Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems: Measuring Monophthong Naturalness

CIVEMSA 2022

Sven Albrecht, Rewa Tamboli, Stefan Taubert

TU Chemnitz

15.-17.06.2022

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?

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

Currently we are able to control:

- morphosyntactic cues with a rule based approach
- phonetic cues with a phone-based TTS

TTS System

Our TTS pipeline is based on

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

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

Audio Examples

`https://stefantaubert.github.io/CIVEMSA-2022`

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

TTS System - Training

Tacotron

- dataset: LJ Speech
- training set: 23 hours 25 minutes
- validation set: 30 minutes
- 500 epochs
- batchsize: 64
- learning rate: 0.001

For WaveGlow we used a public available pretrained model from Nvidia.

Measuring Vowel Spaces

Quantification workflow

1. forced alignment using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017)
2. automated vowel formant measurements in Praat[2]
3. Hampel filtering of outliers (Hampel, 1974)
4. speaker intrinsic, vowel extrinsic, formant intrinsic normalization (Lobanov, 1971)
5. finding convex polygon hull (all data points, central 75%, individual phonemes) using algorithm by Eddy (1977)
6. calculating surface area of polygons and overlap percentage[3]

[2] Praat script available at: `https://mytuc.org/lnvn`
[3] R Script available at: `https://mytuc.org/mthv`

Results: Vowel Space

Table 1: Vowel Space Overlap

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

Results: Vowel Space

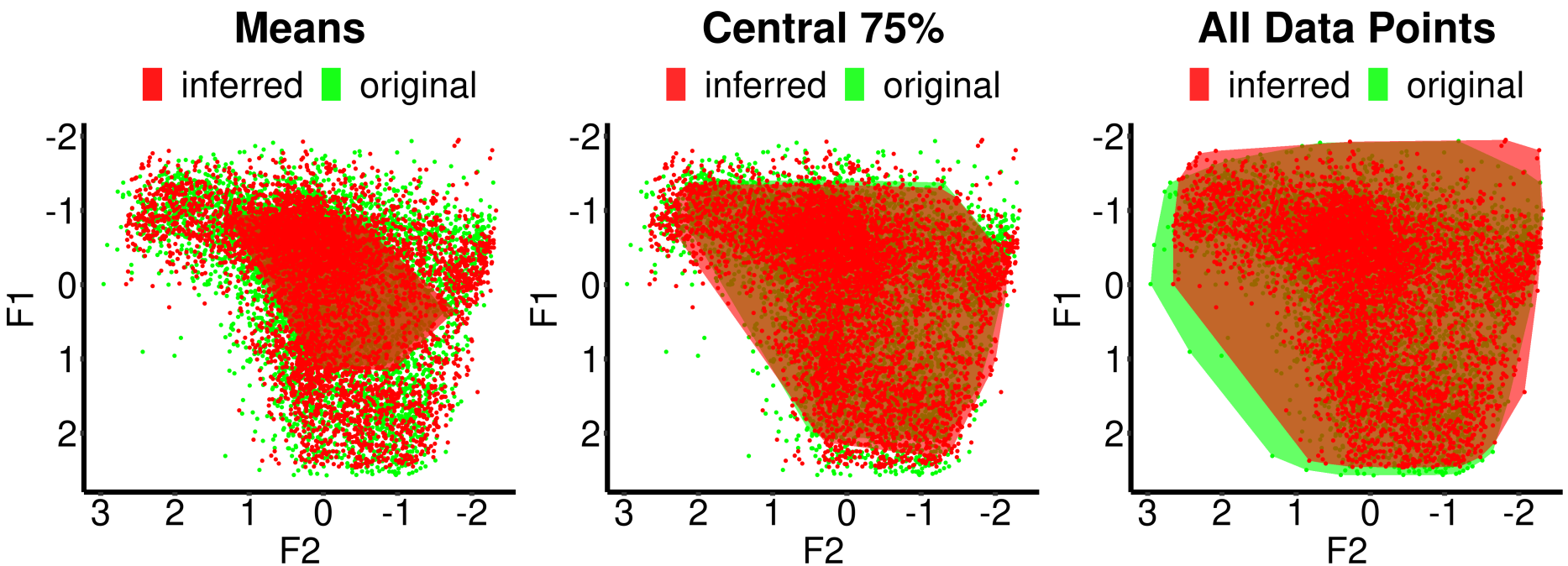

Figure 1: LJ Speech Vowel Space, plotted as F1 - F2 space (Lobanov Normalized, Hampel Filtered)

Results: Vowel Space by Phoneme

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%

Results: Vowel Space by Phoneme (all data points)

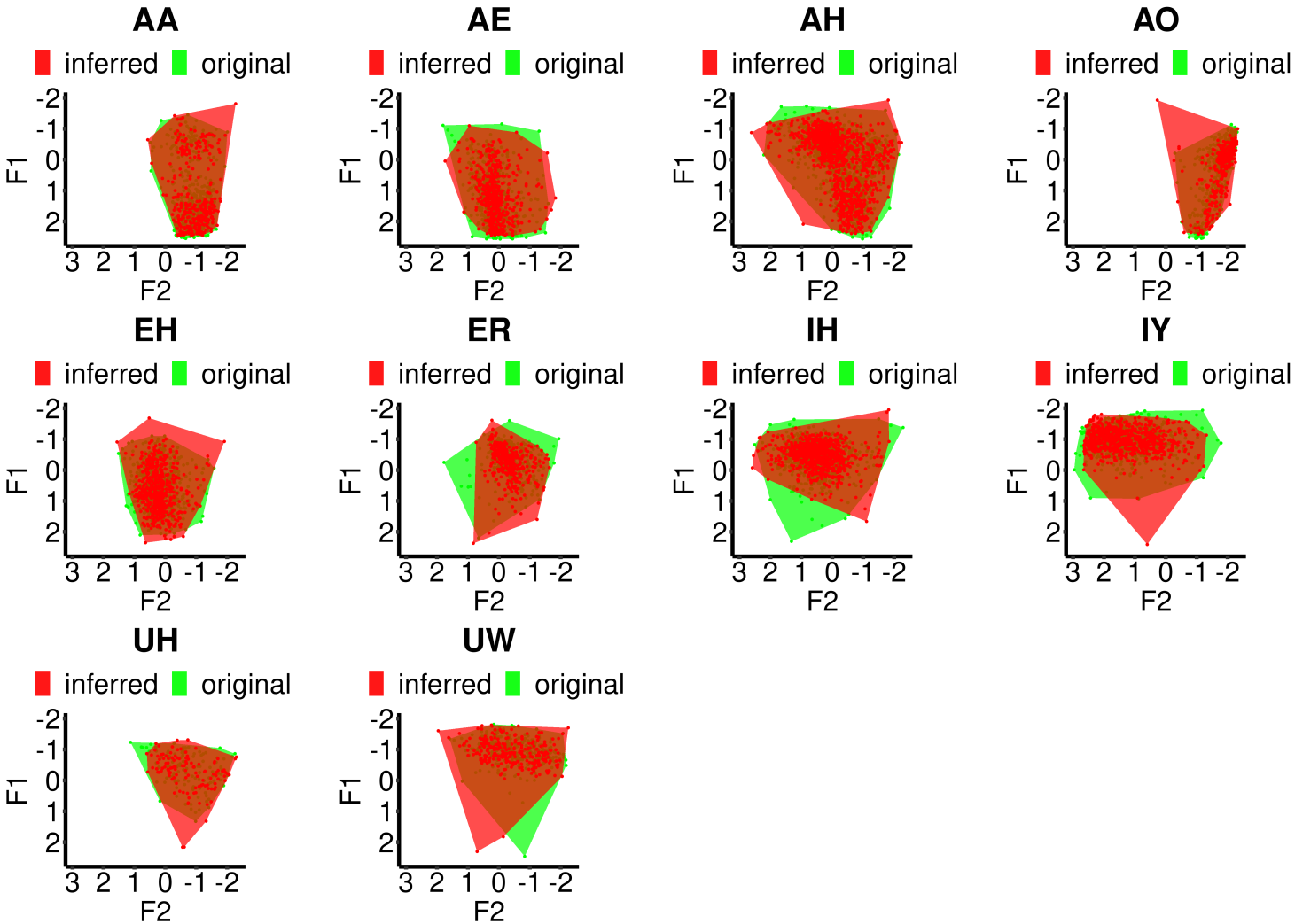

Figure 2: LJ Speech Phoneme Space All Data Points (Lobanov Normalized, Hampel Filtered)

Results: Vowel Space by Phoneme (central 75%)

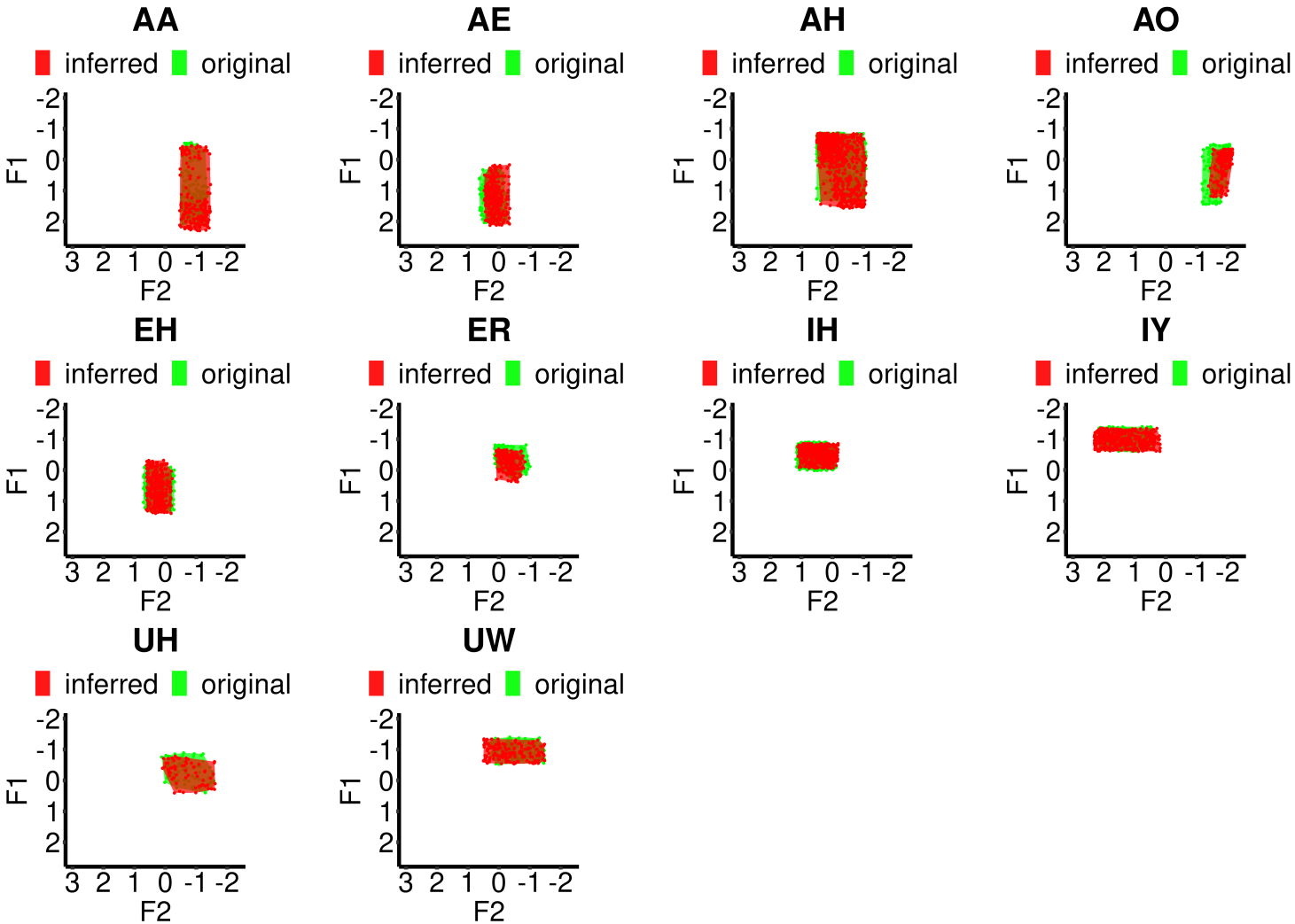

Figure 3: LJ Speech Phoneme Space Central 75% of All Data Points (Lobanov Normalized, Hampel Filtered)

Discussion

Limitations of our approach:

- phoneme alignment is not 100% accurate, though MFA performs well
- quality of the pronunciation dictionary directly impacts transcriptions
- currently only monophthongs, no diphthongs
- no differentiation between stressed and unstressed vowels
- minimum threshold of neural TTS training required for our measurements to work
- output of TTS pipeline (Tacotron 2 + WaveGlow) measured

more research needed, esp. comparing our metric to subjective measures (mean opinion score)

Conclusion

- vowel formant measurements provide a good basis a linguistically quality metric for TTS systems
- vowel space plots provide a good estimate of the quality of synthesized speech
- overlap percentages of 91.50% to 97.70% for the whole vowel space and 56% to 98% for individual vowels
- comparing the vowel spaces of individual vowels provides insights into how well certain vowels can be synthesized
- vowel formant metric can help in targeted optimization of TTS system

References

Eddy, W. F. (1977). Algorithm 523: CONVEX, a new convex hull algorithm for planar sets [Z]. *ACM Transactions on Mathematical Software (TOMS)*, *3*(4), 411–412.

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

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

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

Prenger, R., Valle, R., & Catanzaro, B. (2019). WaveGlow: A Flow-based Generative Network for Speech Synthesis. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 3617-3621). 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). Calgary, AB: IEEE. doi: 10.1109/ICASSP.2018.8461368

Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems: Measuring Monophthong Naturalness

CIVEMSA 2022

Sven Albrecht, Rewa Tamboli, Stefan Taubert

TU Chemnitz

15.-17.06.2022

HYBRID SOCIETIES

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation