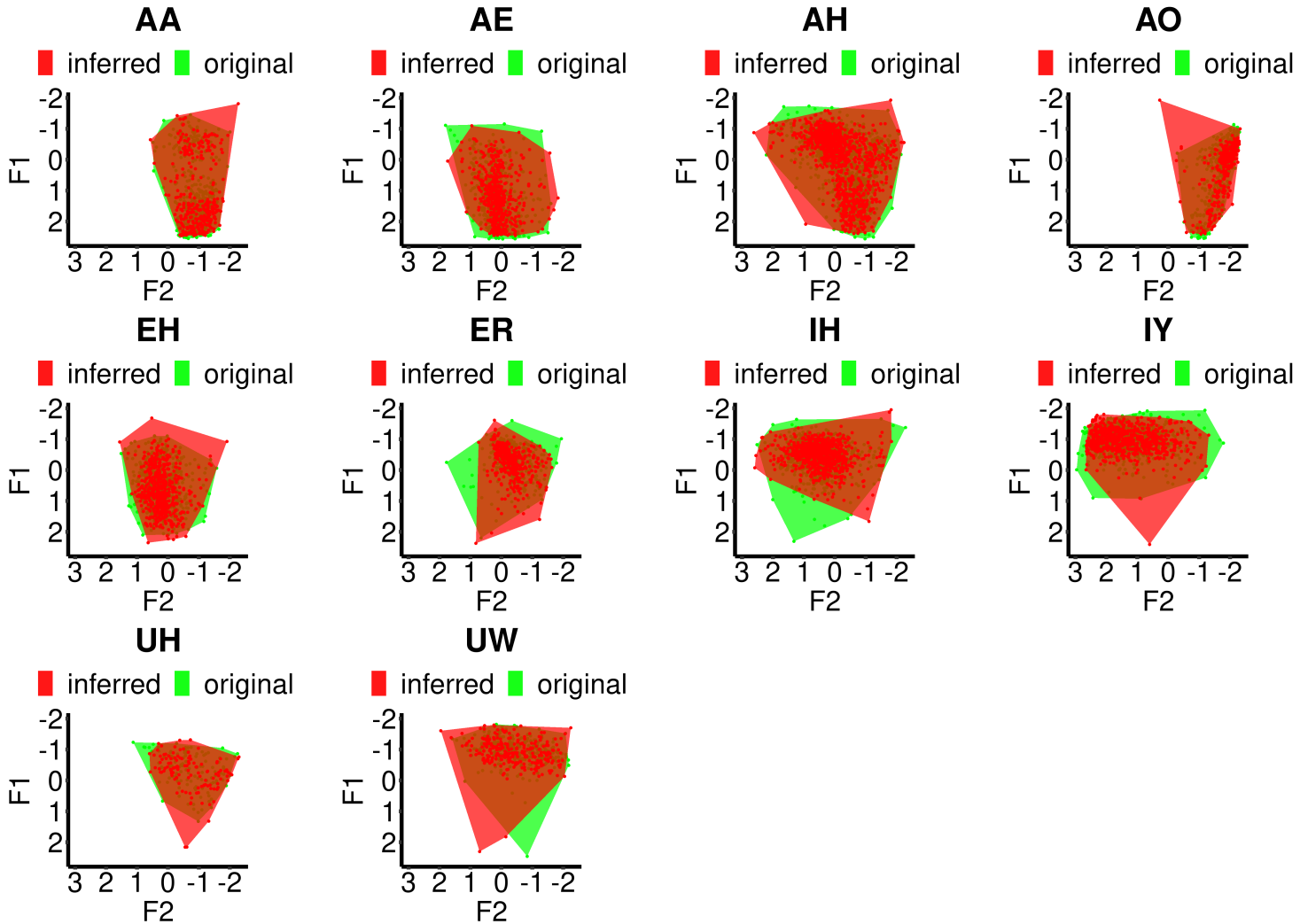

Figure 11: LJ Speech Phoneme Space All Data Points (Lobanov Normalized, Hampel Filtered)

Vowel Formant TTS Quality Metric

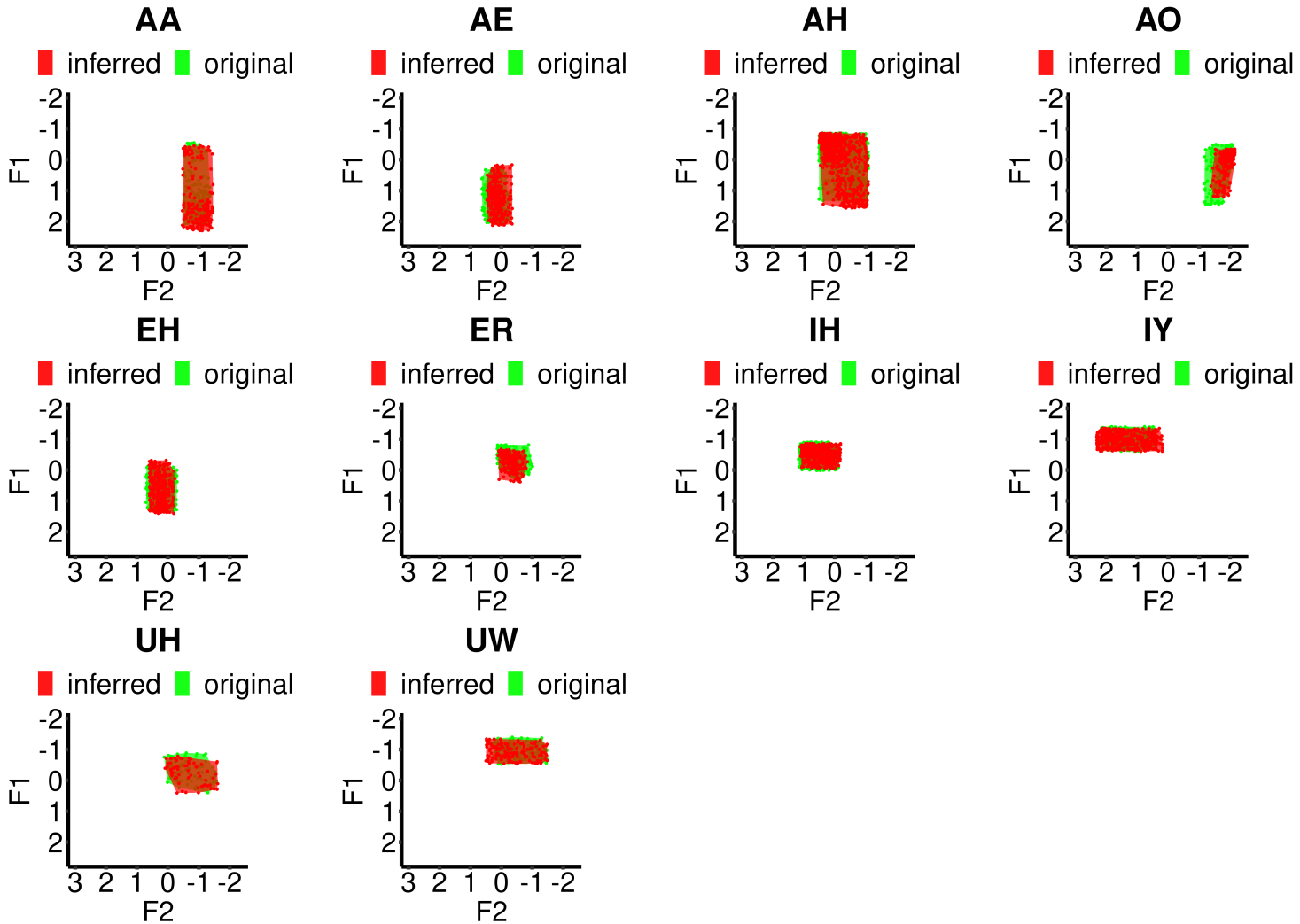

Figure 12: LJ Speech Phoneme Space Central 75% of All Data Points (Lobanov Normalized, Hampel Filtered)

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%

Speech Rhythm Theory

- conventionally, languages have been categorized as syllable or stress timed (Pike, 1945)
- the third category of mora timed languages has been abandond (Fuchs, 2023)
- categorization based on equal length of units could not be supported empirically (Dauer, 1983)
- speech rhythm is a gradable phenomenon (White et al., 2012)
- speech rhythm metrics, calculated from speech segmented into vocalic and consonantal intervals
- adjacent intervals of the same type merged together
- interval measures
  - ΔV, ΔC, %V (Ramus et al., 1999)
  - VarcoV, VarcoC (Dellwo, 2006)
- pairwise variability indices
  - nPVI-V (Low et al., 2000)
    $nPVI = 100 \times \left(\sum_{k=1}^{m-1} \left|(d_k - d_{k+1}) \,/\, ((d_k + d_{k+1}) \,/2)\right|\right) \,/\, (m-1)$
  - rPVI-C (Grabe & Low, 2002) $rPVI = \left(\sum_{k=1}^{m-1} \left|d_k - d_{k+1}\right|\right) \,/\, (m-1)$

Speech Rhythm Application

- data of English varieties from all three circles (cf. Kachru, 1992)
  - inner circle: American, (Northern) British, Scottish, Irish English
  - outer circle: Nigerian English (Yoruba and Igbo)
  - expanding circle: Czech, Italian, Chinese (Guangdong, Guangxi, Yunnan) English
- two types of data:
  - reading data: linguistic reading passages (see Deterding, 2006, for a discussion of different reading passages) and TTS training data script (Taubert et al., fc.)
  - interview data: questions based on the list provided by Tagliamonte (2006)
- automated analysis pipeline:
  - forced alignment using the MFA (McAuliffe et al., 2017)
  - transformation of data and measurements in Praat
  - analysis and plotting in R

Speech Rhythm Results

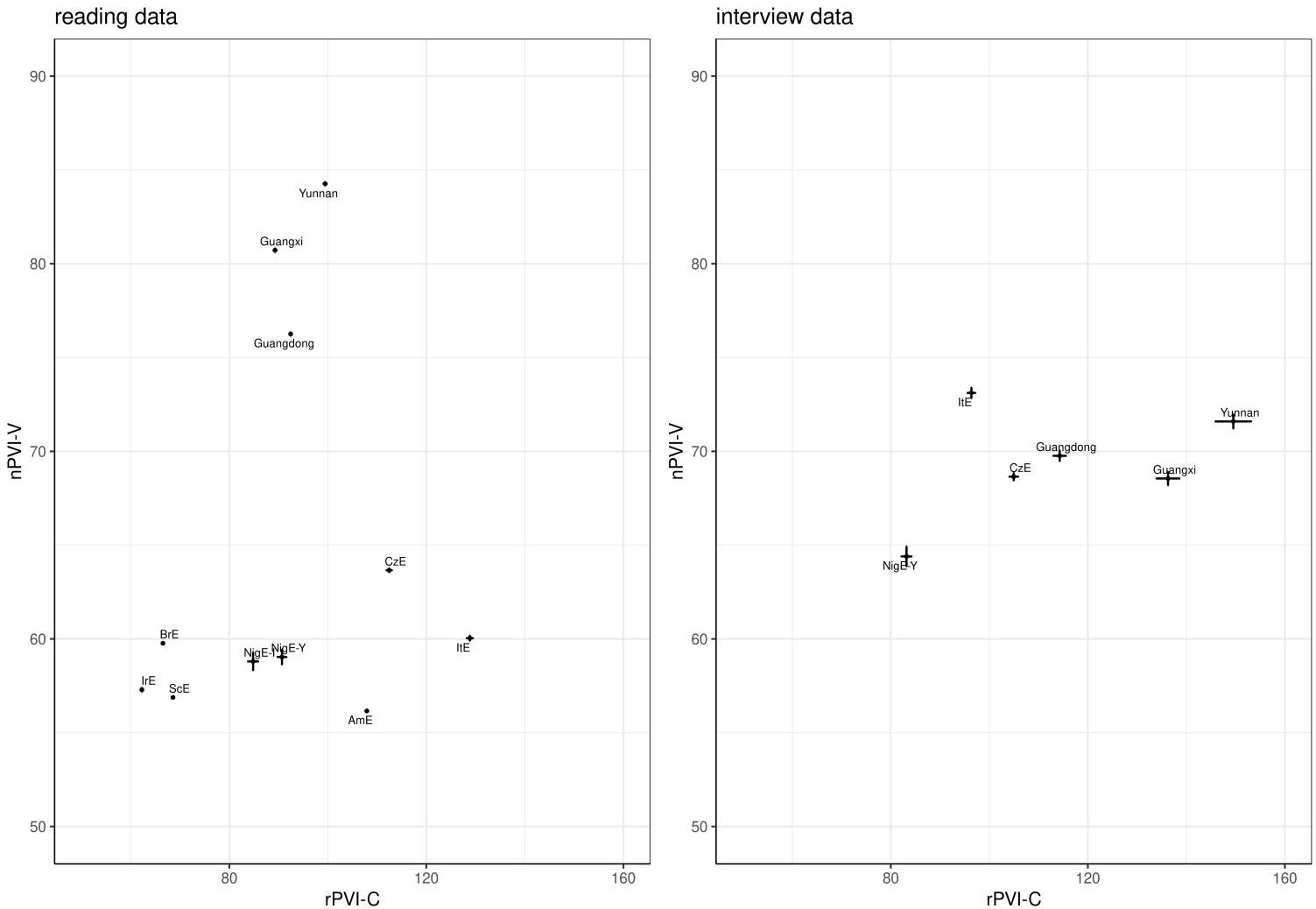

Figure 13: nPVI-V - rPVI-C Chart for reading and interview data

Speech Rhythm Results

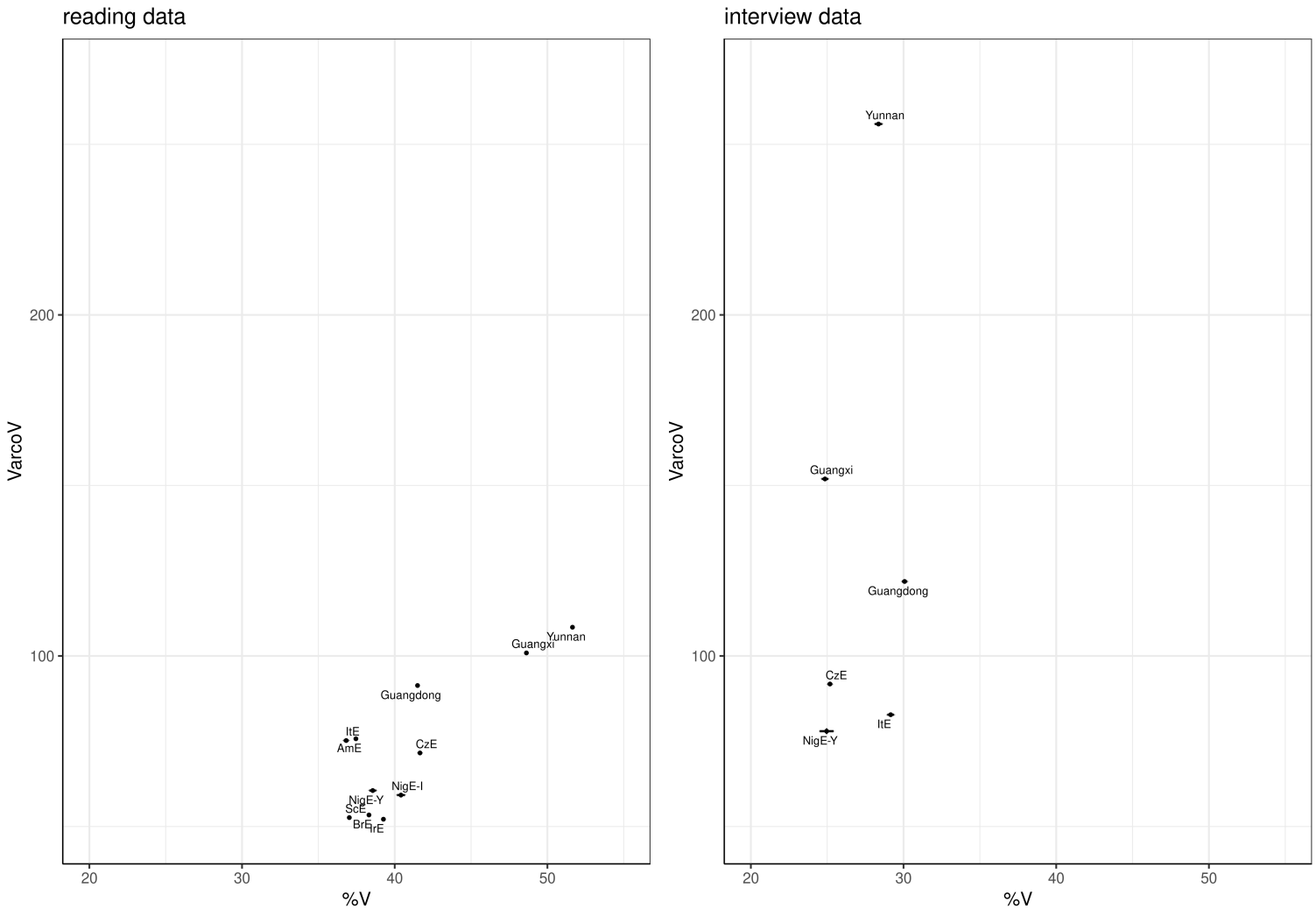

Figure 14: VarcoV - %V Chart for reading and interview data

Speech Rhythm Results

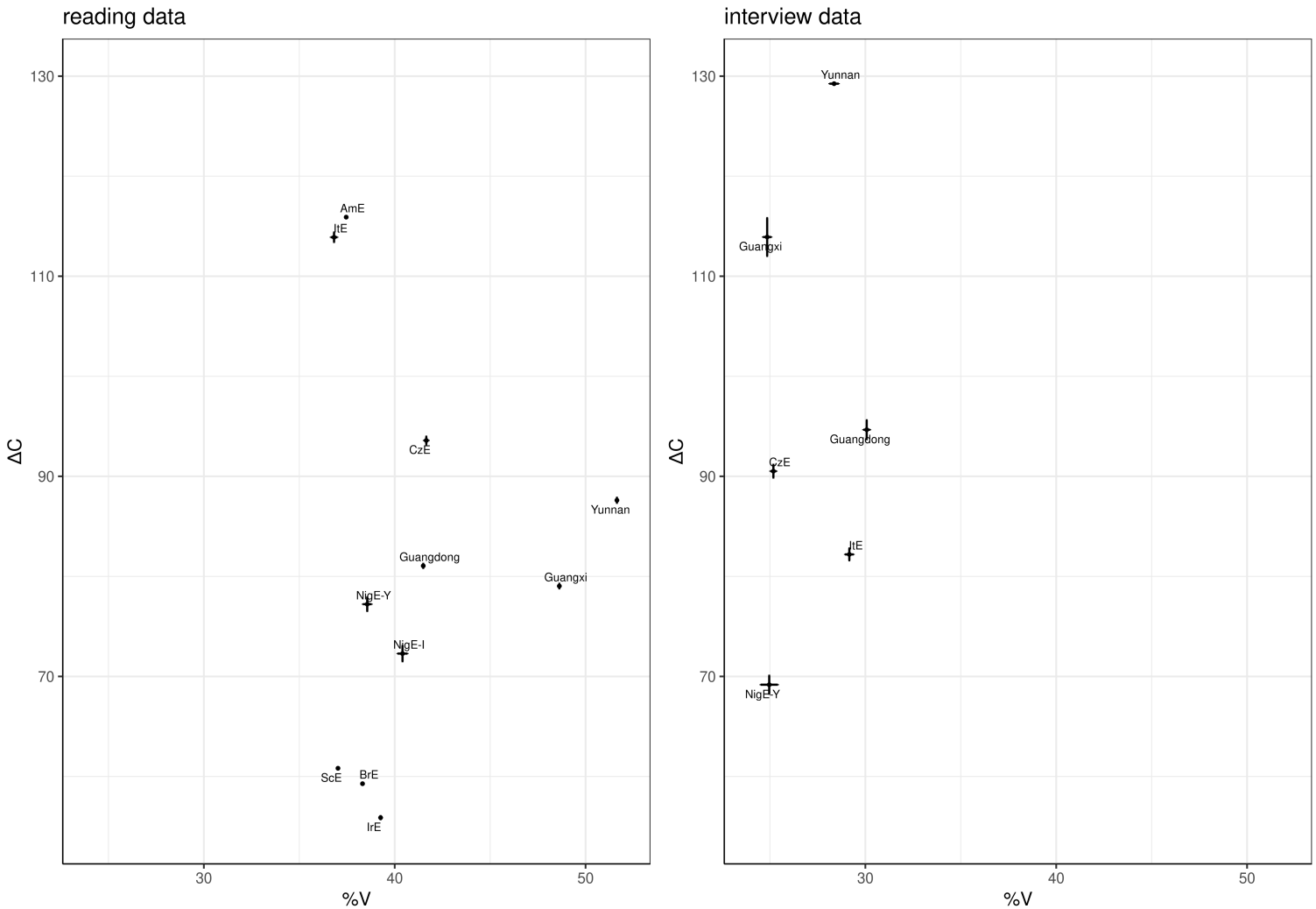

Figure 15: ΔC - %V Chart for reading and interview data

Speech Rhythm Results

Variety	nPVI-V	rPVI-C	VarcoV	VarcoC	ΔC	ΔV	%V
BrE	59.77	66.52	53.38	64.52	59.28	46.31	38.31
ScE	56.88	68.56	52.61	59.11	60.82	42.53	37.03
IrE	57.29	62.25	52.14	60.80	55.88	42.82	39.26
AmE	56.16	107.91	75.72	57.74	115.91	54.18	37.46
NiE-Y	59.04	90.69	60.54	86.98	77.22	53.34	38.56
NiE-I	58.80	84.83	59.21	66.34	72.28	52.01	40.41
CzE	63.66	112.46	71.56	76.88	93.59	75.81	41.65
ItE	60.04	128.84	75.20	76.42	113.90	64.61	36.82
ChE-Guangdong	76.25	92.45	91.34	127.05	81.05	100.14	41.49
ChE-Guangxi	80.72	89.29	100.89	133.74	79.03	117.32	48.62
ChE-Yunnan	84.26	99.46	108.42	140.52	87.61	131.44	51.64

Table 3: Results of all speech rhythm metrics for the reading data

Variety	nPVI-V	rPVI-C	VarcoV	VarcoC	ΔC	ΔV	%V
NiE-Y	64.4	83.19	77.94	107.12	69.17	55.85	24.96
CzE	68.65	104.95	91.77	109.04	90.52	75.27	25.18
ItE	73.11	96.36	82.78	109.75	82.2	87.33	29.16
ChE-Guangdong	69.76	114.29	121.84	94.14	94.66	87.09	30.07
ChE-Guangxi	68.55	136.28	151.90	98.78	113.92	87.51	24.86
ChE-Yunnan	71.59	149.54	255.96	101.29	129.24	105.93	28.36

Table 4: Results of all speech rhythm metrics for the interview data

Speech Rhythm Conclusion

- inner circle varieties clearly stress-timed
- outer circle variety also rather stress-timed
- expanding circle varieties more diverse
  - Czech English more syllable-timed
  - Italian English more stress-timed
  - Chinese English more syllable-timed, with subnational variation
- not all speech rhythm metrics are equally robust
- a minimal threshold of fluency, measured by speech rate, seems to be required
- spoken genre (reading vs. interview) has an impact on speech rhythm metric results

Variety	Speech Rate	Articulation Rate
AmE	4.00	10.71
BrE	3.26	13.91
ScE	3.21	12.52
IrE	3.48	13.85
NiE-Y	3.33	10.74
NiE-I	3.16	9.98
CzE	2.62	8.27
ItE	2.70	8.80
ChE-GD	1.89	8.55
ChE-GX	1.92	7.83
ChE-YN	1.98	7.33

Table 5: Speech rate and articulation rate of read data

Variety	Speech Rate	Articulation Rate
CzE	2.69	7.80
ItE	1.89	8.26
NiE-Y	2.06	11.37
ChE-GD	1.53	8.44
ChE-GX	1.18	8.37
ChE-YN	1.28	7.36

Table 6: Speech rate and articulation rate of interview data

References

Albrecht, S. (2021). Current research on the linguistic features of Chinese English. *World Englishes*.

Albrecht, S., Tamboli, R., Taubert, S., Eibl, M., Rey, G. D., & Schmied, J. (2022). Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems: Measuring Monophthong Naturalness. *2022 IEEE 9th International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 1–6. https://doi.org/10.1109/CIVEMSA53371.2022.9853712

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*.

Chang, J. (2001). Chinese Speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (Second, pp. 310–324). Cambridge University Press.

Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62. https://doi.org/10.1016/S0095-4470(19)30776-4

Dellwo, V. (2006). Rhythm and Speech Rate: A Variation Coefficient for deltaC. https://doi.org/10.5167/UZH-111789

Deterding, D. (2006). The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association*, *36*(2), 187–196.

Fuchs, R. (2023). Analysing the speech rhythm of New Englishes: A guide to researchers and a case study in Pakistani, Philippine, Nigerian and British English. In G. Wilson & M. Westphal (Eds.), *New Englishes, New Methods*. Benjamins.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 515–546). De Gruyter Mouton. https://doi.org/10.1515/9783110197105.2.515

Halliday, M. A. K., & Matthiessen, C. M. (2013). *Halliday's introduction to functional grammar*. Routledge.

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

Kachru, B. B. (1992). Models for non-native Englishes. In B. B. Kachru (Ed.), *The Other Tongue: English across Cultures* (Second, pp. 48–74). University of Illinois Press.

Liang, J., & Li, D. C. (2017). Researching Collocational Features: Towards China English as a Distinctive New Variety. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The State of the Art* (pp. 61–75). Springer.

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.

Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, *43*(4), 377–401. https://doi.org/10.1177/00238309000430040301

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

Pike, K. L. (1945). *The Intonation of American English*. University of Michigan Press.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292. https://doi.org/10.1016/S0010-0277(99)00058-X

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.

Wang, M., & Li, J. (1993). An investigation on English discourse patterns used by Chinese students. *Waiyu Jiaoxue Yu Yanjiu [Foreign Language and Teaching Research]*, 59–64.

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(4), 665–679. https://doi.org/10.1016/j.jml.2011.12.010

Xu, Z. (2008). Analysis of Syntactic Features of Chinese English. *Asian Englishes*, *11*(2), 4–31.

Xu, Z. (2020). Chinese English. In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes* (pp. 265–297). Routledge.

Linguistic Perspectives on Building a Neural TTS System for Teaching and Learning Settings

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

TU Chemnitz

16.05.2023

HYBRID SOCIETIES
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation